



Yu, Boyu (2026) *A data-driven approach for uncovering translational signatures of short open reading frames using ribosome profiling and RNA-Seq data*. PhD thesis.

<https://theses.gla.ac.uk/86143/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

1 **A data-driven approach for**
 uncovering translational signatures
 of short open reading frames using
 ribosome profiling and RNA-Seq
 data

2 Boyue Yu

 Cancer Research UK Scotland Institute
 and
 University of Glasgow

4 Submitted in fulfilment of the requirements
5 for the degree of
6 Doctor of Philosophy

7 September 2024



University
of Glasgow



CANCER
RESEARCH
UK

Scotland
Institute

Abstract

10 Upstream open reading frames (uORFs) are emerging as critical regulatory elements in gene expression, influ-
11 encing diverse cellular processes and disease mechanisms. Despite their significance, current ORF detection
12 workflows are inadequate for capturing uORFs due to several limitations: most existing tools are optimised for
13 long, main coding regions, often overlook non-AUG initiation, and fail to accurately identify uORFs that are
14 short and exhibit low ribosome occupancy. This thesis addresses these challenges by developing an advanced
15 computational workflow specifically tailored for the detection and characterisation of uORFs using ribosome
16 profiling (Ribo-Seq) and RNA sequencing (RNA-Seq) data.

17 The workflow standardises the raw sequencing data pipeline, enhances the detection rate of translated ORFs,
18 and constructs a feature matrix for detailed uORF characterisation. It also includes a series of downstream
19 analyses to investigate uORF functionality within experimental setups. Using this workflow, significant find-
20 ings were uncovered, including codon usage biases embedded within uORFs across different cell cycle-related
21 gene sets, homologous uORF sequences, and the potential regulatory effects of uORFs on downstream coding
22 sequences (CDSs).

23 This research not only improves the accuracy of uORF detection but also provides deeper insights into their
24 regulatory roles, bridging key gaps in current genomic annotations. The developed framework advances our
25 understanding of non-canonical translation and offers valuable tools for future studies exploring the impact of
26 uORFs in gene regulation and disease pathogenesis.

Acknowledgements

28 This is my first time writing a proper acknowledgement, and after extensive academic writing, I feel that, to
29 some extent, I have lost my ability to express my feelings objectively. Nevertheless, I will then try to use my
30 clumsy words to express my deepest gratitude to everyone who has supported me throughout my PhD journey.

31 First and foremost, I am profoundly thankful to my supervisor, Crispin Miller, for his invaluable guidance,
32 encouragement, and insightful feedback throughout my research. This project was relatively open-ended, and
33 with his support and confidence in my abilities, I was able to explore the intricate world of encoding non-
34 coding sequences. I am especially grateful for his emotional support during times when I felt depressed;
35 his conversations always had a unique power to calm and reassure me. Pursuing a PhD is not solely about
36 accumulating knowledge but also about learning how to approach questions and resolve them efficiently. I
37 believe that this is what he has taught me the most. I would also like to express my gratitude to my second
38 supervisor Martin Bushell, my work would not be possible without his invaluable insights in translational
39 control.

40 I am so grateful to have met all the lovely lab members from both the Miller Lab and the Bushell Lab. A special
41 shout-out to Joseph Waldron, who generously donated his sequencing data to my project and introduced me to
42 the world of Ribo-Seq. Meanwhile, huge thanks to Alexandrina Pancheva, who not only kept me on track with
43 thesis writing but also proofread it multiple times, even during her off-work hours. Many thanks to Mayank
44 Sikarwar, who noticed my struggles with depression and consistently reached out to check if I was doing well
45 and to spend time with me. I would like to thank Eva Freckmann for providing all the amazing bakeries during
46 our many coffee hours and for her similarly brilliant comments on my thesis. I would like to thank Holly Hall
47 for her insightful suggestions throughout my entire PhD. I would like to thank Andrew Papanastasiou for not
48 only his critical thoughts on my project but also for answering many of my imaginative astrophysics questions.
49 I would like to thank Robin Shaw for introducing the sequencing workflow and the "Glasgow Introduction
50 101" course. I would also like to thank Ifedayo Ojo, who is both a great colleague and a friend, supporting
51 me during my last year of PhD. Additionally, I would like to thank Aditya Chandru, who shared so many good
52 times and chats with me and is a great coach for ice skating. Also, I want to thank Britt van Abeelen as my PhD
53 pal in this group and as the student representative who speaks out for all the students. Meanwhile, I would like
54 to thank Naveed Khan, who is the guardian of the high-performance computer and a brilliant chef, feeding our

55 group with both computing resources and amazing food. Additionally, I would like to thank Peter McHardy
56 and Iain White for taking care of the entire networking and computing infrastructure of the institute.

57 I would also like to thank Cancer Research UK for making my PhD possible. This work would not have been
58 possible without their generous funding.

59 Last but not least, I must thank my wife, Qingcheng, who bravely came to the UK with me during the COVID-
60 19 pandemic. She has been working abroad for almost four years. To cope with the different time zones, her
61 alarm is always set to six in the morning. I cannot appreciate enough her sacrifices and support, both mentally
62 and physically.

63 Additionally, I need to acknowledge ChatGPT for its assistance in proofreading and refining the language of
64 this thesis according to Dritjon Gruda's paper (2024). No AI-generated content was included in the substantive
65 parts of this research, and all original ideas and analyses presented in this thesis are my own.

Author's Declaration

67 I declare that the work presented in this thesis and during my PhD was performed by myself, with the exception
68 of wet-lab mouse organoid experiments conducted by Dr. Joseph Waldron, who granted me permission to
69 present his work in the thesis.

70 This dissertation has not been submitted, in whole or in part, in any previous application for a degree.

71 Boyu Yu

72 September, 2024

List of Figures

74	1	Schematic representation of the 5' UTR elements	16
75	2	Translation regulation of ATF4	18
76	3	Canonical eukaryotic translation initiation	19
77	4	Schematic diagram of potential translation reinitiation in eukaryotic cell	20
78	5	Overview of Ribo-Seq and RNA-Seq	21
79	6	Chaos game representation	26
80	7	Support vector machine	26
81	8	Illustration of kernel function in support vector machine	27
82	9	Random forest workflow	27
83	10	Receiver operating characteristic curve example	39
84	11	Proposed ORF identification workflow overview	44
85	12	Read structures in sequencing data	46
86	13	Snapshot of Ribo-Seq alignment comparison	48
87	14	Alignment tools performance analysis	49
88	15	Elbow analysis of Short Translatable Unit (STU) coverage	50
89	16	Screening and filtering of Small Translatable Units (STUs).	51
90	17	ORF types	55
91	18	Distribution of key feature scores across ORF types	57
92	19	mTORC1 and MNK signalling in translation initiation	60

93	20	ORF identification rate comparisons in the mouse project	65
94	21	Local read depth across positive and negative uORF samples	67
95	22	ORF identification rate comparisons in the human project	68
96	23	Mouse translated CDS comparison between computational prediction and panel of translating	
97		CDS	71
98	24	Portal design and data flow	73
99	25	Portal interface	74
100	26	Principal component analysis representation of the feature matrix	78
101	27	Codon usage differences between G2-M and G1-S uORFs	97
102	28	Codon usage differences between uORFs and CDSs	98
103	29	Null model for the codon usage comparison	99
104	30	GC content trends across mouse transcriptome and within transcripts	100
105	31	Violin plot of GC content distribution	100
106	32	Synonymous codon mutations in all TCGA projects	101
107	33	Conservation landscape of uORFs	102
108	34	Homologous hits of an uORF (Racgap1) on mouse chromosome ideogram with corresponding	
109		GSEA	102
110	35	Batch effect visualisation of RNA-Seq and Ribo-Seq data	103
111	36	Differential expression analysis	103
112	37	Correlation between RNA-Seq and Ribo-Seq reads in CDS and uORFs	104
113	38	Correlations between uORF and CDS ribosome occupancy changes	104
114	39	Gene set enrichment analysis of genes where its uORFs have potential regulatory effects on CDS	105
115	40	Feature importance rankings in random forest models for predicting cell viability	106
116	41	Impact of information gain on cluster clarity	106
117	42	ROC analysis of Random Forest models discriminating translated CDSs	107
118	43	Feature importance rankings in random forest models for predicting translating CDSs	108
119	44	Random forest model performance and UMAP visualisation of ORF classification	108
120	S1	BLAST homologous hits	133

121	S2	uORF identification workflow of Year 1	134
122	S3	ORF analysis workflow of Year 2	135
123	S4	Homologous hits of an uORF (Ptrh2) on mouse chromosome ideogram with corresponding	
124		GSEA	135
125	S5	Homologous hits of an uORF (Nek4) on mouse chromosome ideogram with corresponding	
126		GSEA	136
127	S6	Codon usage differences between uORFs and CDSs	136

List of Tables

129	1	Ribo-Seq read length offset settings	37
130	2	List of installed R packages	40
131	3	Performance metrics for ORF identification methods	72
132	4	Cell cycle related uORF sequence pattern discriminator performance	78
133	5	Null model for sequence pattern discriminator	79
134	6	Features used for Machine Learning models	93
135	7	Performance metrics for RF model evaluation	94
136	8	Gene information and periodic stage	132

Contents

138	Abstract	1
139	Acknowledgements	2
140	Author's Declaration	4
141	List of Figures	5
142	List of Tables	8
143	1 Introduction	13
144	1.1 Translation initiation and ribosome re-initiation	14
145	1.2 5' UTR	15
146	1.3 Upstream open reading frame	17
147	1.4 Computational approaches	19
148	1.4.1 Ribosome profiling and alignment tools	20
149	1.4.2 ORF identification tools	22
150	1.4.3 Machine learning	25
151	1.5 Summary	29
152	2 Methods	31
153	2.1 Data sources	31
154	2.1.1 MNK mTOR mouse sequencing	31
155	2.1.2 iPSC human experiment	34

156	2.2	Software and development environment	35
157	2.2.1	Raw data processing	35
158	2.2.2	ORF identification algorithms	36
159	2.2.3	Feature Engineering	37
160	3	Development of the BYU Workflow for Translated uORF Identification	41
161	3.1	Standardising raw sequence data pre-processing	43
162	3.1.1	Adapter removal	45
163	3.1.2	De-duplication	45
164	3.1.3	UMI removal	45
165	3.1.4	Genome alignments tool benchmarking	45
166	3.2	Translated ORF candidates screening prioritises translatability and transcriptional coverage . .	47
167	3.2.1	Forming and filtering the small translatable units (STUs)	48
168	3.3	Constructing a multi-dimensional feature matrix for ORF characterisation	52
169	3.3.1	Homologous sequence analysis	52
170	3.3.2	Sequence conservation landscape calculation	53
171	3.3.3	Secondary structure measurement based on minimum free energy	53
172	3.3.4	ORF type	54
173	3.3.5	Kozak score	55
174	3.3.6	External annotation and standardisation	57
175	3.3.7	Conclusion	58
176	4	Benchmarking and Validation of the BYU Workflow in Organoid and iPSC Models	59
177	4.1	Introduction	59
178	4.1.1	Mouse Experiment	59
179	4.1.2	Human sequencing dataset	62
180	4.2	ORF identification rate	63
181	4.2.1	Existing algorithms show low sensitivities and replicate agreements for uORFs	63
182	4.2.2	BYU achieves higher sensitivity and replicate consistency	67

183	4.2.3	Comparison against the background demonstrates effective noise filtration	69
184	4.2.4	Comparative validation against gold standard datasets confirms higher sensitivity . . .	69
185	4.2.5	Summary	73
186	4.3	R Shiny data portal and SQLite backend	73
187	4.4	Summary	74
188	5	Systematic Characterisation of uORF Candidates: Sequence Pattern, Evolution Conservation,	
189		and Translation Regulatory	76
190	5.1	Chaos game representation reveals distinct sequence patterns in cell cycle-related uORFs . . .	77
191	5.2	Codon analysis reveals phase-dependent AT3/GC3 biases in uORFs	80
192	5.2.1	Tumour-associated synonymous mutations in uORFs exhibit an AT3 codon bias	83
193	5.3	Conservation landscape analysis highlights the evolutionary selection of functional uORFs . .	84
194	5.4	Homologous sequence analysis identifies conserved uORFs in co-regulated gene families . . .	86
195	5.5	Differential expression analysis	89
196	5.6	Machine learning	92
197	5.6.1	Datasets and features for machine learning	92
198	5.6.2	Feature integration enables clustering of ORF types in UMAP	92
199	5.6.3	Predictive modelling of translated CDSs	93
200	5.6.4	Predictive modelling of functionally validated uORFs	94
201	5.6.5	Summary	95
202	5.7	Conclusion	95
203	6	Discussion	109
204	6.1	Overview of the research	110
205	6.2	Improved workflow for uORF detection: from translation to function	111
206	6.2.1	Rationale and philosophical underpinnings of the BYU workflow	112
207	6.3	Key findings in downstream analysis	114
208	6.3.1	Sequence patterns emerging in uORF of cell cycle related genes	114
209	6.3.2	Codon basis in uORFs of different cell cycle genes and between uORF and CDS . . .	115

	CONTENTS	12
210	6.3.3 Homologous uORF identified	115
211	6.3.4 uORFs as potential regulators of CDS translation	116
212	6.4 Conclusion	116
213	References	118
214	Appendices	132

Chapter 1

Introduction

In recent years, post-transcriptional regulation has been shown to be a key step in controlling gene expression. A sequence of tightly regulated events, including mRNA processing, localisation, stabilisation, and translation, is required for the proper synthesis of proteins (Audic and Hartley 2004; Zhao, Roundtree, and He 2017). The variety of regulatory events controlling gene expression is revealed by the differing responses of cells to specific conditions. The 5' and 3' UTRs in mRNAs have been shown to contain functional elements serving as targets for the post-transcriptional regulation pathways (Vejnar et al. 2019). These include functional motifs that act as binding sites for RNA-binding proteins, miRNAs, and other trans-acting factors. Meanwhile, the secondary structure motifs in UTRs, such as pseudoknots, hairpins, and other cis-regulatory elements, can also regulate the initiation of translation and mRNA stability (Faure et al. 2016; Leppek, Das, and Barna 2018). In addition, UTRs play an important role in pre-mRNA processing, including the control of alternative splicing and the determination of poly-A tail position and size (Manning and Cooper 2017; Mockenhaupt and Makeyev 2015).

UTRs are often exploited by diseases such as cancer, underscoring their regulatory importance (Schuster and Hsieh 2019). For instance, a frequent mutation in the 5' UTR of Ribosomal Protein S27 (RPS27) has been identified in melanoma, where a C to T substitution at the transcription start site potentially increases RPS27 expression, contributing to tumour progression (Dutton-Regester et al. 2014). Similarly, a 3' UTR signature in triple-negative breast cancer has been linked to event-free survival, and 3' UTR shortening generally correlates with poor prognosis in solid tumours (Schuster and Hsieh 2019). These findings suggest that UTR mutations could function as driver mutations, though they are often under-annotated compared to coding regions (Steri et al. 2018).

Genome-Wide Association Studies (GWAS) have linked millions of genetic variants with various disorders, yet only around 4% of these variants affect protein sequences, while about 88% occur in noncoding regions, including UTRs (Manning and Cooper 2017; Steri et al. 2018). Although some UTR elements, such as RNA binding protein (RBP) binding sites, are conserved, other motifs like specific secondary structures are highly

polymorphic and gene-specific (Bae and Miura 2020; Matoulkova et al. 2012; Telenti et al. 2016). A complete understanding of the elements and motifs in UTRs has yet to be achieved, and how these functional groups interact to regulate post-transcriptional events, including translation associated with tumourigenesis, remains unclear.

1.1 Translation initiation and ribosome re-initiation

Translation initiation is a sophisticated process of recruiting ribosomes onto the mRNA and scanning for the start codon (Figure 3). In eukaryotic cells, more than 12 initiation factors normally participate in this process (Lehninger, Nelson, and Cox 2005). Initially, the eukaryotic initiation factors (eIF) 1, 1A and 3 bind to the 40S ribosome subunit. Meanwhile the eIF2 forms a ternary complex (eIF2-TC) with tRNA^{Met} and GTP (eIF2-GTP-tRNA^{Met} ternary complex). This eIF2-TC availability itself is critical for the translation initiation rate (Adomavicius et al. 2019). Only the GTP bound eIF2 has high affinity to tRNA^{Met}. Upon recognition of the start codon, the GTP on eIF2 is hydrolysed to GDP, and the eIF2-GDP complex is released from the ribosome (Adomavicius et al. 2019; Starck et al. 2016). For translation to be continued, this inactive eIF2-GDP must be recharged back to eIF2-GTP by the guanine nucleotide exchange factor (GEF) known as eIF2B. This recycling of eIF2-TC mechanism is the primary control point for the Integrated Stress Response (ISR) to control the global translation. In response to various cellular stressors, multiple protein kinases become activated and phosphorylate a single serine (Serine 51) on the alpha subunit of eIF2 (eIF2 α) (Jackson, Hellen, and Pestova 2010; Proud 2019). The P-eIF2 α then acts as a competitive inhibitor, with a higher affiliation to eIF2B, effectively restricting it in the non-productive complex (Parra et al. 2018). Since the cellular concentration of eIF2 is significantly higher than eIF2B, even a small fraction of eIF2 α being phosphorylated is sufficient to trap the limited pool of eIF2B and shut down translation. While this suppresses general protein synthesis, it selectively favours the translation of a range of stress-responsive genes, such as ATF4. More details of ATF4 activation will be introduced later.

Then eIF3 stimulates the binding of eIF2-TC to the 40S, forming the 43S preinitiation complex and further promoting the attachment of the 43S complex to mRNA. In particular, the eIF1 is critical in ensuring the fidelity of initiation at codon AUG. eIF1 allows 43S complexes to distinguish between AUG codons and non-AUG triplets or those in a suboptimal context (Jackson, Hellen, and Pestova 2010). Furthermore, it facilitates the discrimination against AUGs located within 8 nucleotides of the 5' cap and actively dissociates any ribosomal complexes that have assembled incorrectly at these sites (Haimov, Sinvani, and Dikstein 2015). With the co-operation of eIF1A, the 43S complex is switched to a scanning-competent exposed conformation. To establish the stable base pairing of the ribosome and mRNA, eIF1 needs to be displaced thereby closing this 'open' conformation of 43S complex. The level of eIF1 has been shown to be important to AUG/non-AUG initiation switching (Kearse and Wilusz 2017; Andreev, Gary Loughran, et al. 2022).

Before this attachment, the mRNA is activated and circularised by the eIF4F complex, consisting of eIF4E, 4G,

274 and 4A. eIF4E recognises and binds the 5' cap, while eIF4G and 4A associate, and poly(A)-binding protein
275 (PABP) binds eIF4G, interacting with the 3' poly(A) tail to stabilise the mRNA. The 43S complex scans the
276 mRNA for the start codon, with eIF4A acting as an RNA helicase to unwind secondary structures, aided by
277 eIF4G and eIF4B. eIF1 and eIF1A then help recognise the translation initiation site through stable codon-
278 anticodon base-pairing. Recognition of the start codon triggers GTP hydrolysis in eIF2, leading to dissociation
279 of the GDP-eIF2 complex as well as other initiation factors, and the binding of the 60S ribosomal subunit. With
280 the assistance of eIF5B, the 60S subunit binds the 40S, forming the 80S ribosome, marking the completion of
281 translation initiation (Jackson, Hellen, and Pestova 2010; Adomavicius et al. 2019).

282 After initiation, translation proceeds through elongation until reaching a stop codon, where eukaryotic release
283 factors 1 and 3 facilitate termination, releasing the 60S subunit. The 40S subunit may remain attached to the
284 mRNA, resuming scanning then re-initiating the translation upon new start codon identification. However,
285 translation cannot reinitiate without acquiring an eIF2-TC due to the absence of initiator tRNA. As mentioned
286 earlier, the abundance of eIF2-TC determines the speed of translation reinitiation. Furthermore, while the
287 initial translation of a uORF is largely controlled by canonical initiation factors such as the eIF4F complex
288 recruiting the 43S pre-initiation complex, and local sequence features like Kozak sequences, the subsequent
289 fate of the ribosome depends heavily on the mechanics of termination and re-initiation. With the advent of
290 translation complex profile sequencing (TCP-Seq), studies on yeast and human uORF have revealed some
291 potential mechanisms of ribosome reinitiation following uORF translation termination (Mohammad et al. 2021;
292 Wagner, Herrmannová, et al. 2020). It is suggested that several initiation factors, including eIF3, eIF4E,
293 and eIF4G, may remain attached to the 40S subunit throughout the initiation, elongation, and termination
294 of uORFs (Bohlen, Fenzl, et al. 2020). However, the efficiency or the ability of the re-initiation after uORF
295 translation is discovered to be dependent on the duration of uORF translation, which is determined upon the
296 uORF length and translation elongation rates (Munzarová et al. 2011; Kozak 2001). Following the disassembly
297 of the 80S complex, the bound eIF3 can stimulate the binding of eIF2-TC to the remaining 40S subunit, thereby
298 reinitiating translation with the involvement of other initiation factors (Figure 4). However, if the uORF length
299 increases, the stability of the 40S and eIF3 interaction might decay (Szamecz et al. 2008; Kozak 2001). Kozak
300 has shown that when uORF length increased from 13 to 33 codons, the translation is dropped significantly
301 (2001). Therefore, a long uORF might act like a trap, as the extended elongation time would allow eIF3
302 and other factors to dissociate. Without these factors, the 40S subunit cannot recruit fresh eIF2-TCs and then
303 dissociates from the transcript, preventing translation of the downstream main coding sequence.

304 1.2 5' UTR

305 The 5' UTR, defined as the sequence between the transcription start site (TSS) and the start codon of the main
306 open reading frame (ORF), plays a crucial role in post-transcriptional regulation (Leppek, Das, and Barna
307 2018). The median length of the 5' UTR ranges from 50 to 220 nucleotides across eukaryotes, but within
308 a species, the length can vary significantly (Araujo et al. 2012). For instance, some human mRNAs have

309 ultra-short 5' UTRs facilitating scanning-free translation initiation, while some may have long 5' UTRs with
 310 complex regulatory elements, indicating a higher level of control (Araujo et al. 2012; Leppek, Das, and Barna
 311 2018; Elfakess et al. 2011). Meanwhile, genes with longer 5' UTRs, such as TP53, RARB and those encoding
 312 transcriptional regulators and protooncogenes, are often hotspots for tumourigenic mutations (Davuluri 2000).

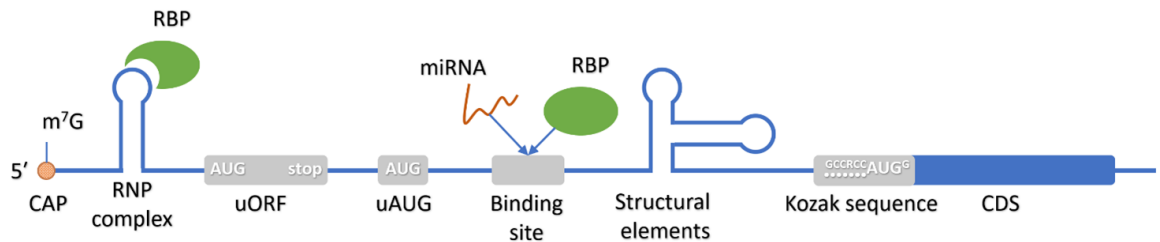


Figure 1: **Schematic representation of the 5' UTR elements** illustrates the key structural and functional elements present in the 5' untranslated region (UTR) of an mRNA. The 7-methylguanylate cap (m^7G) at the 5' end of the mRNA, and the Kozak consensus sequence, which contains the start codon for the main ORF, are the boundaries of the 5' UTR in a mature mRNA. In between, there are multiple cis- and trans-acting regulatory elements, and the number and order of motifs in the 5' UTR can vary. RNA-binding proteins (RBPs) can interact with RNA binding sites or form ribonucleoprotein (RNP) complexes, thereby regulating translation initiation. Upstream ORFs (uORFs) and upstream AUGs (uAUGs) can be found in some mRNAs, which often result in poor translation efficiency. Several structural elements, such as pseudoknots, hairpins, and RNA G-quadruplexes, are also present within the 5' UTR, mostly inhibiting translation initiation. The Kozak sequence is a conserved translation initiation site (mostly GCCRCCAUGG in mammals).

313 Recent studies have discovered multiple functional elements located in the 5' UTR which are responsible for
 314 mediating post-transcriptional regulation (Figure 1) (Leppek, Das, and Barna 2018; Wongfieng et al. 2017).
 315 These elements include the 5' cap and the Kozak consensus sequence, which are highly conserved across
 316 eukaryotes, as well as elements that facilitate cap-dependent and cap-independent (IRES) translation initia-
 317 tion (Kumar, Srinivas, and Patankar 2015; Araujo et al. 2012; Leppek, Das, and Barna 2018). For example,
 318 Glioma-associated oncogene homolog 1 (Gli1), associated with the Hedgehog signalling pathway, is com-
 319 monly found aberrantly activated in multiple cancers such as glioblastoma and breast cancers (Doheny et al.
 320 2020; Palaniswamy et al. 2010). Palaniswamy and colleagues identified three transcripts of the Gli1 gene that
 321 have different TSSs with 5' UTR polymorphisms, and found that the shortened 5' UTR contributed to increased
 322 protein expression (Palaniswamy et al. 2010). Interestingly, their results also showed that these 5' UTR vari-
 323 ants did not affect mRNA stability or the degradation pattern, suggesting that regulation may occur through the
 324 modulation of translational efficiency. They also found that both the secondary structure of the 5' UTR and the
 325 cap to upstream open reading frame (uORF) distance may play a role in regulating translation (Araujo et al.
 326 2012; Palaniswamy et al. 2010).

327 Upstream ORFs and upstream AUGs are commonly observed in the 5' UTR of poorly translated genes (Davuluri
 328 2000). uORFs and uAUGs have also been found to down-regulate the expression of the main ORF by
 329 interfering with the ribosome recognition of the start codon in the main coding sequence (CDS) (Kumar, Srinivas,
 330 and Patankar 2015). The binding sites for RNA-binding proteins, microRNAs, long noncoding RNAs, and
 331 other trans-acting elements are scattered throughout the 5' UTR. These binding events can be classified into
 332 two types of interactions: some serve as part of the essential translation assembly, such as eukaryotic initiation
 333 factors (eIFs), while others are involved in promoting or repressing translation, such as Uchl1 and Musashi1
 334 (Araujo et al. 2012; Leppek, Das, and Barna 2018). These linear elements are not the only important regulatory

335 factors; structural motifs also play a critical role in regulating translation. Structural elements in the 5' UTR
336 can directly mediate translation initiation, mRNA stability, and decay. They have also been found to regulate
337 the expression of other regulatory elements, such as the binding of RBPs (Taliaferro et al. 2016).

338 1.3 Upstream open reading frame

339 Despite being identified over 30 years ago (Wethmar et al. 2014), uORFs have only recently been recognised
340 for their prevalence due to advances in techniques like ribosome footprinting. A uORF is a sequence, as
341 illustrated in Figure 1, in the 5' UTR that can be translated into a peptide, starting at an AUG (or near-cognate)
342 start codon upstream of the main coding region and ending at a stop codon. uORFs have been predicted to
343 be present in around 50% of transcripts in the human genome (McGillivray et al. 2018; Calvo, Pagliarini, and
344 Mootha 2009). Fundamentally, uORFs are similar to downstream main coding regions, denoted by a start
345 codon and an in-frame stop codon. However, unlike the main CDS, uORFs are found to be less conserved in
346 the translation initiation sites, where noncanonical translation initiation events were found to predominate for
347 the uORFs (Andreev, Gary Loughran, et al. 2022; Kearse and Wilusz 2017; Nicholas T. Ingolia, Lareau, and
348 Weissman 2011). Kozak sequence is a consensus sequence surrounding the start codon that enhances ribosome
349 recognition and initiation of translation. Due to the less presence of Kozak sequence and the canonical AUG
350 start codon in uORF, the initiation efficiency of uORFs may be impaired (Ivaylo P. Ivanov et al. 2011; Diaz
351 de Arce, Noderer, and C. L. Wang 2018).

352 Upstream ORFs can serve as an important regulatory element by affecting the translation of the downstream
353 main ORF. Short peptides encoded by uORFs can regulate translation via various mechanisms, such as inter-
354 acting with specific small molecules in the cellular environment, causing ribosome stalling at the uORF thereby
355 repressing downstream CDS translation (Andrews and Rothnagel 2014; Rahmani et al. 2009). Meanwhile, by
356 influencing ribosome scanning and initiation processes on the main ORF, uORFs can lead to attenuation, en-
357 hancement or not significantly alteration of protein synthesis depending on the context (Calvo, Pagliarini, and
358 Mootha 2009). By regulating the reinitiation of ribosomes after uORF translation, uORFs can play a crucial
359 role in controlling the levels of protein produced from an mRNA (Hinnebusch, I. P. Ivanov, and Sonenberg
360 2016). By initiating translation upstream, ribosomes may never reach the main ORF, typically resulting in
361 reduced translation of the downstream protein. Indeed, studies found uORFs to be prevalent translational re-
362 pressors in vertebrates (Dasgupta and Prensner 2024; W. Xiao et al. 2022). The scanning ribosome begins
363 translating the uORF, and upon termination at the uORF stop codon, the large and small ribosomal subunits
364 dissociate. This prevents those ribosomes from reinitiating at the main ORF, thereby lowering production of
365 the main protein. Additionally, uORF termination can trigger nonsense-mediated mRNA decay if reinitiation
366 is inefficient, further reducing the expression of the gene (Russell et al. 2023).

367 However, emerging evidence indicates that the impact of uORFs on translation is more delicate than a sim-
368 ple universal off switch. Rather than acting as a binary mechanism that uniformly regulate cellular protein

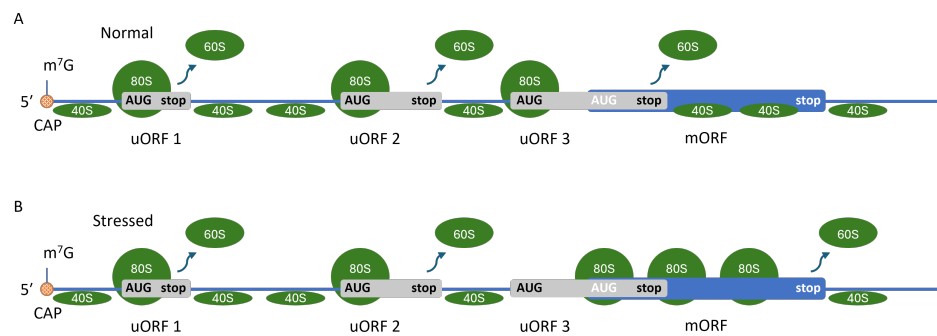


Figure 2: **Translation regulation of ATF4.** Three uORFs are present in the 5' UTR of human ATF4. A) shows that under normal conditions, all uORFs are translated, with the overlapping uORF3 repressing the translation of the main coding region. B) shows that under stress conditions, due to decreased eIF2 α activity levels, ribosome recharging speed is low; therefore, uORF3 is bypassed, and translation reinitiates at the main coding region.

369 levels, the specific moderating ability of a uORF is highly variable and establishes a context based regulatory
 370 landscape. While translation of uORFs generally represses the downstream CDS, the degree of the modera-
 371 tion could be influenced by contextual features such as the Kozak sequence of uORF, uORF conservation, and
 372 distance from cap(Calvo, Pagliarini, and Mootha 2009). And a subset of uORFs serve as context-dependent
 373 regulatory modules. For instance, distinct uORFs within the same transcript, such as those in the MDM2 onco-
 374 gene, can exhibit different suppressive activities based on their specific sequence and structural contexts (Jin
 375 et al. 2003). As in some cases, uORFs may even permit downstream translation under specific conditions.
 376 The human activating transcription factor 4 (ATF4) is a well studied example of uORF mediated translational
 377 control. ATF4 is a key component in the integrated stress response (ISR) pathway. When cells are exposed to
 378 stresses such as proteotoxic stress, eIF2 α will be deactivated thereby activating the ATF4 translation (Chukka,
 379 Wetmore, and Thakor 2021). Three uORFs were found in the human ATF4, two of these are isolated while the
 380 third uORF overlaps with the main CDS (Figure 2) (Bohlen, Harbrecht, et al. 2020).

381 Under normal conditions, the translation of ATF4 is repressed due to the translation of the third uORF (Bohlen,
 382 Harbrecht, et al. 2020). All uORFs are recognised and translated by scanning ribosomes. However, when
 383 ribosomes reinitiate at uORF 3, they translate it and terminate at the uORF 3 stop codon, which prevents the
 384 translation of the main ORF, suppressing the ATF4 protein production. While under stress conditions, eIF2 α is
 385 phosphorylated and deactivated, therefore the levels of active eIF2-GTP-tRNA^{Met} ternary complexes decrease
 386 (Chukka, Wetmore, and Thakor 2021). The decreased availability of ternary complexes delays reinitiation after
 387 uORF 2, causing some ribosomes to scan through uORF 3 and initiate at the downstream ATF4 main ORF start
 388 codon (Bohlen, Harbrecht, et al. 2020). More details about the reinitiation can be found in Chapter 1.1. This
 389 regulatory mechanism demonstrates how uORFs can act as molecular switches, modulating gene expression
 390 in response to cellular conditions by affecting ribosome behaviour during translation initiation and reinitiation
 391 (Qing et al. 2012). Meanwhile, it is important to note that for many transcripts, uORF translation does not
 392 significantly repress or regulate the downstream main CDS. Therefore, fine characterising the uORFs identified
 393 is necessary.

394 Mutations within uORFs can dysregulate gene expression, contributing to potential oncogenesis by altering the

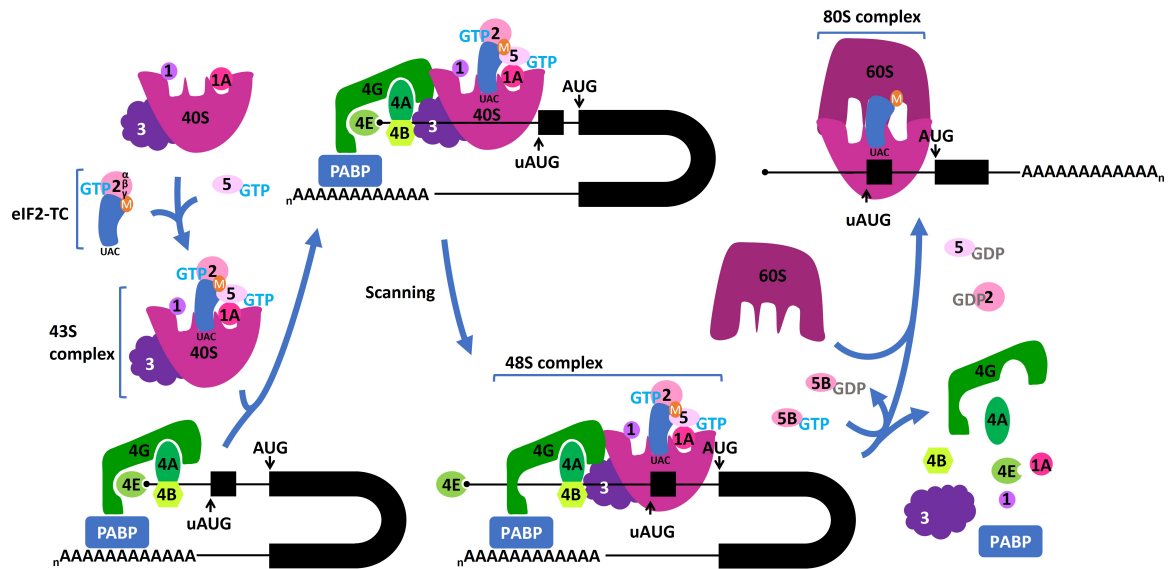


Figure 3: **Canonical eukaryotic translation initiation.** This schematic provides an overview of canonical eukaryotic translation initiation with the highlight of key assemblies. The eukaryotic initiation factor 2 ternary complex (eIF2-TC), which consists of eIF2 α , β , γ , GTP and initiator tRNA, is formed and recruited onto the 40S ribosome subunit along with the eIF 1, 1A and 3, forming the 43S pre-initiation complex. Meanwhile, the mRNA is activated for translation by the binding of eIF4F complex (consisting of eIF4E, eIF4G, and eIF4A) to the 5' cap structure and the poly(A)-binding protein (PABP) at the 3' poly(A) tail, circularising the mRNA and facilitating ribosome recruitment. Then the 43S complex is recruited onto the 5' end of the mRNA and initiate the scanning for the start codon, forming the 48S complex. Once the AUG codon is recognised, several initiation factors are released, and the 60S ribosomal subunit joins the 40S subunit, forming the 80S initiation complex, ready for elongation.

balance between cell proliferation and apoptosis (Schulz et al. 2018). These mutations can create new uORFs or disrupt existing ones, altering the translation efficiency of critical genes involved in cancer development. For instance, a novel uORF can be induced via a point mutation in the tumour suppressor gene CDKN2A, increasing the risk of hereditary melanoma (Schuster and Hsieh 2019). Meanwhile, Schulz et al. have identified mutations in uORF of EPHB1 and MAP2K6 gene from colon adenocarcinoma lead to an enhanced translation of downstream main ORF (Schulz et al. 2018). These observations suggest that uORF mutations directly influence specific cancer phenotypes. As demonstrated above, this mutation abolishes the uORF and leads to the upregulation of these kinases, thereby driving unchecked cellular proliferation and survival, which further results in tumorigenesis.

1.4 Computational approaches

The development of high-throughput sequencing technology has led to the rapid accumulation of genomics and transcriptomics data. Data clusters for UTR research are available in multiple databases including Ensembl and UCSC, however, these UTR databases are currently poorly annotated and not regularly updated (Z. Huang and Teeling 2017; Steri et al. 2018). Moreover, there is currently no comprehensive uORF database and no definitive pipeline for reliable uORF annotation. To analyse these extensive sequencing datasets accurately, bioinformatics tools have become essential, and machine learning algorithms such as Support Vector Machines (SVM) and Random Forest (RF) have been increasingly used in sequence function prediction (Hanyu Zhang

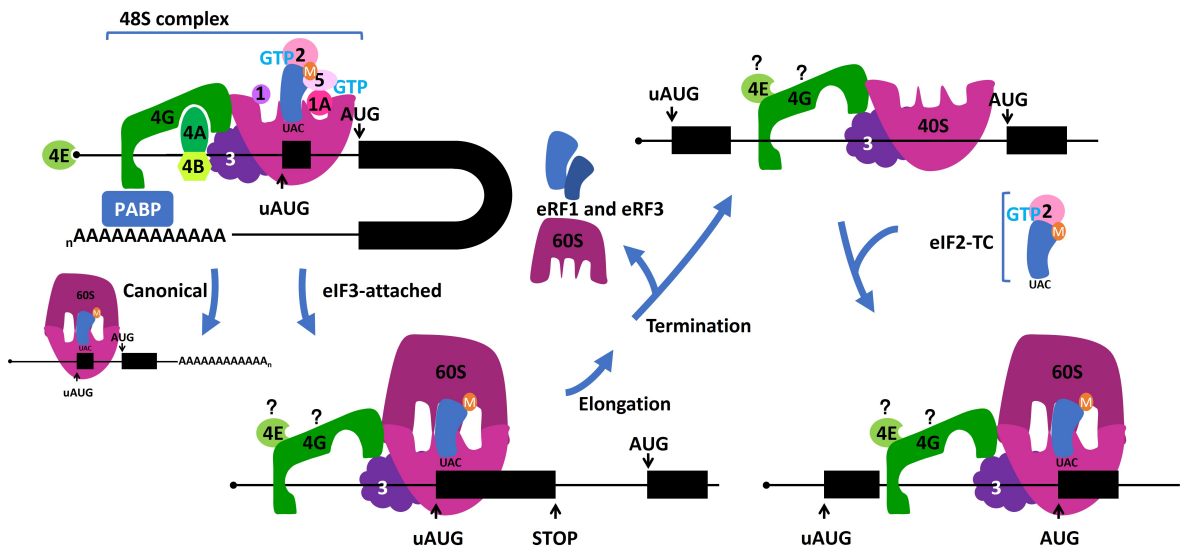


Figure 4: **Schematic diagram of potential translation reinitiation in eukaryotic cell.** Reinitiation occurs when a ribosome resumes scanning and starts translation after terminating translation at a previous stop codon. Unlike the canonical translation initiation process, there is evidence suggesting that some eukaryotic initiation factors, including eIF 3 and potentially eIF 4E and 4G (the question mark suggested 4E and 4G remains hypothesised which need further experiments evidences), may remain attached on the ribosome 40S subunit while forming the 80S ribosomal initiation complex. Then after the eukaryotic release factor (eRF) 1 and 3 terminate the translation and dissociate the ribosome 60S subunit, the 40S may still attached onto the mRNA and continue scanning. The with the success recruitment of new eukaryotic initiation factor 2 ternary complex (eIF2-TC), consisting of eIF2 α , β , γ , GTP and initiator tRNA, the translation started the reinitiation.

et al. 2019). In this section, I will briefly review the computational applications used for ORF detection.

1.4.1 Ribosome profiling and alignment tools

Ribosome profiling (Ribo-Seq) is a high-throughput deep sequencing technique that bridges the gap between the transcriptome (measured by standard RNA-Seq) and the proteome (Figure 5). The idea behind Ribo-Seq is simple and elegant, as it is the sequencing of ribosome protected fragments (RPF), which is fundamentally RNA-Seq with a modified library construction workflow.

Compared to a standard RNA-Seq workflow, an extra step is added to arrest translation prior to lysis using inhibitors such as cycloheximide or via flash-freezing, preventing ribosome dissociation. The fundamental deviation from RNA-Seq then occurs at the fragmentation stage. In a typical RNA-Seq protocol, total mRNA is extracted and randomly fragmented. In contrast, Ribo-Seq replaces this with a nuclease digestion step (Nicholas T. Ingolia 2016). During translation, the ribosome covers a footprint of typically 28 to 32 nucleotides; however, recent studies have confirmed that larger fragments (~ 60 nt) can occasionally be observed when protected by di-ribosomes (Nicholas T. Ingolia 2016; Nicholas T Ingolia et al. 2012).

Deep sequencing of these RPFs produces millions of short reads that indicate ribosome positions on transcripts at the moment of lysis. By mapping RPFs back to the transcriptome, researchers can determine which mRNAs are being translated and where the ribosomes were bound. While Ribo-Seq offers valuable insights into the connection between the transcriptome and the proteome, it remains an evolving method, and the raw sequencing

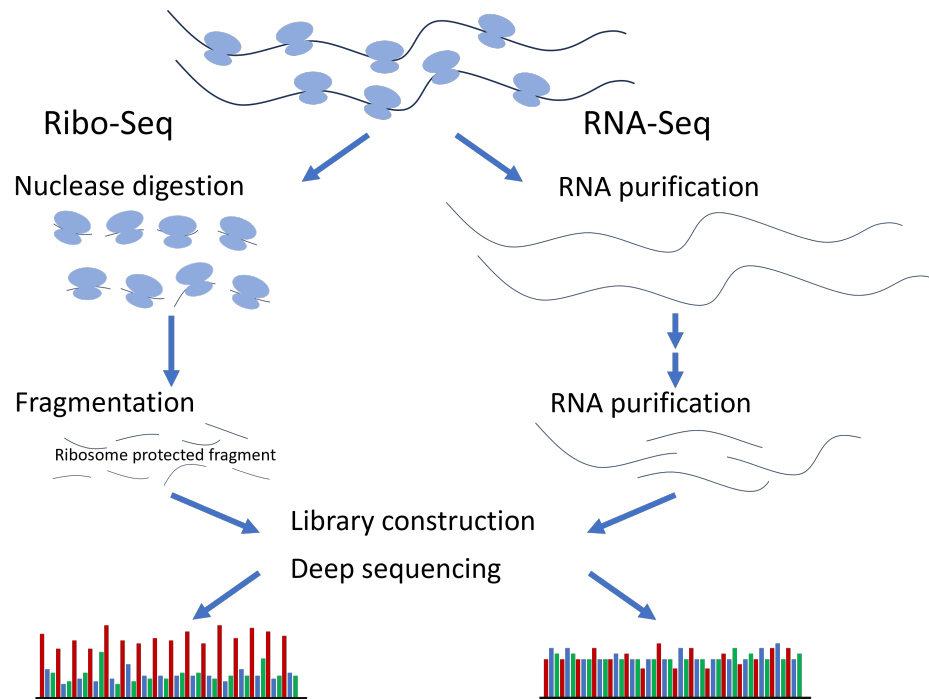


Figure 5: **Overview of Ribo-Seq and RNA-Seq.** Ribo-Seq is fundamentally similar to RNA-Seq, with the primary difference being in the library preparation. In Ribo-Seq, the library is constructed from ribosome protected mRNA fragments. The Ribo-Seq workflow involves isolating mRNA from cell lysates, followed by nuclease digestion to degrade unprotected mRNA, leaving only the ribosome-protected fragments. These fragments are then purified, and a sequencing library is constructed. The final step is then deep sequencing.

data may exhibit more noise compared to well-established RNA-Seq data (Andreev, O'Connor, et al. 2017).
Currently, Ribo-Seq data processing workflows heavily rely on adaptations from RNA sequencing workflows.
The short length of the reads makes accurate mapping difficult.

Three major mapping tools are currently employed in the field of ribosome profiling: BBMap, HISAT2, and STAR (Wagner, Bohlen, et al. 2022; Patraquim et al. 2020; Nicholas T. Ingolia, Lareau, and Weissman 2011).
And, there is still debate on the best practices for Ribo-Seq, as the major aligners are designed for mapping longer reads (normally > 30 nt) (Ziemann, Kaspi, and El-Osta 2016). Roughly speaking, the design of an aligner was aimed at both accuracy and efficiency. Because the genome size can be huge, the human genome, for example, is around 3 billion base pairs long, and each sequencing experiment produces millions of reads, it is impossible to do the pairwise alignment between reads and the genome as was done manually before computational methods were introduced. Modern aligners address the challenges by indexing the reference genome and using seeding-and-extension strategies. This involves splitting the query sequences into short seeds, locating potential mapping sites using the reference index, and then performing detailed pairwise alignments (Donato et al. 2021; Ziemann, Kaspi, and El-Osta 2016).

In this way, the requirement for computing resources is massively reduced. The core step of the aligner is the reference indexing, and it is this step that distinguishes each tool's performance and accuracy. The major reference indexing methods can be categorised into hash table based and suffix array based (Donato et al. 2021). Suffix array-based indexing can be further divided into uncompressed suffix arrays based and Burrows-Wheeler Transform based (BWT). BBMap, STAR, and HISAT2 then represent these three methods, respectively, with

448 BBMap using hash table based indexing, STAR using uncompressed suffix arrays, and HISAT2 employing
449 BWT based indexing (Kim, Langmead, and Steven L Salzberg 2015; Dobin et al. 2013; Bushnell, Rood, and
450 Singer 2017).

451 The hash table-based indexing in BBMap utilises hash functions to generate a hash table of the k-mer subse-
452 quences of a genome. A hash function is a mathematical algorithm that transforms the input (k-mer sequences
453 here) into a fixed-size string of bytes. The output is unique to each distinct input, meaning that even one base
454 change in the input will result in a drastically different hash value. Using the hash table then allows rapid
455 lookup for the sequencing read seeds. However, because hash tables compare exact matches, they may struggle
456 to handle reads that span exon-exon junctions. The uncompressed suffix array indexing in STAR, as stated in
457 its name, stores all the suffixes of the reference genome in lexicographical order, allowing fast binary search-
458 based matching of the read seeds. This method provides quick access to the matching regions but comes with
459 a substantial cost in memory requirements. The BWT method, originally introduced as a data compression
460 method that reorganises strings by grouping repeated substrings, allows effective compression and speeds up
461 pattern matching of read seeds. Using BWT-based indexing algorithms can address the limitation of high mem-
462 ory usage in the uncompressed suffix array, but they may introduce slight reductions in accuracy due to their
463 reliance on compression.

464 1.4.2 ORF identification tools

465 Identifying translated ORFs from Ribo-Seq data, especially short ORFs like uORFs, requires specialised com-
466 putational tools. Over the past decade, numerous of ORF calling algorithms have been developed, each with
467 its own strategy and strengths. These algorithms scan the Ribo-Seq read for signatures of active translation,
468 and a common strategy is to identify the three-nucleotide periodicity: by identifying regions where Ribo-Seq
469 reads are enriched in one reading frame, the algorithm can then predict the presence of a translating ORF. For
470 example, RiboTaper was one of the early methods that used the three-nucleotide periodicity of Ribo-Seq reads
471 to statistically define translated ORFs. It applies a spectral test across reading frames to distinguish genuine
472 signals from noise (Calviello, Mukherjee, et al. 2016). RiboCode similarly scans for regions with a triplet
473 periodicity of based on the Ribo-Seq data and was designed to detect even small ORFs by integrating evidence
474 across replicate experiments (Z. Xiao et al. 2018). RibORF uses a machine learning based approach to classify
475 ORFs as translated or not according to features like read periodicity and uniformity (Ji 2018).

476 Despite these developments, computational challenges remain in ORF calling, which motivated me to develop
477 the new workflow presented in this thesis. First, the heterogeneity of ribosome footprint data (varying read
478 lengths, sequence biases, and experimental noise) can hinder the reliable detection of short ORFs. False posi-
479 tives can arise from random noise that coincidentally displays a periodic pattern, whereas false negatives occur
480 when low level translating ORFs fail to produce a clear periodic pattern. Meanwhile, when a uORF overlaps
481 with the main coding sequence (in a different reading frame), their ribosome footprints intermix, thereby sab-
482 otaging the triplet periodicity. In such cases, conventional tools may struggle to decide whether reads belong

to a uORF or the main ORF, potentially missing one or the other. Furthermore, existing ORF identification tools typically adopt their own methods for defining and classifying ORFs. For example, Ribotaper annotates ORFs initiating in the 5' UTR but overlapping with the CDS as 'non-CCDS', whereas RiboCode combines ORFs starting within the CDS and ending in the 3' UTR into a single category called 'overlapping ORFs'. This lack of standardisation complicates cross-study comparisons and meta-analyses, as researchers cannot easily compare ORF catalogues from different studies if the underlying algorithms employ distinct ORF definitions. Fortunately, efforts are currently underway to establish standardised annotations for translated open reading frames, and I have incorporated these proposed standards into the development of my own workflow (Mudge et al. 2022). Another significant issue is that many pipelines have been developed and benchmarked on relatively high-coverage Ribo-Seq datasets, and their performance may deteriorate with lower sequencing depth or noisier clinical samples which is more common if applied to future clinical usage.

Here, I introduce two state-of-the-art ORF prediction tools: ORFquant (developed based on Ribotaper) and Ribotricer which I applied in the rest of my research for comparison with my proposed workflow. To ensure a clear understanding of how decisions were made, no deep learning algorithms were considered. I selected these two tools after reviewing the benchmarking of currently established tools alongside the relevant literature (Gelhausen et al. 2022; Calviello, Hirsekorn, and Ohler 2020; Calviello and Ohler 2017).

ORFquant

ORFquant is a bioinformatic method developed to annotate and quantify translation at the level of single ORFs using ribosome profiling data (Calviello, Hirsekorn, and Ohler 2020). ORFquant is specifically designed to address the challenges associated with identifying and quantifying translated ORFs across multiple transcript isoforms of genes, providing insights into the complexity of the ORFs. Their workflow consists of three major steps, transcript filtering, ORF identification and ORF quantification.

ORFquant begins by analysing the Ribo-Seq data to extract single-nucleotide positions corresponding to the ribosome's peptidyl site (P-site positions) and junction reads from the Ribo-Seq alignment. For the P-site position identification, ORFquant requests user-defined offset values for the Ribo-Seq reads. Then these reads are mapped onto the gene annotations. ORFquant filters the transcripts to retain a subset that sufficiently explains the observed P-site while minimising the occurrence of exons and junctions with no Ribo-Seq signal. This filtering applied a greedy algorithm that calculates 'positive features' (exons or junctions containing at least one mapped Ribo-Seq read) and 'negative internal features' (regions with no reads located between the first and last positive features of a transcript). In this context, a transcript candidate is discarded if all of its positive features are fully contained within another transcript isoform that either possesses more total positive features or fewer negative internal features. This filtering is based on an "Occam's razor" strategy, which is a philosophical concept where the fundamental statement is "Entities should not be multiplied beyond necessity" (Domingos 1999).

After transcript filtering, ORFquant identifies translated ORFs within the selected transcripts. MultiTaper

Method (MTM) was then used to select in-frame signal with triplet periodicity, as the hallmark of active translation elongation. MTM is a time-frequency decomposition technique developed specifically to analyse non-stationary and potentially noisy signals, such as Ribo-Seq data. While avoiding an in-depth discussion of the mathematical intricacies, MTM essentially involves averaging multiple Fourier transforms that utilise different tapers. This approach effectively reduces the variance of the spectral estimate, resulting in a more accurate and reliable representation of the periodicity present in the sequencing data (Thomson 1982). ORFquant only considers ATG as a potential start codon, and ORFs are selected based on their P-site signal in the reading frame, ensuring that the average in-frame signal on covered codons is greater than 50%.

Once ORFs are identified, ORFquant quantifies them by calculating coverage on unique and shared ORF features (such as exonic bins and splice junctions). A scaling factor is determined based on the coverage on unique ORF features or the amount of overlap between ORFs. This scaling factor represents the fraction of Ribo-seq signal that can be attributed to a specific ORF. The number of P sites is then normalised by ORF length to yield ORFs per million, a metric similar to transcripts per million (TPM) used in RNA-seq.

Ribotracer

Ribotracer is designed for the accurate detection of actively translating open reading frames (ORFs) using ribosome profiling (Ribo-seq) data (Choudhary, Li, and D. Smith 2020). In their workflow four major steps were applied, ORF candidate identification, Ribo-Seq offset calculation, periodicity assessment and phase score calculation.

Ribotracer begins by identifying all possible ORFs within annotated transcription units. This is done by searching for pairs of start and stop codons across the genome within these annotated genes. This step generates a comprehensive set of candidate ORFs, which can range from very short ORFs (less than 100 codons) to longer ones. The method accounts for eight types of ORFs, based on the ORF start and stop codon positions relative to the main coding region. After identifying candidate ORFs, Ribotracer groups the Ribo-seq reads based on their length. This step is crucial because different read lengths may correspond to different P-site offsets. For each read length, Ribotracer generates a metagene profile using the 5' ends of the mapped reads. This profile is used to infer the P-site offsets by maximising the cross-correlation with the profile of the most abundant read length. The inferred offsets allow Ribotracer to merge the read profiles from different lengths into a single, coherent ribosome reads profile for each ORF.

The core of the Ribotracer algorithm is, similar to ORFquant, the assessment of the triplet periodicity in the Ribo-Seq profile. Ribotracer introduces a novel method for this by projecting a 3D vector which is the absolute read counts at each nucleotide representing the codon into an abstract 2D unit vector representing the p-site alignment in certain nucleotide of those three bases of a codon. This transformation focuses on the qualitative pattern of read distribution across the three frames of a codon (i.e., the 'high-low-low' pattern expected from active translation) rather than the absolute read counts. By down-weighting the influence of individual read counts and focusing on consistency in the periodicity, the noise introduced by the potential ribosomal pause

553 and non-ribosome binding could be significantly reduced.

554 Finally, Ribotricer calculates a phase score for each ORF based on the consistency of the calculated periodicity
555 across the ORF. This score is used to distinguish actively translating ORFs from non-translating regions. The
556 phase score is robust against variations in read coverage and is not biased by the length of the ORF.

557 1.4.3 Machine learning

558 To investigate whether ORFs have distinct feature patterns across different gene sets, or if they can be clustered
559 based on a feature matrix which comprise structural, sequence, and conservation metrics, I trained several
560 classifiers using ORF sequence patterns and annotations extracted from my workflow. With the Ribo-Seq and
561 RNA-Seq data as well as annotations information acquired from other databases such as conservation score
562 from PhastCons (a statistical model grounded in a phylogenetic hidden Markov model) (Siepel et al. 2005), a
563 high dimensional feature database was constructed to describe tens of thousands of ORFs. In this section, I will
564 introduce the sequence pattern extraction algorithm and the machine learning models used.

565 Chaos game representation

566 The Chaos Game Representation (CGR) is a symbolic sequence representation method based on an iterated
567 function system (Löchel et al. 2020). It is a 2D fractal representation of a 1D sequence, constructed as follows:
568 The CGR space here is a square plot with each corner representing the nucleotides A, C, G, T. Each point
569 represents a base in the sequence. The first base is plotted halfway between the centre of the square and the
570 corner representing the first base. Every successive base is then plotted in the middle of the previous base and
571 the corresponding corner. I applied CGR in this study because it provides an alignment free, fixed resolution
572 embedding that summarises k-mer composition and higher order dependencies across ORF sequences of vari-
573 able length. This allows both global base composition and local periodicities (such as the 3-nt structure) to be
574 fed into downstream classifiers, such as support vector machines (SVMs) and random forests (RFs), enabling
575 them to exploit ORF sequence patterns without hand-crafted motif rules. The general concept of CGR used in
576 this study is illustrated in Figure 6 below. This can be represented numerically as follows:

577 Let $\mathbf{P}_i$ be a position in a sequence, with $\mathbf{P}_0 = (0.5, 0.5)$ as the initial position. The position $\mathbf{P}_i$ updates
578 according to the formula:

$$\mathbf{P}_i = \mathbf{P}_{i-1} + \frac{\mathbf{P}_{i-1} - \mathbf{N}_i}{2} \quad (1.1)$$

579 where $\mathbf{N}_i$ is a point from the set $\{\mathbf{A} = (1, 0), \mathbf{T} = (0, 1), \mathbf{C} = (1, 1), \mathbf{G} = (0, 0)\}$, representing different
580 directions or states.

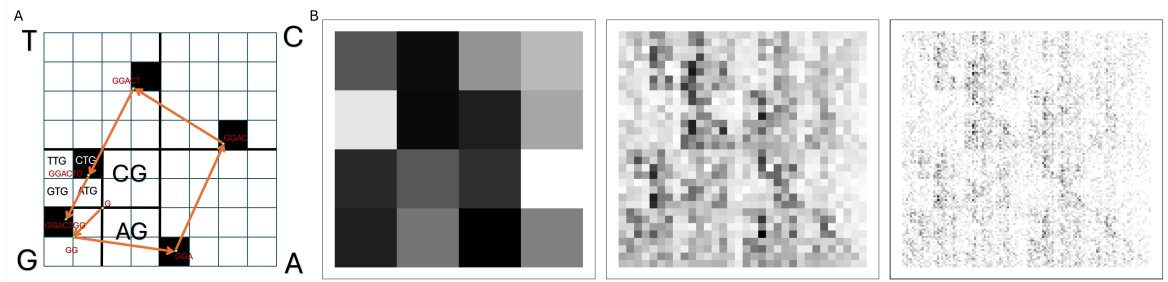


Figure 6: **Chaos game representation.** This set of plots demonstrated how chaos game representation works. A) illustrates the spatial representation of DNA sequences on a coordinate grid. The specific arrangement of points reveals the sequence patterns. B) visualises the sequence complexity and recurring motifs as resolution increases. The resolution corresponds to the k-mer size that each square represents, progressing from 2-mer in the first plot to 8-mer in the last plot, showing finer sequence details at higher resolutions.

Support vector machine

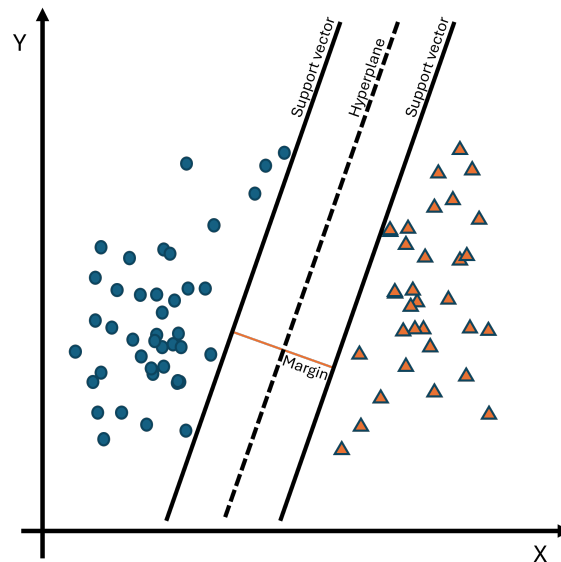


Figure 7: **Support vector machine.** This figure represents the SVM in a binary classification process. The blue circles and orange triangles represent two distinct classes in a 2D feature space. The black dashed line indicates the hyperplane that SVM constructed to maximally separate two classes (maxing the margin). The points closest to the hyperplane then denote the support vectors.

A support vector machine (SVM) is a supervised machine learning algorithm for binary classification. The idea behind the algorithm is to find the optimal hyperplane that separates data points belonging to different classes with the maximum margin (Figure 7). The margin is defined as the distance between the hyperplane and the closest data points from each class, known as support vectors. These support vectors are critical to the model, as they determine the decision boundary. However, when working with data in high dimensional spaces, the relationship between features might be complex and the data might not be linearly separable. In such cases, SVM employs kernel functions, which are essentially mathematical transformations, to convert the data into a higher dimensional space where linear separation might be possible. As shown in Figure 8, the kernel function $\phi(x) = x \bmod 2$ maps the original one-dimensional dataset to a transformed space where a linear decision boundary can be easily identified.

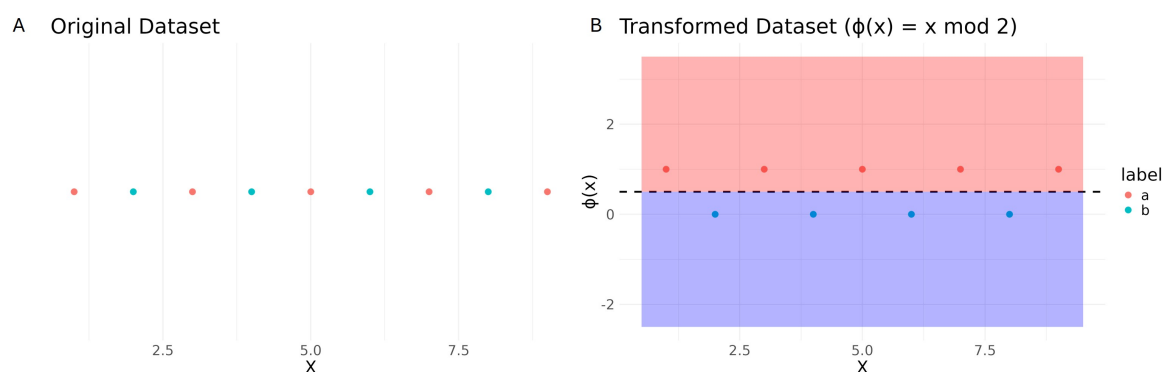


Figure 8: **Illustration of kernel function in support vector machine (SVM).** This set of plots illustrates how the kernel function facilitates the separation of data in SVM. The dataset represents a one dimensional vector from a high dimensional space. A) represents the original dataset: the data points are plotted along the X-axis with two distinct labels, 'a' in red and 'b' in blue. It is clear that it is linearly inseparable in this space. B) is the transformed dataset that a kernel function $\phi(x) = x \bmod 2$ is applied to map the data into a higher-dimensional space where a linear separation can be achieved.

SVM was applied for the analysis of the sequence pattern future matrix generated using CGR methods in Chapter 5.1. Due to its high accuracy and strong theoretical foundation for avoiding overfitting. It is particularly well-suited for my problems where the sequence pattern datasets is small and with high feature dimensions.

Random forest

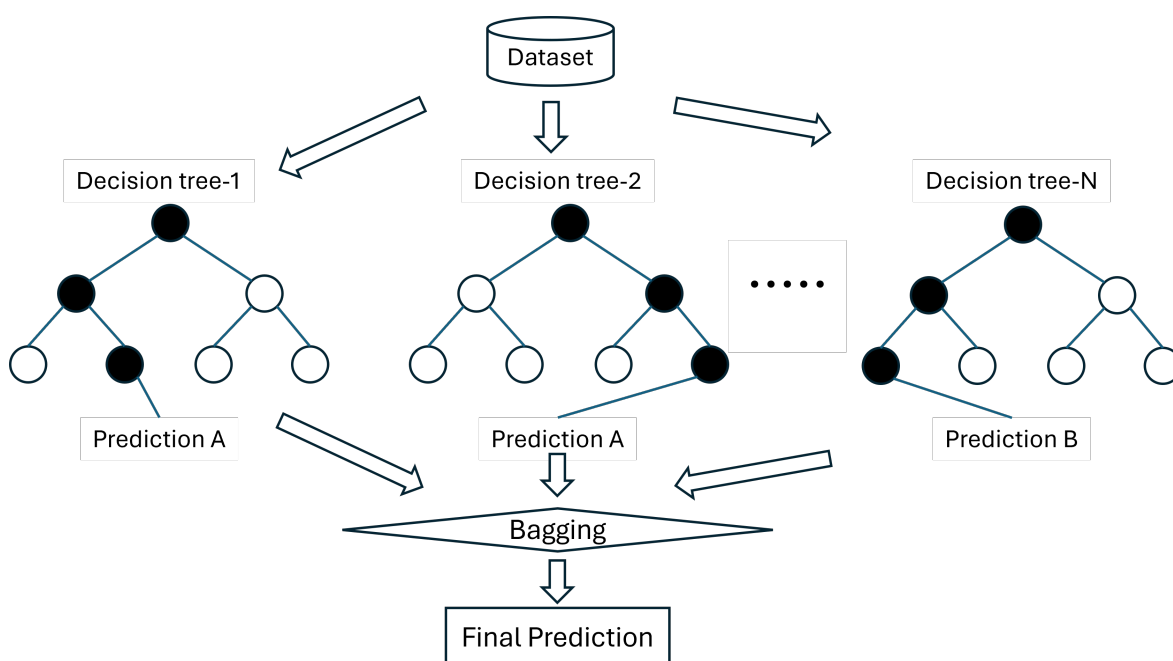


Figure 9: **Random forest workflow.** This figure illustrates the process of constructing a Random Forest model. The dataset is used to train multiple CART trees (Classification And Regression Tree). Each tree is trained on a random subset of the dataset and independently makes a prediction. The predictions are then aggregated through a process known as bagging (Bootstrap Aggregating), which combines the predictions from all decision trees to produce a final prediction.

Random forest (RF), introduced by Breiman, is an ensemble learning method which combines multiple learners to enhance the overall model performance (Figure 9) (2001). The learners in RF are Classification And

Regression Trees (CART tree). Each decision tree in RF is constructed independently, but with two levels of randomness: random sampling of the data (bootstrap sampling) and random selection of features at each node of the decision tree. This ensures that each decision tree in the forest is trained on a slightly different version of the dataset and reduces the risk of overfitting by ensuring that the trees do not all rely on the same dominant features, increasing diversity among them. This dual-layer randomness is key to RF's ability to generalise well across diverse datasets. Once all decision trees are built, RF aggregates their predictions (bagging) and the final prediction is determined through majority voting. By combining the outputs of multiple trees, RF achieves better performance than a single decision tree, as the combined prediction smooths out errors that individual trees may make, thereby reducing variance.

The theoretical foundations of RF include principles like the law of large numbers and the bias-variance trade-off. As the number of trees grows, the law of large numbers suggests that the aggregated predictions of the forest will stabilise and become more reliable. Meanwhile, the bagging technique that RF employs helps to mitigate the effects of variance, enhancing the model's robustness without greatly impacting bias.

RF was used in Chapter 5.6 for distinguishing the experiment-proven uORF that is vital for cell survival. The reason for choosing RF is that it can effortlessly handle interactions between features and is non-parametric, so one does not have to worry about outliers or whether the data is linearly separable. And it can manage irrelevant features. For example, consider two features, a conservation score and the minimum free energy (MFE) of local RNA structure. A decision tree can easily handle a scenario where $\text{conservation} \leq c_1 \Rightarrow \text{type A}$; $c_1 < \text{conservation} < c_2$ and MFE in an intermediate band $\Rightarrow \text{type B}$; then MFE at extreme values $\Rightarrow \text{type A}$ again. This allows the model to represent non-linear, even disjoint, class regions, reflecting a pattern in which one ORF type occurs at low conservation, a second type dominates in the middle band, and the first type reappears at the extremes of MFE.

In Chapter 5.6, RF was chosen for its ability to handle feature interactions and its non-parametric nature, which means it does not require assumptions about the data distribution or linearity. It effectively manages irrelevant features as well, allowing it to distinguish between experimentally verified uORFs critical for cell survival. For instance, RF can seamlessly navigate scenarios where different ORF types show distinct patterns across conservation and minimum free energy feature dimensions.

To evaluate feature importance in RF, two measures are commonly used: mean decrease in Gini (MDG) and permutation importance (mean decrease in accuracy, MDA). MDG summarises how much a feature reduces the Gini impurity at the splits where it is used; these reductions are accumulated and averaged across all trees, yielding a ranking by contribution to node purity. Gini impurity measures how mixed the classes are within a node; lower impurity means a purer, more homogeneous node. MDA estimates a feature's predictive contribution by randomly permuting its values in hold out (out-of-bag) data and measuring the resulting drop in model accuracy; a larger drop indicates greater importance. MDG is fast and built-in but can be biased towards variables with many potential split points, whereas MDA is directly tied to predictive performance but can be reduced when features are highly correlated. In this work, importance scores are interpreted with these caveats

634 in mind.

635 1.5 Summary

636 This introduction has outlined the critical role of upstream open reading frames (uORFs) as key regulatory
637 elements that modulate gene expression through the control of translation. uORFs are prevalent throughout
638 the genome and play significant roles in developmental regulation and potential tumorigenesis. Recent ad-
639 vancements in ribosome profiling have uncovered the hidden complexities of uORFs, offering unprecedented
640 insights into their functional significance. However, current ORF prediction algorithms often face challenges in
641 accurately identifying these small, noncanonical uORFs, highlighting the need for more robust computational
642 approaches tailored specifically for their precise detection and quantification.

643 To address these challenges, the central goal of this thesis is to test the hypothesis that the fundamental evidence
644 of ORF translation is primarily dependent on the agreement of transcription and ribosome occupancy, instead
645 of the strict complex periodicity signals. Existing tools often depend on rigid periodic patterns which can
646 be difficult to detect in short, low-abundance uORFs. This research moves away from initial assumptions,
647 proposing a two-stage framework which detects then curates. This approach prioritises high sensitivity to
648 capture a broad spectrum of potential translation events first, followed by a data driven characterisation using
649 orthogonal biological features to distinguish functional elements.

650 This thesis addresses these challenges by developing an novo ORF identification workflow aimed at a better
651 detection of translating ORFs (especially uORFs) based on Ribo-Seq and RNA-Seq data. The goal is to inte-
652 grate strengths from existing methodologies and introduce new analytical steps to overcome current limitations
653 (such as to reduce noise without discarding genuine uORF signals). By doing so, I aimed to provide a clearer,
654 more complete picture of the proteomes. And two major objectives are as follows:

655 Objective 1: To design and implement a translating ORF identification workflow that addresses gaps in the
656 current state-of-the-art. This new workflow is intended to improve the sensitivity of detecting upstream and
657 non-canonical ORFs by leveraging data-driven features. In Chapter 3, I have detailed the development of my
658 proposed approach and provided a snapshot of the workflow implementation on a demonstration dataset.

659 Objective 2: To apply the new workflow to ribosome profiling datasets in order to characterise the landscape of
660 translation in distinct biological contexts, and to investigate the regulatory roles of uORFs in those contexts. In
661 Chapter 4, the workflow is employed on two case studies, a mouse study using our lab data and a human study
662 using the published datasource. Through these applications, I demonstrate the utility of the proposed approach
663 for real biological data and draw insights into how uORFs modulate gene expression in different experimental
664 settings.

665 By accomplishing these objectives, this research will enhance our ability to decode the translation landscape,
666 particularly the hidden proteomes encoded by uORFs, providing a foundation for future studies to explore the

⁶⁶⁷ functional significance of the newly discovered peptides and regulatory elements.

Chapter 2

Methods

My PhD project involves developing a computational workflow to identify and characterise uORFs using Ribo-Seq and RNA-Seq data, which will be detailed in the following results chapters. Here in this chapter, I will go through the external software and algorithms I have applied in my workflow as well as the preparation protocols of the datasets I used from both our lab and online publications.

2.1 Data sources

2.1.1 MNK mTOR mouse sequencing

The mouse sequencing data was obtained from Dr Joseph Waldron in the Bushell lab. The experiments were designed to explore the role of mitogen-activated protein kinase-interacting kinases (MNK) and mammalian target of rapamycin (mTOR) in the KRAS signalling pathway. These experiments included six biological replicates of APC^{-/-} KRAS^{G12D} mouse colon organoids, which were divided into three treatment groups and a control group. This specific genotype is characterised by elevated eIF2 α phosphorylation compared to APC^{-/-} KRAS^{WT}. These organoids, to some extent, simulate a basal activation of the Integrated Stress Response (ISR) node makes this organoid model sensitive to perturbations in the translational machinery, thereby making it ideal for investigation of translation regulatory mechanisms (Adomavicius et al. 2019). All organoids were derived from female mice. Due to library preparation failures, replicates 1, 2, and 5 were ultimately excluded from analysis. The treatment groups consisted of Rapa (rapamycin), EFT508 (MNK1/2 inhibitor), and Rapa&EFT508 (combined treatment with both drugs). The setup of inhibitor combination was grounded in the therapeutic synergy described by Knight et al (2021). They have demonstrated that while KRAS mutant colorectal cancer often show resistance to monotherapy with mTOR inhibition, additional MNK inhibition could sensitise these cells to mTOR blockade by preventing the phosphorylation of eIF4E and reducing the translation of specific oncogenic drivers such as c-MYC. Organoids were treated with eFT508 (30 nmol/L) and

rapamycin (250 nmol/L) for 6 hours prior to harvesting. The organoid cells were then collected by scraping, rinsed with PBS, pelleted by centrifugation, snap frozen in liquid N₂, and stored at -80°C for future downstream experiments. Each biological replicate was conducted using a six-well plate, with six individual organoid plugs plated in each well. One hour before the lysis, cells were moved to -20°C and then placed on ice as well as the lysis buffer (15 mM Tris-HCl pH 7.5, 15 mM MgCl₂, 300 mM NaCl, 1% Triton X100, 0.05% Tween 20, 2% n-Dodecyl β -Maltoside (Thermo Fisher, 89903), 0.5 mM DTT, 100 μ g/ml cycloheximide, 1X cOmplete EDTA-free Protease Inhibitor Cocktail (Roche, 11836170001), 200 U/ml RiboLock RNase Inhibitor (Thermo Fisher Scientific, EO0381)). For each cell pellet, 500 μ l lysis buffer was added and incubated on ice for 5 mins followed by centrifugation. For Total RNA-Seq, 25 μ l aliquots were taken from the lysate and added to 975 μ l TRIzol for RNA extraction based on the manufacturer's instructions. Cytoplasmic RNA integrity was then assessed using an Agilent TapeStation, and RNA concentration together with sample purity were determined using a NanoDrop spectrophotometer. Ribosomal RNA was removed from 1 μ g of cytoplasmic RNA using the RiboCop v2 kit, and sequencing libraries were generated from the rRNA-depleted RNA with the Corall Total RNA-Seq Library Prep Kit v1 and further employed for 13 PCR amplification cycles. Then for Ribo-Seq, the remaining lysate were treated with 10 μ l Ambion RNase I at 22°C for 15 mins and the digestion was terminated by adding 20 μ l SUPERaseIn RNase Inhibitor (20 U/ μ l). The resulting material was then layered onto a 10–50 % sucrose gradient containing 15 mM Tris-HCl (pH 7.5), 15 mM MgCl₂, 300 mM NaCl, and 100 μ g/ml cycloheximide, prepared using a BioComp gradient station and equilibrated at 4°C for at least 1 h before use. And then centrifuged at 38,000 rpm for 2 h at 4°C. Gradients were fractionated in 1 ml increments using a BioComp gradient station with a Gilson FC 203B fraction collector. Fractions corresponding to the 80S ribosome peak were pooled, and RNA was extracted using acid phenol chloroform, followed by two chloroform washes. Ribosome-protected fragments (28–34 nt) were size-selected on a 15 % TBE-Urea gel. Unlike for total RNA preparation, ribosome-protected RNA was processed without prior rRNA depletion. Samples were heated to 80°C for 90s, then rapidly chilled on ice before undergoing 5' phosphorylation and 3' dephosphorylation using 1 μ l T4 PNK (NEB, M0201S), 5 μ l 10 $\times$ PNK buffer, and 1 μ l SUPERaseIn RNase Inhibitor (20U/ μ l) at 37°C for 35min. For the final 20min of the reaction, 5 μ l 10mM ATP was added. RNA was subsequently purified by acid phenol chloroform extraction and precipitated with isopropanol as described above. Small RNA libraries were then prepared according to the NEXTFLEX Small RNA-Seq Kit v3 (Inc. 2019) with Dr Joseph Waldron's modification. In short, eight steps were undertaken, I will highlight the key information and the enhancement that Dr Joseph Waldron undertook, but not include all the details.

(A) NEXTFLEX 3'4N Adenylated Adapter Ligation RPF RNA (1ng–2 μ g) was denatured and then immediately chilled on ice. The NEXTFLEX 3'4N adenylated adapter (full amount without dilution) was ligated to the 3' end of RNA using a ligation buffer and enzyme mix. The reaction was incubated at 25°C for 2 hours. This step adds a 3' adapter with four random bases (4N) to reduce ligation bias.

725 **(B) Excess 3' Adapter Removal** Unligated 3' adapters were immediately removed using the NEXTFLEX
726 Adapter Depletion Solution after the ligation and followed by a bead-based cleanup. Two steps were high-
727 lighted. Firstly, the ethanol incubation of the bead was carefully monitored to avoid over incubation (no longer
728 than 3 min). As suggested by Dr Joseph Waldron, an automatic pipette and multichannel were applied to this
729 setup as more than 4 samples were used. Then, during the final wash with Nuclease-free Water, 13.5 μ l was
730 used.

731 **(C) Excess Adapter Inactivation** Remaining active adapter molecules were enzymatically inactivated to
732 prevent cross-ligation in subsequent steps. To achieve this, NEXTFLEX Adapter Inactivation Buffer (2 μ l) and
733 Enzyme (0.5 μ l) were added into the purified ligated RNA (11.5 μ l) from the previous step. Then incubated
734 sequentially at 12°C and 50°C for 15 mins and 20 mins.

735 **(D) NEXTFLEX 5'4N Adapter Ligation** Immediately after the previous step, similar but slightly different
736 from step A, the 5' adapter with 4 random bases were added. The pre-heated NEXTFLEX 5'4N adapter (1.5 μ l)
737 was ligated using a NEXTFLEX5' Ligation Buffer (7.5 μ l) and NEXTFLEX5' Ligation Enzyme mix (2 μ l),
738 incubated at 20°C for 1 hour. This step completes the addition of sequencing-compatible adapters with random
739 bases at both ends as the UMI, which will be used for deduplication and then trimmed during the data analysis.

740 **(E) Reverse Transcription – First Strand Synthesis** The ligated RNA (25 μ l) was reverse-transcribed to
741 generate cDNA using M-MuLV Reverse Transcriptase (2 μ l) and NEXTFLEX RT Buffer (13 μ l). Reactions
742 were incubated at 42°C for 30 minutes followed by 90°C for 10 minutes to inactivate the enzyme.

743 **(F) Bead Cleanup for larger RNA** Following first-strand synthesis, cDNA (40 μ l) was subjected to bead
744 cleanup using the alternative step F bead clean up. 10 μ l of Adapter Depletion Solution was added to each
745 sample, followed by the addition of 40 μ l NEXTFLEX Cleanup Beads and 90 μ l isopropanol. Samples were
746 incubated for 5min at room temperature, then until the solution cleared, and the supernatant discarded. Beads
747 were washed twice with 180 μ l freshly prepared 80% ethanol, ensuring that ethanol incubation did not exceed
748 30s per wash. After a 3min air dry, beads were resuspended in 20 μ l nuclease-free water, incubated for 2min,
749 and magnetised until the solution was clear. A total of 18 μ l of the eluate was transferred to a new tube for PCR
750 amplification.

751 **(G) PCR Amplification** Purified cDNA (18 μ l) was amplified using NEXTFLEX Small RNA PCR Master
752 Mix, Universal Primer, and unique NEXTFLEX Barcoded Primers (for indexing, therefore each sample needs
753 to be treated with different primers). Then the pre-heated (above 80°C) thermocycler was loaded and set for
754 PCR. The amplification cycle was set to 10-11 cycles.

755 **(H) PAGE Size Selection and Cleanup** The libraries were resolved on a 6% TBE-PAGE gel. The gel is
756 prepared with the recipe of 1.5ml AccuGel 40% (19:1), 2ml 5X TBE, 6.5ml miliQ water, 100 μ l APS (25%)
757 and 10 μ l TEMED. Then pour the gel and leave the polyacrylamide to polymerise for more than 30 mins. After
758 the gel has been cast, it is pre-run for 10-15 mins before loading samples. Then load 20 μ l of sample and keep
759 10 μ l for backup. In an adjacent lane, load 4 μ l of Ready to Load Low MW Ladder. After running the gel, bands
760 corresponding to 150bp (representing small RNA inserts plus adapters) were excised, crushed, eluted, and
761 purified using NEXTFLEX Cleanup Beads. Final library quality was assessed using the Agilent Bioanalyzer
762 High Sensitivity DNA assay and quantified by Qubit dsDNA HS Assay.

763 2.1.2 iPSC human experiment

764 The human sequencing data and the list of ORFs which have statistically significant impact on cell viability and
765 proliferation were from Chen et al.'s work (2020). They have conducted a study on using induced pluripotent
766 stem cells (iPSCs) to identify and functionally characterise non-canonical ORFs in the human genome. They
767 applied ribosome profiling to detect actively translated regions outside the main coding region, RNA sequenc-
768 ing to map those regions to transcriptomes, and finally use CRISPR-based functional screening to assess the
769 impact of identified ORF knockouts on cell growth and viability. They have identified 570 non-canonical ORFs
770 including 230 uORFs could affect normal cell growth.

771 Ribo-Seq and RNA-Seq in iPSCs

772 The iPSC used in the Ribo-Seq and RNA-Seq experiments was from Coriell Biorepository (#GM25256). These
773 cells were cultured under feeder-free conditions on growth factor-reduced Matrigel in mTeSR medium. The
774 cells were enzymatically dissociated using Accutase to produce single-cell suspensions for passaging, with the
775 addition of a ROCK inhibitor (Y-27632) to promote survival.

776 For ribosome profiling, samples were prepared for sequencing as described in a recently updated protocol of
777 ribosome profiling (McGlinchey and Nicholas T. Ingolia 2017). In summary, cell were harvested rapidly, and
778 lysed. Lysates were then treated with RNase I to digest RNA not protected by ribosomes, and 80S ribosomes
779 were isolated via ultracentrifugation. Ribosome-protected RNA fragments were purified, size-selected (28-
780 34 bp), ligated with a pre-adenylated DNA linker as the unique molecular identifiers (UMIs), and reverse
781 transcribed. These cDNA libraries were prepared for Illumina sequencing. The library was sequenced on an
782 Illumina HiSeq 4000 sequencer with a single-end 50 base pair run.

783 CRISPR-based functional screening design

784 The CRISPR-based screening was performed to systematically knock out noncanonical ORFs identified by
785 ribosome profiling. For this, the authors designed a Cas9 single-guide RNA (sgRNA) library to target 1098
786 upstream open reading frames (uORFs). Out of these targeted uORFs, the screen identified 230 hits that resulted

787 in statistically significant growth phenotypes. For each ORF, up to 10 sgRNAs were designed, focusing on
788 maximizing on-target activity and minimizing off-target effects. Control sgRNAs targeting noncoding regions
789 upstream of the CDSs and non-targeting sgRNAs were included as controls.

790 The sgRNA-induced knockout phenotypes were assessed by measuring changes in sgRNA abundance over
791 time. At the beginning of the experiment, genomic DNA was extracted, and sgRNA-encoding regions were
792 PCR-amplified and sequenced. After 10 doublings, cells were harvested again, and sgRNA abundance was
793 similarly measured. Then the phenotype score (γ) for each sgRNA was calculated by determining the change
794 in its abundance relative to the control sgRNAs. ORF phenotypes were quantified by averaging the γ values of
795 the three most depleted sgRNAs targeting each ORF. Mann-Whitney U tests were performed to determine the
796 statistical significance of the depletion or enrichment of sgRNAs compared to control sgRNAs. ORFs with a
797 false discovery rate (FDR) of less than 0.05 and a phenotype score below -0.02 were classified as essential for
798 cellular growth.

799 2.2 Software and development environment

800 The workflow BYU was built using R 4.4.0 with the Bioconductor version 3.19, and the packages listed in Table
801 2. The following subsections will introduce the settings used for the major tools used in this workflow. Other
802 parameters remain as the default values. This project was developing in the High Performance Computing
803 cluster inside the Cancer Research UK Scotland Institute.

804 2.2.1 Raw data processing

805 The workflow begins with processing raw FASTQ sequencing data using FastP (0.23.2), which removes arte-
806 facts introduced during sequencing experiments. After cleaning the sequencing data, different aligners are
807 applied and benchmarked to map the reads to the genome. The tools used for alignment include STAR (2.7.3a),
808 BBDMap (38.98), and Hisat2 (2.2.1), each tested for performance and accuracy in handling the sequencing data.

809 FastP

810 Fastp is designed for a fast all-in-one tool box for pre-processing the raw FastQ files (S. Chen et al. 2018). It
811 was used to perform adaptor removal, de-duplication, UMI removal sequentially.

812 Adapter trimming: Automatically detects and trims adapters, which are synthetic sequences added during
813 library preparation that can interfere with accurate read alignment, ensuring clean and reliable sequence data.

814 The `--adaptor_sequence` was taking FASTA file input to trim multiple adapters.

815 De-duplication: Specified `--dedup` to enable duplication rate evaluation, which removes duplicate reads intro-
816 duced during PCR amplification, thus preventing biased results.

817 UMI trimming: UMIs (unique molecular identifiers) are short sequences introduced during library preparation
 818 to uniquely tag each original RNA molecule, allowing precise quantification and accurate differentiation of
 819 PCR duplicates, ensuring that amplification bias does not affect the analysis. Based on QC report and the
 820 UMI design information from the experiment, UMI can be removed via the `--umi_loc=read1 --umi_len=x`
 821 where read1 is for single end sequencing and x is the UMI length.

822 **STAR (v2.7.10a)**

823 For indexing genome: The `--sjdbOverhang` is important in improving the accuracy of spliced alignments.
 824 It specifies the length of the genomic sequence flanking each splice junction that will be used during the
 825 alignment. The optimal value for `--sjdbOverhang` is typically calculated as `ReadLength - 1`. For align-
 826 ment: `--outFilterMultimapNmax 3`, limits the number of loci a read can map to, controlling multi-mapping
 827 reads; `--outFilterMismatchNmax 3`, sets the maximum number of mismatches allowed per read alignment;
 828 `--outSAMprimaryFlag AllBestScore`, marks all alignments with the best score as primary alignments.

829 **Hisat2 (v2.2.1)**

830 For indexing genome:

831 `python extract_splice_sites.py <x_annotation.gtf> > <x_splice_sites.txt>`
 832 extracts the splice site information from the GTF file; hisat2-build then takes `--ss x_splice_sites.txt`
 833 argument allows the incorporation of splice site information, enhancing spliced read alignment. For alignment:
 834 `--known-splicesite-infile x_splice_sites.txt` allows the input of a file containing known splice
 835 sites which generated before, enhancing alignment accuracy around annotated splice junctions.

836 **BBMap (v39.01)**

837 For indexing genome: `k=10` defines the k-mer size for the hash table construction. A smaller k-mer size
 838 increases sensitivity to indels and mismatches but can increase memory usage. For alignment: `ambiguous=all`
 839 refine how BBMap treats multi-mapping reads. This setting retains all top-scoring sites.

840 **2.2.2 ORF identification algorithms**

841 Two ORF identification algorithms were demonstrated in this thesis, as I have introduced in the Introduction. I
 842 will introduce the file structure and the settings used.

843 **ORFquant (v1.02)**

844 ORFquant is a bioinformatic method developed to annotate and quantify translation at the level of single open
 845 reading frames using ribosome profiling (Ribo-seq) data (Calviello, Hirsekorn, and Ohler 2020).

The workflow is implemented in R and involves three key functions. The first one, **prepare_annotation_files**, takes two inputs: a genome sequence file in .2bit format and a gene annotation file in .gtf format. It processes these files to create a comprehensive list of annotated transcripts and genomic features, stored in a .gtf_Rannot file, which is used for downstream ORF annotation. The second function, **prepare_for_ORFquant**, requires the previous .gtf_Rannot file, a sorted Ribo-Seq alignment file in .bam format, and a plain text file detailing read lengths and their corresponding p-site offsets. This function generates a file, **for_ORFquant_file**, prepared for the following ORFquant processing step. Finally, the main function, **run_ORFquant**, takes the **for_ORFquant_file** and the .gtf_Rannot file as inputs. It processes these to identify ORFs, outputting the results in an Rdata format.

For this PhD project, I have run the ORFquant on multiple human and mouse datasets. There are no customizable settings related to the data processing besides the Ribo-Seq read length offsets. I have used the offset settings as shown in Table 1 below for all the data processing. This offset table was initially provided by Dr Joseph Waldron, then I further validated on CDSs for both human and mouse data.

Read length	Offset	Compartment
29	12	nucl
30	12	nucl
31	13	nucl
32	13	nucl
33	13	nucl
28	12	chrM
55	12	chrM

Table 1: **Ribo-Seq read length offset settings.** The read length is the Ribo-Seq reads, the offset indicates the p-site position on the read and the compartment states either it is nuclear encoded or mitochondria encoded.

Ribotricer (v1.3.2)

Ribotricer is designed for the accurate detection of actively translating open reading frames (ORFs) using ribosome profiling (Ribo-seq) data (Choudhary, Li, and D. Smith 2020).

Ribotricer is running in Python (> 3.7) and has two main functions for ORF detection. Initially, as mentioned above, Ribotricer detected and annotated all possible ORFs using the function **prepareorfs**. This function takes the gene annotation file in .gtf format and the whole genome sequence FASTA file then generates the index of all ORF candidates. The second function which is the main translating ORF detecting function **detectorfs** takes the ORF candidate index generated by the first function and the sorted Ribo-Seq alignment file in .bam format then calculates the phase score and identifies the translating ORFs with the output of ORFs in .tsv file.

2.2.3 Feature Engineering

To further characterise the established translating ORF candidates library, I compiled a set of features for each identified ORF. This feature matrix captures various attributes of the ORFs, from sequence-based features to

external database driver annotations. For treatment effective analyses, RNA-Seq (gene level) and Ribo-Seq (ORF level) counts were modelled with DESeq2 (v1.44.0)

Differential expression analysis

To assess treatment effects in the mouse experiments mentioned before in this chapter, differential expression (DE) was performed at gene level for RNA-Seq, and at ORF level for Ribo-Seq. All analyses used the same reference build and gene model as elsewhere in the Methods. Libraries comprised a control and three treatments (MNK inhibitor, mTOR inhibitor, combined MNK and mTOR inhibitor). Each group has three biological replicates followed the experimental design used throughout this study. This analysis was set to test each treatment versus control. For RNA-Seq, exon read counts were summarised to the gene level using strand-specific rules matching the library protocol. For Ribo-Seq, counts were obtained by mapping reads falling within BYU-defined ORF coordinates on the sense strand. Candidates with low counts were not removed by any explicit pre-filter. Instead, filtering was handled by DESeq2 independent filtering, based on the mean of normalised counts. DE was performed using DESeq2 (v1.44.0) in R. A generalised linear model (GLM) with `design ~ batch + condition` was fit, where batch captures technical variability such as organoid preparation and condition encodes the four experimental groups. GLM is a statistical framework that extends ordinary linear regression to handle data that is not normally distributed. This GLM framework separates biological signal (condition) from technical noise (batch and library size) while properly accounting for count variability, so that replicate information increases confidence without inflating false positives. Wald tests on these coefficients then produce P values, which are adjusted for multiple testing using the Benjamini–Hochberg procedure. Reported effect sizes are in log2 fold-changes.

Machine learning algorithms

In this section, I will introduce how to evaluate the machine learning (ML) model, SVM and RF, which I have introduced in the Introduction. The SVM models were trained using the ‘ksvm’ function from the ‘kernlab’ package in R, employing a radial basis function (RBF) kernel. The penalty parameter C was set to 1, and the kernel width parameter σ was set to 0.1. The RF models were trained using ‘randomForest’ package, with the number of trees set to 200. These implementations were benchmarked to provide an optimal balance between performance and computational efficiency. The evaluation is based on the accuracy, sensitivity and specificity. Sensitivity, also known as the true positive rate, indicates how well the model can detect true positives (TP) while avoiding false negatives (FN). Specificity, or the true negative rate, on the other hand, reflects how the model performance of correctly classifying true negatives (TN) while minimising false positives (FP).

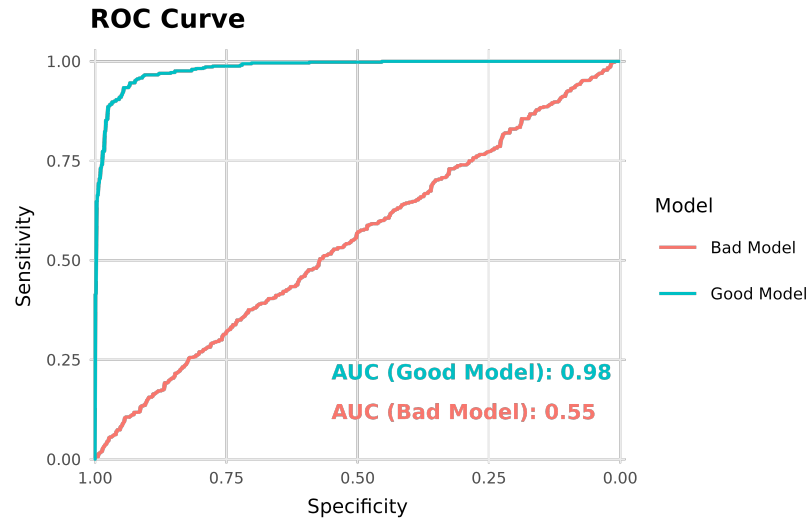


Figure 10: **Receiver operating characteristic curve example.** These plots show the ROC (Receiver Operating Characteristic) curves for two models based on synthetic, randomly generated data. The X-axis is the specificity and the Y-axis is sensitivity. And the area under the curve (AUC) represent the performance of the model. AUC ranges from 0 to 1 where 1 indicates the best performance in classification. The 'Good Model' (blue line) simulates a strong predictive model, while the 'Bad Model' (red line) represents a model making random guesses.

901 To visualise the model's performance, a receiver operating characteristic (ROC) curve is used, as shown in
 902 Figure 10. On this curve, the X-axis represents 1 - specificity, and the Y-axis represents sensitivity. The
 903 overall model performance is quantified by calculating the area under the curve (AUC), which ranges from 0
 904 to 1 and the higher the AUC is, the better the ML model performance is. To understand how the RF model
 905 evaluates feature importance, importance analyses were conducted through assessing two primary metrics:
 906 Mean Decrease Gini (MDG) and Mean Decrease Accuracy (MDA).

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + FP + TN + FN}, & \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Specificity} &= \frac{TN}{TN + FP}, & \text{Sensitivity} &= \frac{TP}{TP + FN} \end{aligned} \quad (2.1)$$

Package	Version
AnnotationHub	3.11.5
Biostrings	2.71.6
clusterProfiler	4.12.0
coRdon	1.21.0
data.table	1.15.4
dplyr	1.1.4
DESeq2	1.44.0
ensemblDb	2.27.1
eulerr	7.0.2
GenomicAlignments	1.39.5
GenomicScores	2.16.0
ggplot2	3.5.1
ggpubr	0.6.0
ideogRam	0.1.1
kernlab	0.9.32
kmer	1.1.2
kaos	0.1.2
liftOver	1.27.0
maftools	2.19.0
ORFquant	1.02
parallel	4.4.1
pROC	1.18.5
randomForest	4.7.1.1
rtracklayer	1.63.3
Ribotricer	1.3.2
Rsamtools	2.19.4
stringr	1.5.1
TCGAbiolinks	2.31.4
umap	0.2.10.0
universalmotif	1.21.4
UpSetR	1.4.0
reshape2	1.4.4

Table 2: List of installed R packages.

Development of the BYU Workflow for Translated uORF Identification

With recent advances in sequencing technology, we are now able to understand translation at the codon level, providing us with valuable insights into the connection between the transcriptome and proteome. For example, over the past decade, translation events have been observed in non-coding regions, and microproteins have been found encoded in the 5' and 3' UTRs, significantly extending our understanding of microproteomics (Anderson et al. 2015; Patraquim et al. 2020). However, as discussed in the Introduction, the discovery of translated small ORFs (sORFs), such as upstream ORFs (uORFs), has been sporadic and unsystematic, primarily due to the lack of a robust ORF-detection workflow that is capable of accurately capturing ORFs with short lengths. Existing methods are predominantly designed for canonical ORF discovery. I therefore hypothesise that these methods are more sensitively tuned towards long ORFs, potentially leading to a high false-negative rate for sORFs due to their shorter length and weaker periodicity signals.

The concept of developing my own ORF detection workflow emerged during the first two years of my PhD, a period primarily spent utilising predictions made by others. During this initial phase, I employed published ORF annotations and sequencing data to identify translated ORFs. Additionally, I developed feature extraction modules, attempting to utilise machine learning algorithms for distinguishing sequence pattern differences among ORFs. At that time, I found many of the existing ORF annotations lacked substantial support for non-canonical ORFs as they only considered AUG initiated ORFs.

As mentioned in the Introduction, numerous sORFs in non-coding regions, such as uORFs, are non-AUG initiated. Consequently, current ORF annotations do not adequately cover these regions. Meanwhile, the state-of-art ORF identification algorithms place significant emphasis on the three-nucleotide phase of Ribo-Seq reads (Calviello, Hirsekorn, and Ohler 2020; Choudhary, Li, and D. Smith 2020; Gelhausen et al. 2022). However, it has been suggested by Ingolia that the periodicity of reads may be inconsistent when the local read depth

931 is low causing low signal-to-noise ratio (Nicholas T. Ingolia 2016). Meanwhile, this limitation could also be
932 affected by the data quality and the sequencing depth. If the dataset has a shallow coverage, then the periodic
933 signal could be much difficult to be distinguished from the background noise, suggesting that current tools may
934 be less resilient to variations in experimental design. Statistical modelling demonstrates that for sORFs, the
935 limited number of codons is often insufficient to establish a statistically significant periodic signal, even when
936 translation is occurring (Michel et al. 2012; Aspden et al. 2014). Many of those algorithms are designed to
937 identify canonical ORFs (CDSs) with an AUG start codon (Calviello, Hirsekorn, and Ohler 2020).

938 Nevertheless, extensive evidence of non-AUG translation initiation sites has been discovered in the human
939 genome (J. Chen et al. 2020). Furthermore, noncanonical translation initiation events were found to predomi-
940 nate in non-coding regions such as 5' UTRs (Andreev, Gary Loughran, et al. 2022). Crucially, the rigid exclu-
941 sion criteria employed by these periodicity-dependent algorithms create a significant blind spot in the current
942 annotation landscape. As these tools operate on a binary discarding any region that fails to meet a strict period-
943 icity threshold, it becomes impossible to characterise the population of ORFs being excluded. Consequently,
944 the field currently faces an unquantified false negative rate, leaving us with no insight into the biological nature
945 of the potentially translated elements that are systematically missed by standard workflows. Therefore, it is
946 crucial to develop an ORF discovery method that can accurately capture and analyse the intricate landscape of
947 translational regulation to advance our understanding of the 5' UTR. Meanwhile, existing methods primarily
948 focus on filtering ORFs based on the triplet periodicity in Ribo-Seq data, which reflects the codon by codon
949 pattern of ribosome movement along ORFs. However, many sORFs exhibit low translation level, resulting in
950 reduced ribosome occupancy. This lower signal-to-noise ratio leads to poor periodicity and, consequently, a
951 decreased rate of discovery.

952 To resolve these limitations, I developed a workflow for uORF detection to capture the nuance of the uORF
953 landscape. My workflow, by contrast, integrates a multidimensional feature matrix, incorporating sequence
954 conservation, start codon context, and physicochemical properties, to identify potential non-canonical ORFs
955 that would otherwise be rejected due to poor phasing. This approach facilitates a more comprehensive charac-
956 terisation of uORF functionality and their potential roles as regulatory motifs or therapeutic targets.

957 Furthermore, the iterative development of my workflow over the years has led to a modular design. This not
958 only enables the quick integration of other tools but also offers the flexibility to adapt to various experimental
959 designs. In this chapter, I will present my workflow for ORF detection (Figure 11). The data discussed in this
960 chapter originates from a mouse project conducted in the Bushell lab by Dr Joseph Waldron et al. The experi-
961 ments aimed to explore the role of mitogen-activated protein kinase-interacting kinases (MNK) and mammalian
962 target of rapamycin (mTOR) in the KRAS signalling pathway, focusing on translation initiation. These exper-
963 iments included three biological replicates of APC^{-/-} KRAS^{G12D} mouse colon organoids, which were divided
964 into three treatment groups and a control group. For this chapter, I primarily utilised data from the control
965 group, as it exhibited the highest gene expression rate due to active mTOR signalling.

966 My approach to identifying translating ORFs follows the philosophical principle of '*Entia non sunt multi-*

plicanda praeter necessitatem’ which expresses the idea of Occam’s Razor, meaning hypotheses should be formulated with the fewest possible assumptions. Accordingly, I hypothesise that given accurate sequence alignment data, regions that identified as transcribed by RNA-Seq, and demonstrating ribosome occupancy above certain noise thresholds according to Ribo-Seq data, are likely engaged in ribosome-related events such as translation. The further determination of the offset of the Ribo-Seq reads and the RPF periodicity could be helpful but could be too stringent and resulting in a low discovery rate as also suggested in many publications (Ahmed et al. 2019; Erhard et al. 2018; Z. Xiao et al. 2018). My workflow is designed to test whether these regions indeed represent translating ORFs. The workflow can naturally be divided into three key steps: 1) ensuring accurate sequence alignment through careful evaluation and selection of optimal aligners during raw data processing; 2) identifying genomic loci exhibiting co-aligned reads, applying read-depth thresholds, and assessing translatable lengths to establish potential ORF candidacy; and 3) determining potential translating ORFs by constructing a comprehensive feature matrix incorporating internal experimental data and external knowledge bases.

3.1 Standardising raw sequence data pre-processing

The initial step in the ORF discovery workflow is to process the raw FASTQ read data from Ribo-Seq and RNA-Seq experiments. The Ribo-Seq data offer insights into the ribosome occupancy status of ORF candidates, whereas the RNA-Seq data reveal the levels of transcripts. Meanwhile, given that RNA-Seq and Ribo-Seq data represent the same transcriptome from different perspectives, their signals are expected to overlap and corroborate one another, providing mutually supporting evidence during further ORF identification. This raw data processing module is exclusively designed for single-end sequencing data, as the majority of available sequencing data is single-ended. However, it can be easily modified to accommodate double-end sequencing. My aim was to develop a workflow that is robust to differences in library preparation protocol, enabling the use of raw FASTQ files sourced from multiple data repositories, including in-house experiments and published data from the Sequence Read Archive (SRA). Owing to the varying library preparation protocols employed in these experiments, the linear layout and composition of the sequence reads, such as the arrangement of the biological insert relative to adapters and barcodes, can be different (Figure 12). Furthermore, it is often the case that provided metadata does not include complete details of all experiment configurations. Therefore, a workflow capable of identifying and adapting to various settings is necessary. To further align and analyse the data, it is essential to remove artefacts introduced during the sequencing process. Given the variability in experiment kits, there can be differences in read length, adapter choice, and UMI (Unique Molecular Identifier) design across experiments.

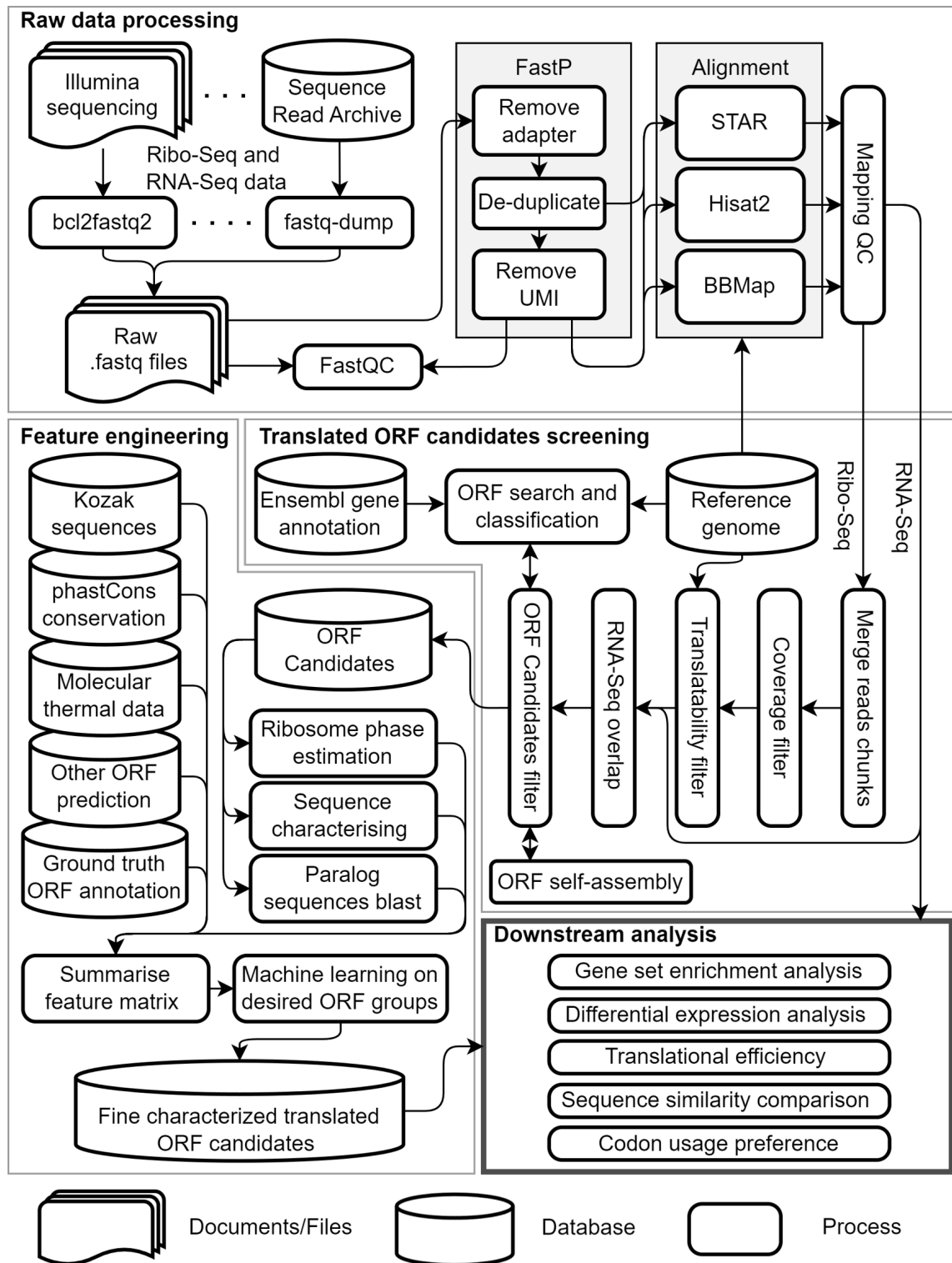


Figure 11: **Proposed ORF identification workflow overview.** The workflow consists of three major stages: 1) Raw data processing, which includes cleaning and aligning sequencing data to the reference genome; 2) Translated ORF candidate screening, involving the selection and filtering of candidate ORFs using aligned sequencing data and predefined criteria (coverage, translatability, and RNA-Seq overlap); 3) Feature engineering, where filtered ORF candidates undergo detailed characterisation using both internal analyses and external databases. Machine learning algorithms are subsequently applied to group and select the most biologically relevant ORF candidates. Finally, selected ORFs proceed to a series of downstream analyses designed to investigate their functional significance extensively.

998 3.1.1 Adapter removal

999 In the process of raw data processing, one of the first steps is the removal of sequencing adapters. These are
1000 artificial sequences added during library preparation. Adapters serve to bridge and link oligonucleotides in the
1001 flow cell to the target sequence during PCR. Consequently, different platforms and sequencing kits use different
1002 adapters (Figure 12). It is important to note that in projects employing both Ribo-Seq and RNA-Seq, even when
1003 the same kit is purportedly used according to the experiment reports, the actual adapters used may differ.

1004 3.1.2 De-duplication

1005 De-duplication is essential to eliminate duplicate reads that might artificially inflate the representation of cer-
1006 tain sequences, potentially biasing the analysis. Various de-duplication methods are applied in RNA-Seq and
1007 Ribo-Seq, broadly categorized into two types: those with, and those without, Unique Molecular Identifiers
1008 (UMIs). UMIs are short sequences can be introduced during library preparation that uniquely tag each original
1009 RNA molecule, enabling precise quantification of transcript abundance. De-duplication with UMIs is straight-
1010 forward; duplicates are removed since UMIs uniquely tag mRNA abundance before PCR amplification. In
1011 contrast, for de-duplication without UMIs, reads are sorted by genomic alignment position, and duplicates are
1012 flagged based on identical start and end positions. While UMIs enable accurate identification of PCR dupli-
1013 cates, in practice many bulk RNA-Seq experiments are performed without using UMIs.

1014 3.1.3 UMI removal

1015 The handling of UMIs is a crucial step in processing sequencing data. UMI processing begins with the verifi-
1016 cation of the correct UMI design, which can be determined from metadata or from downstream quality reports.
1017 As shown in Figure 12, different experiment design with different sequencing protocol may produce different
1018 sequence structure.

1019 Consequently, the processes for adapter and UMI removal should be designed to be adaptable to various sce-
1020 narios. Fastp (S. Chen et al. 2018), a recently developed all-in-one FastQ pre-processing tool, was chosen
1021 for its capabilities in supporting automatic adapter and UMI identification, as well as for its fast and accurate
1022 deduplication. Furthermore, the real-time quality control process applied to the processed reads facilitates the
1023 removal of UMIs from reads generated in experiments with unknown UMI designs.

1024 3.1.4 Genome alignments tool benchmarking

1025 After pre-processing, the cleaned Ribo-Seq and RNA-Seq reads were prepared for genomic alignment. My
1026 initial analyses, which used HISAT2 to map data from the previously mentioned mouse project revealed a sig-
1027 nificant issue: many genomic loci with Ribo-Seq read coverage were not supported by corresponding reads

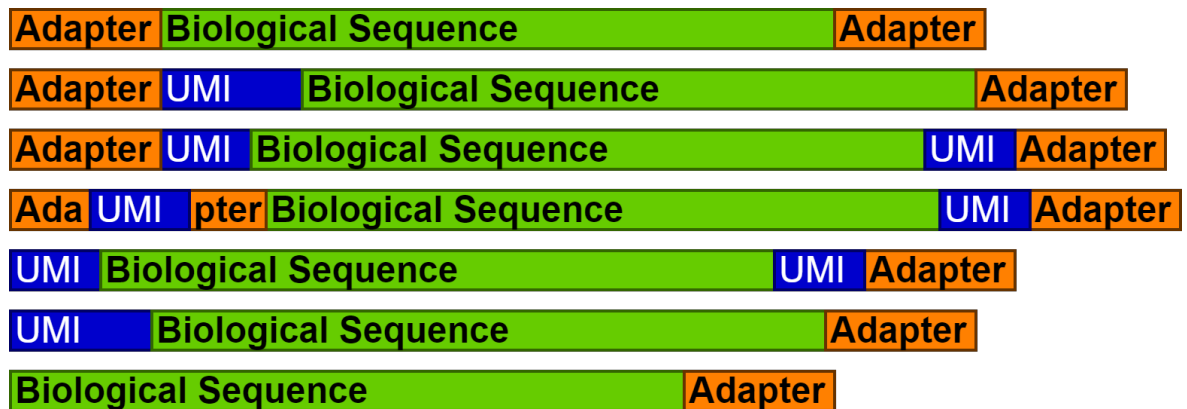


Figure 12: **Read structures in sequencing data.** Each row represent a typical raw sequencing reads obtained from RNA-Seq or Ribo-Seq experiments. Each sequencing read can contain any of three distinct components: adapter (orange), Unique Molecular Identifier (UMI, blue), and biological sequence of interest (green). Note that the UMI is not always employed in sequencing. The presence and position of each component vary depending on the library preparation protocol applied.

1028 in the matched RNA-Seq data. An example of this discrepancy is shown in Figure 13. I hypothesised that
 1029 this discrepancy might reflect a technical artefact introduced by the mapping tool. Ribo-Seq reads are typi-
 1030 cally much shorter than conventional RNA-Seq reads, generally spanning approximately 26–34 nucleotides.
 1031 This limited read length presents a substantial challenge for genomic alignment, as many existing aligners are
 1032 optimised for longer reads, potentially leading to alignment inaccuracies or reduced performance when han-
 1033 dling short-read data (Ziemann, Kaspi, and El-Osta 2016). While high translation efficiency on low-abundance
 1034 transcripts is possible, the complete lack of RNA-Seq support may suggest careful evaluation is needed. This
 1035 discrepancy could result from differences in sequencing depth between the Ribo-Seq and RNA-Seq libraries.
 1036 Because Ribo-Seq is an enrichment-based method, it captures reads on actively translated regions, therefore an
 1037 effectively translated but low abundant mRNA could generate a high-coverage Ribo-Seq peak due to this en-
 1038 richment effect, while remaining undetected by the broader random sampling of standard RNA-Seq. However,
 1039 I hypothesised that a large proportion of these unsupported Ribo-Seq signals might represent technical artefacts
 1040 introduced by the mapping tool, rather than genuine translation events. To resolve this, I designed a bench-
 1041 marking and quality control strategy to evaluate aligner performance. The primary objective was to minimise these
 1042 false-positive mappings and maximise the concordance between the Ribo-Seq and RNA-Seq datasets, ensuring
 1043 that identified translation events were supported by transcriptional evidence. To investigate this, I compared the
 1044 performance of three widely-used alignment tools: STAR (v2.7.10a), HISAT2 (v2.2.1), and BBDMap (v39.01).
 1045 The benchmarking used matched Ribo-Seq and RNA-Seq datasets from control experiments conducted by Dr
 1046 Joseph Waldron et al. (unpublished). In this benchmarking, I used the RNA-Seq libraries as a conservative
 1047 reference for the Ribo-Seq and treated absent coverage in a region as a putative biological absence instead of a
 1048 result of poor sequencing.

1049 All datasets were aligned to the mouse reference genome (GRCm39). Each aligner was run with parameters
 1050 optimised for short-read mapping. Such as, the maximum number of mismatches allowed was reduced to better
 1051 suit the Ribo-Seq read length. A complete list of the specific parameters used for each tool can be found in the
 1052 Methods. For the purposes of this comparison, a genomic region was defined as ‘covered by RNA-Seq’ if at

1053 least one RNA-Seq read (after noise filtering) aligned to it.

1054 I firstly assessed the performance on the RNA-Seq data to establish a robust transcriptional baseline. In this
1055 quality control step, all three aligners performed comparably, achieving over 98% uniquely mapped reads. This
1056 confirmed that the RNA-Seq dataset was of sufficient quality to serve as a reference for RNA expression. The
1057 key challenge was the alignment of the short Ribo-Seq reads. My analysis focused on the consistency between
1058 the Ribo-Seq and RNA-Seq alignments, with the goal of minimising Ribo-Seq reads mapping to regions with
1059 no transcriptional evidence.

1060 As mentioned previously the example shown in Figure 13 illustrates the core problem, how HISAT2 pro-
1061 duced putative alignments in a region completely devoid of RNA-Seq reads, whereas STAR's alignments cor-
1062 responded well with the RNA-Seq coverage. This observation was quantified across the entire dataset (Figure
1063 14A). STAR produced the most conservative and consistent results, with the lowest proportion of Ribo-Seq
1064 reads mapping to regions unsupported by RNA-Seq. Conversely, both HISAT2 and BMap mapped a signifi-
1065 cantly higher fraction of Ribo-Seq reads to unsupported regions.

1066 Further analysis supports the hypothesis that these unsupported reads are artefacts. As shown in Figure 14B,
1067 both HISAT2 and BMap tended to map these unsupported Ribo-Seq reads to un-annotated regions (according
1068 to Ensembl 109). Conversely, STAR produced the most conservative mappings, with alignments predominantly
1069 confined to annotated gene regions that were well-supported by RNA-Seq evidence. Furthermore, Figure
1070 14C shows that the Ribo-Seq reads lacking RNA-Seq support consistently exhibited lower local read depth,
1071 a characteristic feature of sequencing noise rather than genuine translation events. Among those, STAR still
1072 demonstrated the highest performance on mapping the reads to the RNA-Seq supported region.

1073 Taken together, these data show that the choice of aligner has a profound impact on the quality and reliability
1074 of Ribo-Seq mapping. STAR consistently demonstrated the highest accuracy, maximising the overlap between
1075 Ribo-Seq and RNA-Seq data while minimising spurious alignments to regions without transcriptional evidence.
1076 I, therefore, selected STAR for all subsequent analyses in this project.

1077 **3.2 Translated ORF candidates screening prioritises translatability and** 1078 **transcriptional coverage**

1079 After both RNA and Ribo Seqs have been correctly aligned to the genome, the next step is to identify the
1080 potential open reading frames that are undergoing translation. This second procedure of this ORF identification
1081 workflow takes the reference genome, Ensembl gene annotation, and the aligned sequence data as the input,
1082 and then generates the ORF candidates dataset. The identification process consists of two primary stages: first,
1083 the formation and filtering of small sequence segments based on read coverage and translatability, and second,
1084 the assembly of these segments into full-length ORF candidates. The specific implementation of these stages
1085 are detailed in the subsections below.

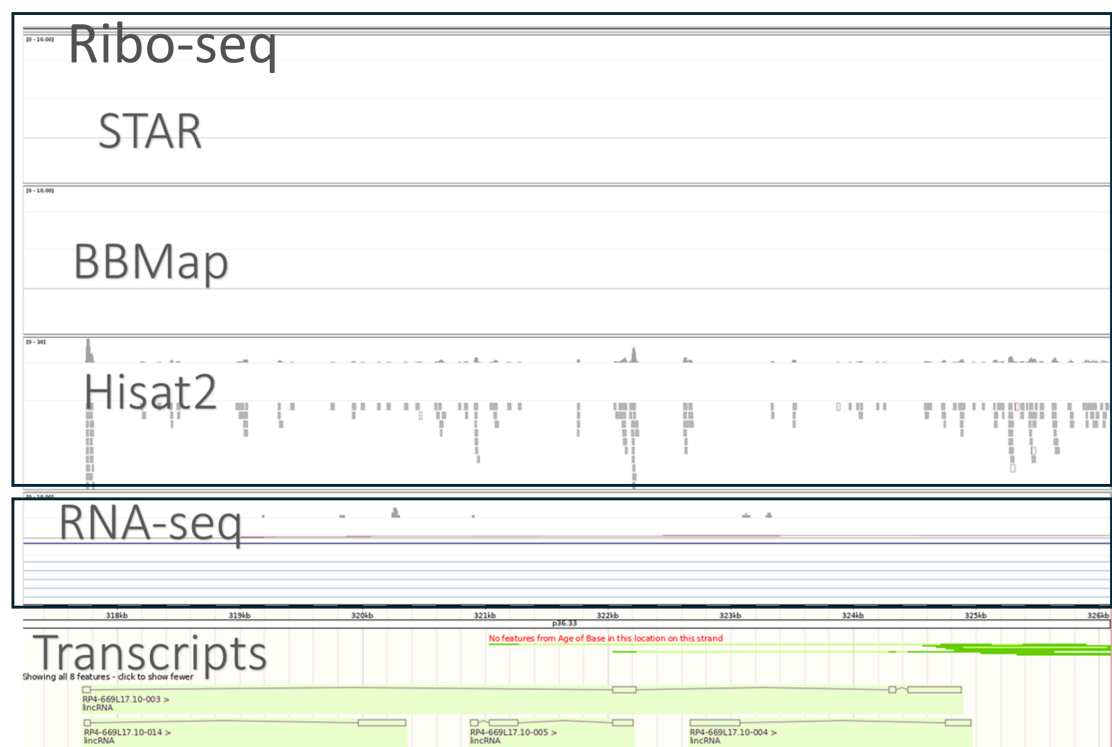


Figure 13: **Snapshot of Ribo-Seq alignment comparison.** Comparison of Ribo-Seq read alignments using STAR, BBMap, and HISAT2 against corresponding RNA-Seq reads within the same genomic region. The top three tracks are the Ribo-Seq alignments the fourth track is the RNA-Seq and the last track is the transcripts annotation. The performance of the aligners on Ribo-Seq data varies due to the short length of Ribo-Seq reads, leading to potential pseudo-mapping issues. As shown in this snapshot, HISAT2 aligns reads to regions not covered by RNA-Seq, highlighting inconsistencies that can arise when aligning Ribo-Seq data. This example highlights the complexities associated with accurately aligning short Ribo-Seq reads and emphasises the necessity for careful selection of alignment software to minimise false mappings.

3.2.1 Forming and filtering the small translatable units (STUs)

The initial step is to convert aligned sequencing reads into biologically meaningful elements. To do this, I defined a genomic structure called a ‘small translatable unit’ (STU), which represents a small, contiguous segment of a potentially translated ORF. Following alignment, Ribo-Seq reads aggregate into discrete sequence chunks. To ensure these chunks represent genuine translation events rather than mapping artefacts or background noise, they undergo a multi-step filtering process. These segments are checked for stop codons but not for start codons, as the objective here is to determine the translatability. The start stop codon pairing is in the ORF assembling step later.

First, regions with insufficient sequencing coverage are filtered out. Coverage is assessed using the absolute count of reads within each chunk. To determine an optimal threshold to separate genuine signal from noise, I plotted the number of identified chunks against their corresponding read counts to identify the inflection point where signal distinguishes from noise(Figure 15). By plotting the number of identified chunks against their corresponding read counts, a clear inflection point was observed at a minimum of 5 reads per chunk. This ‘elbow point’ represents the threshold beyond which increasing the read count requirement yields diminishing returns in distinguishing genuine translational hotspots. Applying this threshold to the mouse control replicate

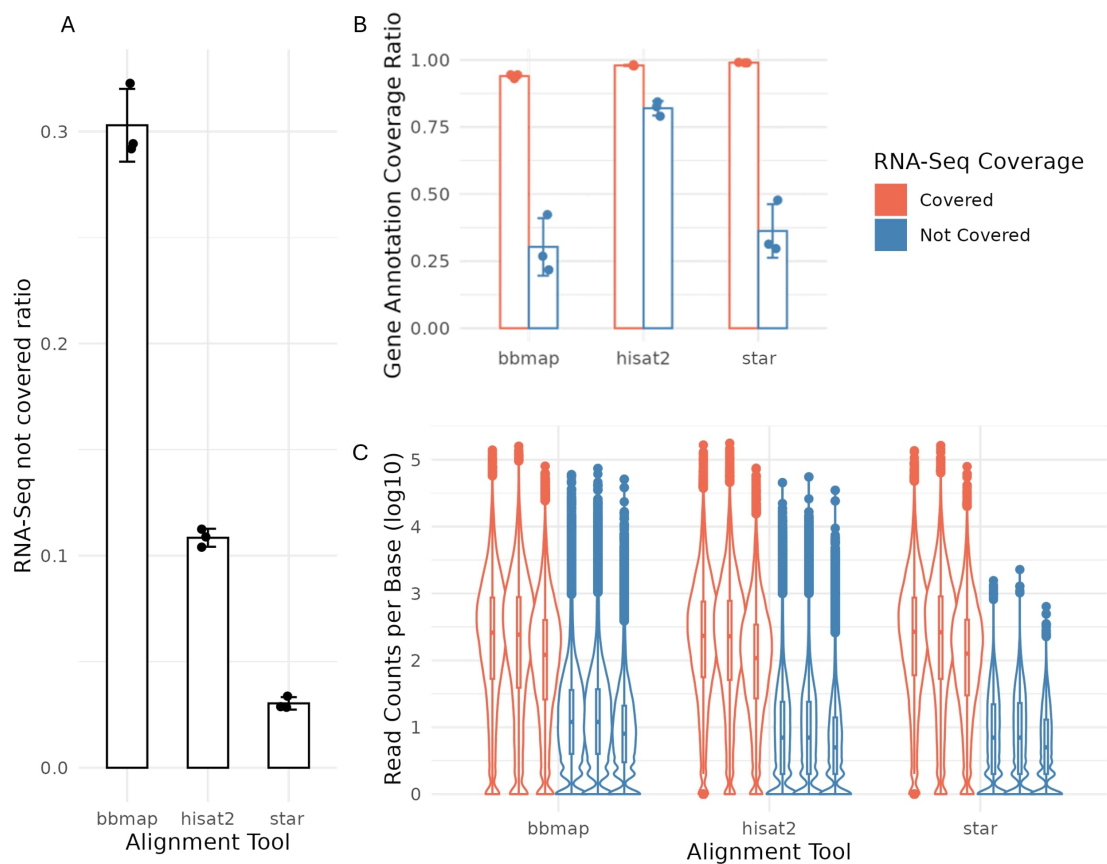


Figure 14: **Alignment tools performance analysis.** Performance metrics comparing alignments from BBMap, HISAT2, and STAR, using control datasets from experiments conducted by Dr Joseph Waldron et al. (unpublished). A) Proportion of Ribo-Seq reads aligned to genomic regions not supported by RNA-Seq alignments. STAR demonstrates the lowest proportion of these unmatched reads, indicating greater consistency between Ribo-Seq and RNA-Seq datasets. B) Alignment coverage of annotated gene regions (Ensembl 109), distinguished by RNA-Seq overlap status ('Covered' vs 'Not Covered'). STAR shows higher annotation coverage for Ribo-Seq reads supported by RNA-Seq evidence while minimising the low annotation coverage in RNA-Seq not supported regions. C) Violin plots depicting the distribution of local read depth within gene annotated regions for different RNA-Seq overlap status. Ribo-Seq reads lacking RNA-Seq support typically exhibit lower read depth, suggesting they likely represent sequencing noise rather than genuine translation events. Overall, STAR demonstrates the best alignment quality, with the highest overlap and alignment accuracy for short Ribo-Seq reads in this mouse dataset.

1101 3 (Ctrl 3) data, for example, reduced the number of chunks from 397,603 to 154,952. Comparing to Panel A
 1102 (STU counts), Panel B evaluates transcript level coverage as for a given minimum read per STU threshold on
 1103 the x-axis, the y-axis shows the number of distinct Ensembl transcripts that contain at least one STU meeting
 1104 that threshold (illustrated using Ctrl 3). This transcript wise elbow also occurs at five reads per STU, indicating
 1105 diminishing returns in transcriptome coverage beyond this point. Therefore, based on this analysis of the mouse
 1106 dataset, I adopt 5 reads/STU as the default coverage filter for the subsequent mouse analyses. Next, these high-
 1107 coverage chunks are tested for translatability. A chunk is defined as an STU if it contains no more than two
 1108 stop codons in any of the three reading frames. In this step, I am not checking for any start codons as it is for
 1109 checking if anything could terminate translation. Then the next step is to determine the ORF with paired start
 1110 and stop codon. The rationale for allowing up to two stop codons is to accommodate the authentic 3' end of an
 1111 ORF (which includes one stop codon) and to capture potential instances of stop codon read-through, a known
 1112 biological phenomenon. This filtering step further refined the dataset for the Ctrl 3 replicate to 127,962 STUs.
 1113 Finally, the filtered STUs are crosschecked with RNA-Seq alignments to ensure they originate from actively

transcribed regions. Only STUs that overlap with RNA-Seq reads are retained for the final ORF assembly. This step reduced the number of candidate STUs in the Ctrl 3 replicate to 98,993. The progressive reduction in chunks at each stage is quantified in Figure 16, illustrating the effectiveness of this filtering cascade in isolating regions with a higher evidence of translation. While this process filters out a large number of initial chunks, it is designed to enrich for true positives, a claim that is substantiated by the high recovery of known translated genes in validation analyses (see Chapter 5, Table 4).



Figure 15: **Elbow analysis of Short Translatable Unit (STU) coverage.** Sequencing data from Dr Joseph Waldron et al. (unpublished) were used to identify an optimal coverage threshold for STUs through elbow analysis. A) (STU-level): number of STUs retained (y-axis) as a function of the minimum Ribo-Seq reads required per STU (x-axis). B) (transcript-level): number of distinct Ensembl transcripts with more than 1 STU passing the same minimum read-per-STU threshold (x-axis); this summarises transcriptome coverage and shows the corresponding elbow. C): genome browser view illustrating Ribo-Seq read clusters; red boxes mark candidate STUs that pass the chosen coverage threshold. In both A and B (illustrated with control replicate Ctrl 3), a clear inflection (elbow) occurs at 5 reads/STU, which was used as the default coverage cut-off in subsequent analyses.

ORF assembling

Once a high-confidence set of STUs is established, the final step of this stage is to extending these evidence blocks into full length candidate ORFs. Two assembly methods are employed depending on the genomic context.

First is the annotation guided assembly. For STUs mapped to annotated transcripts, a gene annotation based method is used. This method first identifies all potential Translation Initiation Sites (TISs), currently set to ATG, TTG, CTG, GTG, and ACG, within the transcript regions covered by the STUs (Andreev, Gary Loughran, et al. 2022). A key criterion is that the TIS itself must be covered by Ribo-Seq footprints, as ribosomes are known to dwell longer during initiation, resulting in higher read density at start codons. For each identified TIS, an in frame stop codon search is performed to define the full ORF.

In contrast to this annotation-guided approach, for STUs in intergenic regions where no transcript model is

available, a selfassembly method is designed to assemble ORFs based on *de novo* transcript assembly from RNA-Seq data. For STUs located in intergenic regions lacking gene models, an annotation-guided approach is impossible. In these instances, the workflow switches to a *de novo* assembly mode, utilising the matched RNA-Seq data to reconstruct the underlying transcript architecture before applying the TIS identification logic described above.

At this stage, ORF identification emphasises breadth, retaining as many plausible translation events as possible for downstream characterisation. This inclusive strategy is underpinned by the earlier multi-step STU filtering in the previous Subsection 3.2.1 (i) a minimum coverage threshold of five reads per chunk to suppress noise; (ii) a translatability rule allowing no more than two stop codons across the three frames to capture genuine ORF termini while tolerating potential read-through; and (iii) an RNA-Seq overlap requirement to ensure transcriptional support. Together, these steps yield a high-confidence STU set that forms the basis of the candidate library.

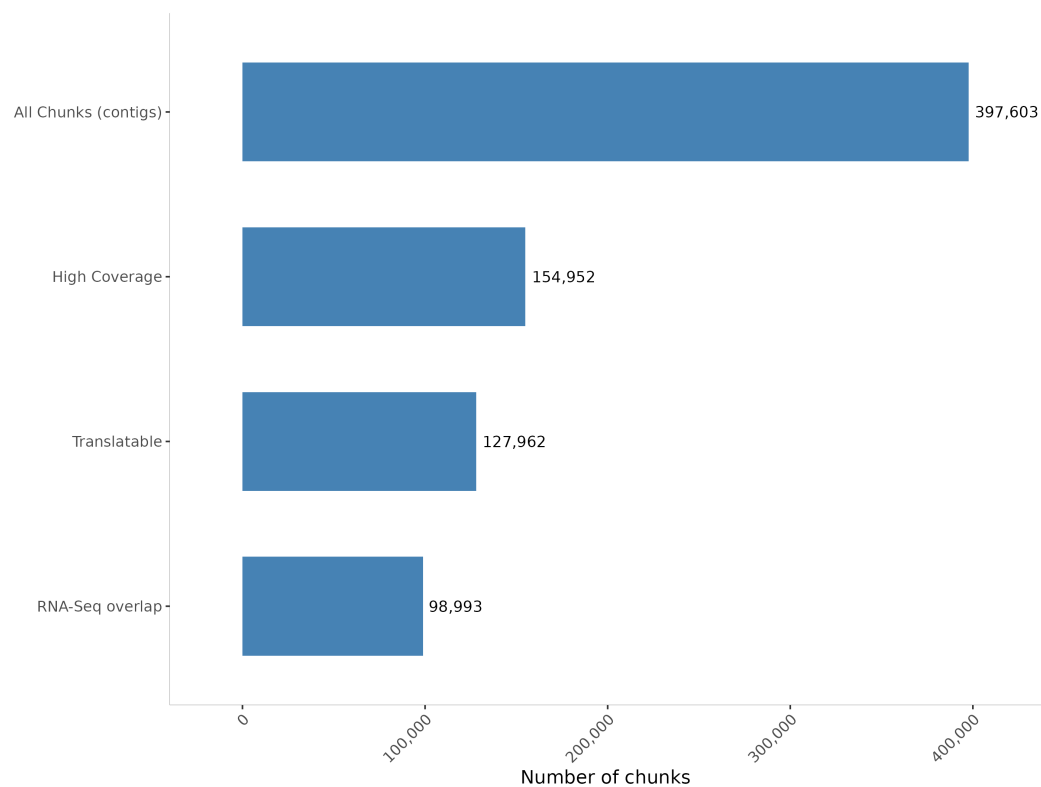


Figure 16: **Screening and filtering of Small Translatable Units (STUs).** Quantification of genomic sequence chunks at each filtering step of the translated ORF candidate screening workflow, derived from mouse experiment control replicate 3 as an example. ‘All Chunks’ represent initial assembled sequence chunks (contigs) formed by aligned Ribo-Seq reads. ‘High Coverage’ chunks are those exceeding the established minimum threshold of 5 reads per chunk, identified via elbow analysis (Figure 15). ‘Translatable’ chunks are those with fewer than three stop codons in at least one reading frame, indicating potential translation capability. ‘RNA-Seq overlap’ shows chunks passing previous filters and overlapping with RNA-Seq alignments, signifying transcriptional evidence. These filtered STUs subsequently enter the ORF assembly and annotation stage, ensuring more reliable candidate identification for downstream analysis.

This strategy aims to maximise the capture of potential translation events, which is reasonable given the strin-

gent filtering applied during STU formation. The process can yield multiple ORF candidates from a single STU, as one STU might be part of several overlapping ORFs initiated from different TISs within the same transcript. For the Ctrl 3 group, this assembly process generated a total of 533,909 ORF candidates from the 98,993 filtered STUs, forming a comprehensive library for subsequent feature engineering and characterisation.

3.3 Constructing a multi-dimensional feature matrix for ORF characterisation

In the third step of my ORF discovery workflow, I hypothesise that further characterisation of identified ORFs, coupled with machine learning algorithms, can effectively define and distinguish between various ORF groups. The feature matrix is designed to provide sufficient descriptive power to separate distinct ORF categories. In practical terms, adding informative and complementary features tends to increase class separability and can yield a stronger classifier, whereas adding uninformative or redundant features can introduce noise and hinder performance. Accordingly, I apply feature selection to prioritise features that contribute unique discriminatory signal and to reduce overfitting, thereby improving prediction accuracy.

By integrating these features from various knowledge bases, I expected to accurately classify and predict the functional significance of ORF candidates. To test these hypotheses, I extracted relevant features from the ORF sequence context and integrated data from genome sequences, wet lab experiments, and external databases. Utilising machine learning algorithms, I aim to validate whether the selected features lead to substantial information gain and improved classification performance, thereby advancing our understanding of small ORFs in translation regulation. The feature categories considered are: homologous sequence analysis, sequence conservation, secondary structure, and Kozak score; these are described in the following Subsections 3.3.1 - 3.3.5 below.

3.3.1 Homologous sequence analysis

Beyond the proteomic information encoded in ORF sequences, additional data embedded in the context of these sequences can provide insights into ORF behaviours during translation and other functionalities, as explored in the Introduction. Given that sequence similarity within the same species has been effectively utilised to infer homology (Pearson 2013), leading to speculation about related functions among coding genes, it is reasonable to posit that similar approaches may be useful in identifying functional relationships between ORFs and other genomic sequences. I employ the Basic Local Alignment Search Tool (BLAST) to identify homologous sequences of ORFs across the genome (Camacho et al. 2009) (Figure S1), as well as to search for sequence similarity among the identified ORFs. For the genome-wide BLAST search, I indexed the local genome FASTA file and built a BLAST database to facilitate parallel queries. For the within-ORF similarity search, another BLAST database was constructed using all the ORF sequences, each uniquely named by the ORF ID previously men-

tioned. Meanwhile, considering the prevalence of repetitive DNA throughout the genome (Ruiz-Ruano et al. 2016), I questioned whether those homologous sequence hits were part of larger repetitive sequence structures. Consequently, I developed a ‘sequence extending analysis’ algorithm to determine if the sequence similarity decreased upon extending the matching region at the 5’ and 3’ ends. This algorithm is based on recursive build-ups. For each uORF BLAST result, it extends all BLAST hit sequences by 100 nucleotides (nt) in one direction and then calculates the similarity score for the extended sequences. The recursion terminates if more than 50% of the extended sequences have a sequence similarity below 80%. Once one direction is fully extended, the algorithm proceeds to extend in the opposite direction. If an uORF can be extended more than 500 nt, it will be classified as an incomplete homologous sequence. Further details regarding the BLAST algorithm will be provided in the Methods. In this context, I gathered and annotated ORFs with information on ORF sizes, start codons, the number of BLAST hits, and the average length of these BLAST hits.

3.3.2 Sequence conservation landscape calculation

Cross-species sequence conservation can provide valuable insights into the functions and evolutionary development of ORFs. High conservation scores may suggest ORF evolution occurred at an early stage of speciation, indicating that their functionality could be fundamental across different species (Magny et al. 2013). For this project, conservation data were obtained from the UCSC Genome Browser, with per-base conservation scores calculated using the phastCons program, a statistical model grounded in a phylogenetic hidden Markov model (Siepel et al. 2005). Currently, this workflow is applied exclusively to humans and mice. As such, two datasets are used: phastCons100way.UCSC.hg38 for the human genome build 38, and phastCons35way.UCSC.mm39 for the mouse genome build 39. A conservation score for each ORF is calculated by aligning the ORF with the per-base conservation genome track and then computing the average per-base conservation. If any part of an ORF aligns with regions lacking a conservation score, the overall conservation score of that ORF will be calculated based on the remaining aligned portions. If an entire ORF does not align with any scored region, its conservation score will be marked as 0.

3.3.3 Secondary structure measurement based on minimum free energy

The secondary structure of ORFs can regulate translation. As mentioned in the Introduction, *in silico* structure prediction is commonly performed by calculating the Minimum Free Energy (MFE). A lower MFE indicates a more stable structure, which suggests reduced translation efficiency (Clote 2005). Therefore, by examining the MFE of each identified ORF, we can hypothesise about their translational preference and any potential regulatory effects on downstream ORFs. The MFE and secondary structure were calculated using the ViennaRNA package, which incorporates molecular thermal data (Lorenz et al. 2011). I extracted all the ORF sequences and inputted them into the RNAfold algorithms in the ViennaRNA package to calculate their MFE and secondary structures. I did not normalise MFE based on the sequence length; instead, the downstream analysis will account for the positive correlation between MFE and sequence length.

3.3.4 ORF type

Genome annotations from Ensembl and RefSeq enhance the characterisation of each ORF by defining their relative transcription level positions, thus elucidating their relationships with canonical coding regions (Cunningham et al. 2022; Morales et al. 2022). Additionally, the reading frames of ORFs and their corresponding transcripts' CDSs can be determined, enabling the deduction of ORF types. Different ORF types can impact gene expression variably. As detailed in the Introduction, ribosomal scanning behaviour and reinitiation can be influenced by ORFs. Therefore, identifying the types of ORFs is crucial for advancing our understanding of their functional roles in regulatory control. The classification of ORF types (Figure 17) is determined by their relative position to the main coding region (CDS). If an ORF is located in transcripts without a CDS, such as noncoding genes, it is labelled as a Novel ORF; this also applies to intergenic ORFs. For ORFs aligned to coding genes, classification begins by determining the start codon position:

- **When the start codon is within the 5' UTR:**

- If the in-frame stop codon is also within the 5' UTR, it is classified as a *uORF*.
- If the stop codon overlaps with the CDS, it is an *overlap-uORF*.
- If it shares the same stop codon as the CDS, it is an *N-terminal extension*.
- If the stop codon is located in the 3' UTR, it is classified as an *overlap ORF*.

- **When an ORF starts at the same position as the CDS:**

- If the stop codon is within the CDS, it is a *C-terminal extension*.
- If it shares the CDS stop codon, it is labelled as *annotated*.
- If it extends into the 3' UTR, it is labelled as an *overlap-dORF*.

- **For ORFs that initiate inside the main coding region:**

- If the stop codon is also inside the CDS, it is named an *inside ORF*.
- If it shares the same CDS stop codon, it is an *N-terminal truncation*.
- If it terminates in the 3' UTR, it is categorised as an *overlap-dORF*.

- **Finally, if the start codon is located in the 3' UTR, the ORF is classified as a *dORF*.**

Continuing with the Ctrl 3 replicate example, previously I assembled a total of 533,909 ORFs, which were then categorised into the following types based on my typing methods: 'Annotated' (21,149), 'C-extension' (7), 'C-truncation' (189), 'dORF' (331,081), 'N-extension' (29,375), 'overlap-ORF' (594), 'overlap-uORF' (43,321), and 'uORF' (108,193). Meanwhile, besides the ORF typing information, Ensembl and RefSeq have annotated the transcripts with their respective biotypes and transcript support levels. The biotypes reflects the transcripts classification like whether the it is protein coding or long non coding. The support levels indicate the confidence in these transcript annotations. These information has further enhanced our understanding of the ORFs discovered within the transcripts region.

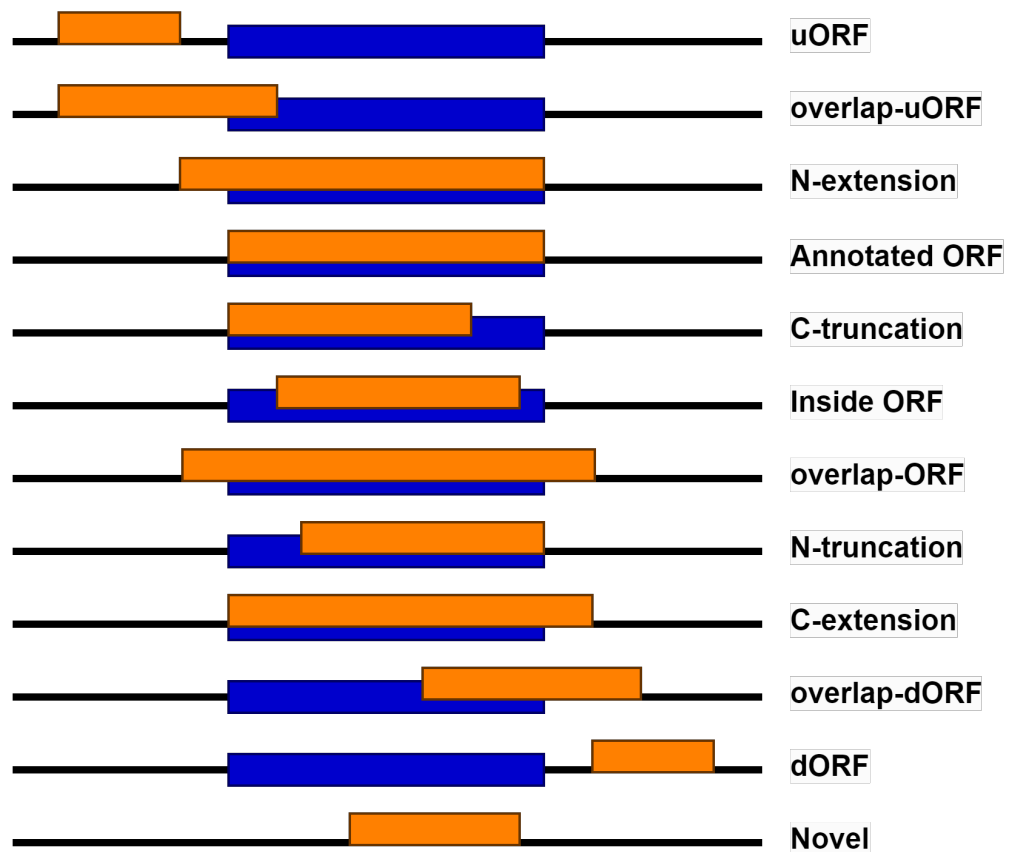


Figure 17: **ORF types**. Illustration of the various types of ORFs classified according to their positions relative to the main coding sequence (CDS). In this diagram, the blue bar represents the CDS, while the yellow bar denotes the ORF of interest. ‘uORF’ and ‘overlap-uORF’ are ORFs located upstream of the CDS, with the latter overlapping the CDS. ‘N-extension’ and ‘C-extension’ are ORFs extending beyond the N-terminus or C-terminus of the CDS. ‘Annotated ORF’ corresponds to an ORF that matches a known CDS. ‘N-truncation’ and ‘C-truncation’ indicate ORFs truncating the N-terminus or C-terminus of the protein. ‘Inside ORFs’ are entirely encompassed within the CDS. ‘Overlap-ORFs’ start upstream and end downstream of the CDS. ‘dORF’ and ‘overlap-dORF’ are ORFs downstream of the CDS, with the latter overlapping the CDS. ‘Novel’ ORFs are those without overlap with known transcripts. It should be noted that ORFs overlapping with the CDS can be challenging to detect due to difficulties in separating the Ribo-Seq signals of the ORF and CDS. This is especially true for C-extension and C-truncation types, where they share the same translation initiation site (TIS) with the CDS but stop at a novel position due to frameshifting or other mechanisms.

3.3.5 Kozak score

The kozak sequence, here defined as -9 to +5 nucleotides relative to the start codon, can influence translation initiation and is highly conserved within species (Hong Zhang et al. 2021). The Ensembl annotation of the transcriptome enables the retrieval and comparison of Kozak sequences within identified ORFs and CDSs. To quantify the preference of these Kozak sequences, I utilise a Position Weight Matrix (PWM) based algorithm for calculating Kozak scores. The PWM is a widely used method for representing nucleotide motifs, quantifying the likelihood of each nucleotide’s occurrence at specific positions within a set of aligned sequences. The matrix’s columns correspond to motif positions, while the rows represent the nucleotides possible at each position. Initially, I calculated a position frequency matrix (PFM) by counting the frequency of each nucleotide at each position from the alignments of annotated ORF Kozak sequences. The calculated PFM for human is as

1253 shown below:

	P_{-9}	P_{-8}	P_{-7}	P_{-6}	P_{-5}	P_{-4}	P_{-3}	P_{-2}	P_{-1}	P_{+3}	P_{+4}
A	20	20	21	21	19	24	46	29	19	22	27
C	27	33	32	23	32	38	10	38	45	15	40
G	35	29	28	39	30	26	37	20	28	49	18
T	18	18	19	17	19	12	7	13	8	14	15

1254 The PWM is then constructed based on the log ratio of probabilities. For each position i and nucleotide b
 1255 (where b is A, C, G, or T), the probability $p_{b,i}$ is calculated from the observed frequency $f_{b,i}$ as:

$$p_{b,i} = \frac{f_{b,i}}{\sum_{b \in \{A,C,G,T\}} f_{b,i}} \quad (3.1)$$

1256 Then the entry $M_{b,i}$ of PWM can then be calculated using:

$$M_{b,i} = \log_2 \left(\frac{p_{b,i}}{q_b} \right) \quad (3.2)$$

1257 where q_b is the background probability of nucleotide b , assumed here to be $q_b = 0.25$. The PWM is computed
 1258 as follows:

	P_{-9}	P_{-8}	P_{-7}	P_{-6}	P_{-5}	P_{-4}	P_{-3}	P_{-2}	P_{-1}	P_{+1}	P_{+2}
A	0.037	0.028	0.037	0.033	0.024	0.052	0.110	0.066	0.019	0.045	0.062
C	0.059	0.079	0.071	0.045	0.074	0.089	-0.036	0.092	0.108	0.002	0.094
G	0.079	0.062	0.059	0.092	0.066	0.052	0.087	0.033	0.062	0.112	0.024
T	0.024	0.028	0.037	0.024	0.037	-0.012	-0.071	-0.012	-0.047	0.002	0.002

1259 Finally, the Kozak score is calculated by summing the PWM entries corresponding to each nucleotide in the
 1260 sequence:

$$K = \sum_i M_{b_i,i} \quad (3.3)$$

1261 The Kozak score represents the similarity to the conserved, preferential Kozak sequence. A higher score
 1262 indicates a greater likelihood of high translational efficiency. ORF candidates possessing a preferential Kozak
 1263 sequence thereby gain increased confidence in being truly expressed.

3.3.6 External annotation and standardisation

Besides the information contained in established databases, additional ORF annotations derived from other prediction tools and published papers can also enhance our confidence in predicting expressed ORFs (Calviello, Hirsekorn, and Ohler 2020; Choudhary, Li, and D. Smith 2020; J. Chen et al. 2020; Calviello, Mukherjee, et al. 2016; Hong Zhang et al. 2021). All the ORF candidates identified in my prediction workflow will be cross-referenced with those shortlisted from these source. Due to the varying naming and mapping strategies used in different predictions and experiments, I developed an interface to standardise the annotation of coordinates to the genome level and create a unique identifier for each ORF (ORF-ID), thereby enhancing cross-panel comparison capabilities. Each ORF-ID follows the format ‘chromosome-start position-end position-ORF size’. All ORF candidate lists from various predictions and experiments are standardised accordingly, cross-referenced, and further analysed, as detailed in the following chapters. This standardisation enhances cross-study comparison and downstream integration.

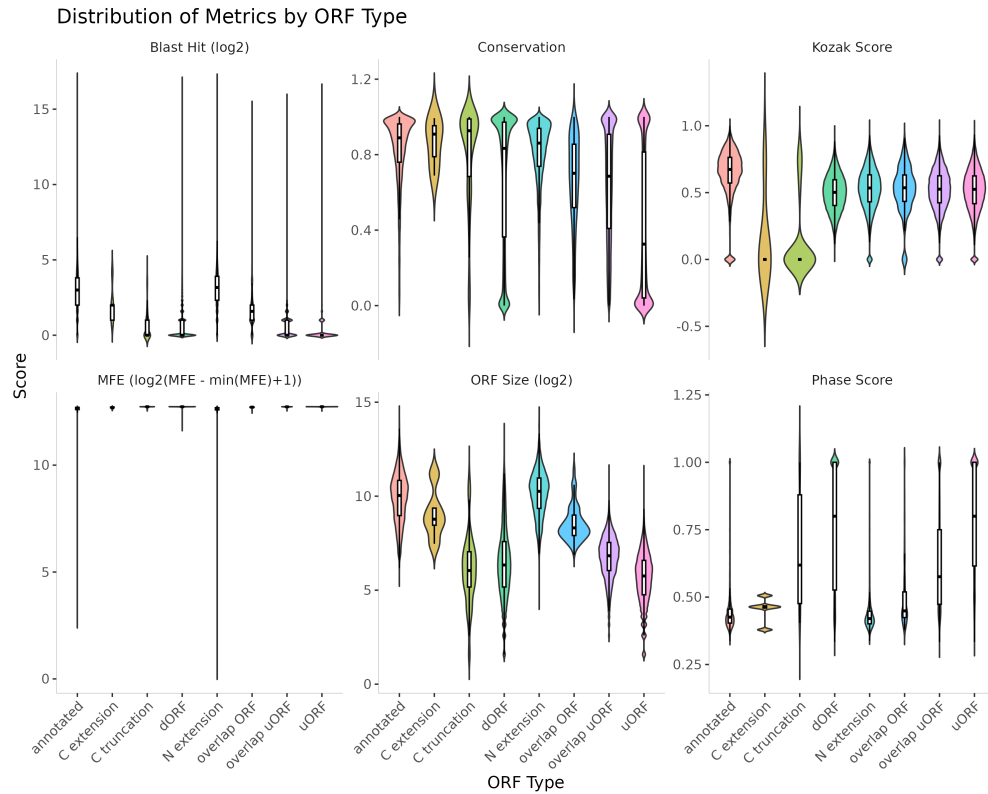


Figure 18: **Distribution of key feature scores across ORF types.** Violin plots illustrating six features used for ORF characterisation, grouped according to different ORF types. ‘Blast Hit ($\log_2$)’ represents the number of homologous sequences identified through genome-wide BLAST searches, indicating potential functional redundancy or evolutionary conservation. ‘Conservation’ reflects average phastCons scores, with higher values suggesting critical functional roles conserved across species. ‘Kozak Score’ quantifies translational initiation efficiency based on nucleotide preferences surrounding start codons. ‘MFE’ (Minimum Free Energy, represented as $\log_2(\text{MFE} - \min(\text{MFE}) + 1)$) denotes RNA secondary structure stability; lower MFE scores indicates more stable structures, influencing the translation efficiency. ‘ORF Size ($\log_2$)’ is the length distribution of ORFs, which reflex the translational dynamics. ‘Phase Score’ assesses ribosome footprint periodicity, where a higher score suggests a genuine translational activity. Collectively, these features create a robust feature matrix utilised for downstream machine learning analyses, enabling accurate prediction and functional classification of ORFs.

3.3.7 Conclusion

At this point in my workflow, a catalogue of candidate ORFs assembled from high confidence STUs has been annotated with a harmonised feature matrix for downstream modelling. As talked above, genome-wide homology search provides summary metrics such as the number and average length of hits, and a sequence-extending check helps distinguish genuine short homologues from matches driven by longer repetitive sequences. Cross-species conservation is integrated by aligning each ORF to base level conservation tracks and aggregating to an ORF-level score. Meanwhile, RNA secondary structure is summarised by the MFE, capturing potential structural influences on translation. Each candidate is then typed by its position relative to the host CDS and accompanied by transcript biotype as well as all the support metadata to provide contextual annotation. The nucleotide context surrounding the start codon is then quantified using the Kozak score for highlighting the preferred initiation motifs. Finally, to support the consistent cross-study comparison, coordinates are normalised to the chosen genome build and a unique ORF-IDs are assigned to every candidate.

By acquiring additional information from various knowledge bases as described above, I established a comprehensive feature matrix to characterise each ORF. Continuing to use the Ctrl 3 replicate, I generated a set of violin plots to demonstrate the variability of those features across different ORF types (Figure 18). The feature matrix facilitates further grouping using machine learning techniques and downstream analysis. A case study demonstrating the application of the proposed workflow will be presented in Chapter 4.

Chapter 4

Benchmarking and Validation of the BYU Workflow in Organoid and iPSC Models

In the previous chapter, I proposed a modularised workflow for the identification and characterisation of translated ORFs, based on Ribo-Seq and RNA-Seq data. The current framework of this workflow evolved from the workflow established in the first two years of my PhD as shown in Figure S2 and S3. The major developments were the individual analytical modules such as the conservation scoring system and homologous sequences identification. During this period, I was using the published uORF predictions (Hong Zhang et al. 2021) to explore their potential functionality in post-transcriptional regulation. Gradually, I introduced more structured analysis modules and ultimately transitioned from relying on online ORF annotations to employing my own de novo predictions. Throughout my PhD studies, I have investigated a range of datasets, which include both those published in academic papers and those generated within our institute, spanning projects involving both human and mouse subjects.

4.1 Introduction

4.1.1 Mouse Experiment

Ribo-Seq and RNA-Seq data from Bushell lab

The primary data presented in this chapter is from a mouse organoid project focused on experiments designed to explore the role of mitogen-activated protein kinase-interacting kinases (MNK) and the mammalian target

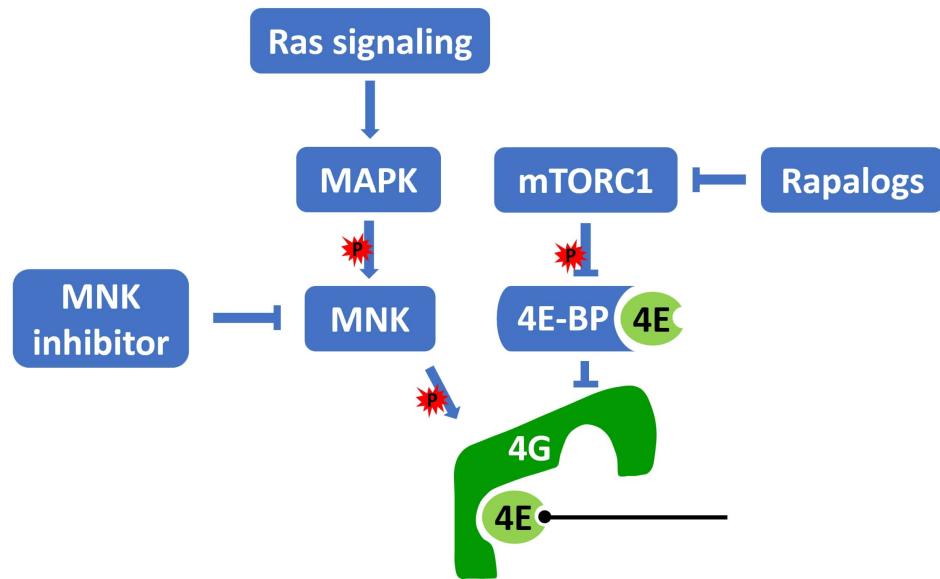


Figure 19: **mTORC1 and MNK signalling in translation initiation.** Simplified schematic representation of the mTORC1 and MNK signalling cascades controlling translation initiation in tumour cells. The diagram illustrates key regulatory elements, including Ras signalling activation of MAPK and MNK, and mTORC1-mediated phosphorylation of 4E-BP, enabling release of eIF4E (4E) and subsequent formation of the translation initiation complex with eIF4G (4G).

of rapamycin (mTOR) in the KRAS signalling cascade related to translation initiation (Figure 19). These experiments involved three biological replicates of APC^{-/-} KRAS^{G12D} mouse colon organoids, divided into three treatment groups and a control group. While the KRAS gene is one of the most frequently mutated oncogenes across all cancer types, it acts as a distinct driver in several highly fatal cancers, including nonsmall-cell lung cancer and colorectal cancer (L. Huang et al. 2021). The mutant KRAS remains perpetually active, continuously stimulating pathways that promote uncontrolled cell growth. Due to its structure, directly targeting KRAS in clinical practice is challenging (Vatansever, Erman, and Gümüř 2019). Consequently, therapeutic strategies often target downstream signalling pathways. In these mice organoid experiments, the treatment groups received eFT508 (an MNK inhibitor), rapamycin (an mTOR inhibitor), and a combination of eFT508 and rapamycin. Detailed information about the experimental design is detailed in the Section 2.1.1.

It has been demonstrated that both the MNK and mTOR signalling pathways converge to regulate the activity of the eukaryotic initiation factor 4E (eIF4E), which is crucial for ribosome recruitment during cap-dependent translation initiation (Saxton and Sabatini 2017; Knight et al. 2021). mTOR is composed of two distinct complexes, mTORC1 and mTORC2, with mTORC1 specifically tasked with regulating translation. mTORC1 phosphorylates the eIF4E-binding protein (4E-BP), preventing its binding to eIF4E and thus facilitating translation initiation (Saxton and Sabatini 2017). Consequently, targeting mTOR with rapalogs such as rapamycin can suppress the formation of the translation initiation complex. Meanwhile, MNK, activated downstream of the MAPK pathway, binds to eIF4G (Knight et al. 2021). This interaction promotes the phosphorylation of eIF4E at serine 209, enhancing the translation of specific mRNAs. However, unlike mTOR, the activation of MNK is not critical for normal cell development but is preferential for certain oncogene activations (Kosciuczuk, Saleiro, and Platanias 2017; Knight et al. 2021). For example, Knight et al. (2021) have shown that in

1333 KRAS mutant colorectal cancer, MNK mediated phosphorylation of eIF4E becomes critical for sustaining the
1334 translation of specific tumorigenic drivers such as c-MYC.

1335 In the treatment group where MNK is inhibited, the phosphorylation of eIF4E is reduced, leading to decreased
1336 translation levels. Some of the established uORF embedded mRNAs has been shown to depend on eIF4E
1337 availability and its phosphorylation status. For example, the translation of key stress-responsive transcripts,
1338 such as ATF4 and CEBPB, is modulated by the release and activity of eIF4E (H.-H. Chen and Tarn 2019).
1339 The experiments initially investigated the regulatory mechanisms of gene expression with a focus on the cap-
1340 dependent translation of the main coding region, this established link prompted further investigation. Given the
1341 widespread identification of uORFs in transcriptomes, I explored their role in regulating CDS expression. Using
1342 the raw Ribo-Seq and RNA-Seq data graciously provided to me, I conducted my own ORF predictions and
1343 compared my detection approach with state-of-the-art algorithms, namely ORFquant (Calviello, Hirsekorn, and
1344 Ohler 2020) and ribotricer (Choudhary, Li, and D. Smith 2020). Subsequently, analyses were conducted on the
1345 identified expressed uORF and CDS candidates to explore the characteristics of uORFs and their relationship
1346 to CDS expression.

1347 **Gold standard ORF predictions from published data**

1348 In order to validate the translated ORF identification algorithm, proteomic evidence is ideal. Although ribosome
1349 profiling captures 80S occupancy, it does not by itself demonstrate that a stable peptide is produced: initiation
1350 can abort early and nascent peptides can be rapidly degraded. However, the mouse project in our lab did not
1351 generate additional proteomic data (such as mass spectrometry) beyond the Ribo-Seq to assess the translation
1352 status of the sample. I therefore used two established datasets from previously published papers as the gold
1353 standard for translated ORFs. These datasets provide essential references primarily for two purposes: first,
1354 they are used for evaluating the specificity and sensitivity of my raw ORF identification pipeline, which was
1355 designed to maximise ORF coverage while maintaining a low false-positive discovery rate; second, these gold
1356 standard ORF datasets are employed to train a random forest machine learning model aimed at distinguishing
1357 different patterns within the feature matrix between ORFs with a high potential for translation (present in the
1358 gold standard) and those with a low potential for translation (absent from the gold standard). It should be noted
1359 that the coverage of these two datasets, like most existing proteomic datasets, mainly focuses on canonical
1360 coding regions and does not comprehensively characterize uORFs or other noncanonical ORFs. Consequently,
1361 for benchmarking and model training I restrict evaluation to annotated CDSs, and then apply the trained model
1362 to uORFs. Predictions for uORFs are interpreted alongside Ribo-seq signals and other orthogonal features
1363 introduced above. This strategy allowed the assessment where the gold standard exists while still enabling
1364 prioritisation of uORFs, which remains undersampled in available proteomic knowledge bases.

1365 To address the requirement for direct peptide evidence, the first gold standard dataset selected was the com-
1366 prehensive mouse tissue spectral library recently published by Lu et al. (2022), where they established a com-
1367 prehensive spectral library for mouse tissues using quantitative mass spectrometry. The library spanned across

1368 36 organs including intestine (corresponds to the colon, the jejunum, the cecum, the rectum, and the stomach).
1369 To achieve this, they collected 36 distinct mouse organs from three male and three female C57BL/6 mice. For
1370 generating the spectral library, peptides from those 36 tissues were fractionated into six fractions to enhance
1371 the proteome coverage and then analysed by data-dependent acquisition mass spectrometry (DDA-MS). Al-
1372 together, they have identified 12,320 proteins of which 8,494 were identified in the intestine. To verify the
1373 reliability and applicability of this newly developed spectral library, validation experiments were performed.
1374 Tissues from multiple organs (79 regions, including one colon region) were extracted from a female mouse and
1375 assessed against this library. This experiment then provided me with a colon specific gold standard proteome
1376 reference (5730 proteins).

1377 However, while proteomics confirms the presence of stable peptides, it often lacks the resolution to define
1378 precise translation initiation sites (TIS), particularly for non-canonical ORFs. To resolve this, I selected a sec-
1379 ondary gold standard which derived from the seminal ribosome profiling study by Ingolia et al. (2011). One of
1380 the outcome of their paper is the establishment of the simplified mammalian ribosome-profiling assay to allow
1381 a robust measurement of protein synthesis directly from ribosome footprinting data. With this assay, they have
1382 characterized translation in an mESC line (E14 mESCs) with the identification of 10,230 CDSs. I then used
1383 this CDS cohort as the gold standard of the translated CDS reference. Another critical advancement introduced
1384 by this research was the use of harringtonine, a compound that inhibits translation initiation by blocking the
1385 formation of peptide bonds soon after ribosomal subunit joining. They observed that harringtonine treatment
1386 could make ribosomes to stall and accumulate at sites of translation initiation sites. Then they reasoned that
1387 they could reliably and systematically identify translation initiation sites across the genome by analysing the
1388 ribosome accumulation patterns observed upon harringtonine treatment. By applying this method to 4,994
1389 highly expressed transcripts in the mESC study, they have identified a total of 13,454 candidate translation
1390 initiation sites including 3,811 annotated CDS, 3,541 uORFs, 1,770 overlap-uORFs. However, a caveat is that
1391 this TIS dataset was derived from mESCs. Therefore, when comparing this reference cohort with my mouse
1392 colon organoid data later in this study, tissue specific biases may be introduced.

1393 Taken together, these two datasets form a complementary and robust validation framework for my ORF iden-
1394 tification algorithm as well as reliable gold standard for machine learning. Ingolia et al.'s dataset provides a
1395 broad, extensively validated benchmark against which to measure general algorithmic performance, ensuring
1396 my pipeline maintains sensitivity and specificity comparable to established methods. Meanwhile, Lu et al.'s
1397 tissue-specific dataset provides critical biological context, allowing me to verify that my algorithm performs
1398 reliably within the specific biological environment of interest (colon-derived organoids).

1399 4.1.2 Human sequencing dataset

1400 A secondary dataset from a human project was also applied to evaluate the proposed ORF detection workflow
1401 in this thesis (J. Chen et al. 2020). Unlike the mouse data, this project was specifically designed for high-
1402 resolution ORF identification. The selection of a human dataset enabled the utilisation of mutational data from

the TCGA. This dataset includes Ribo-Seq, RNA-Seq data and MS-based proteomics data, along with a list of ORFs that have been experimentally validated to significantly impact cell viability and proliferation. These data originate from the work of Chen et al. (2020), who conducted a comprehensive study using induced pluripotent stem cells (iPSCs) to identify and functionally characterise non-canonical ORFs within the human genome. More detailed information about the experimental design, including sequencing setup and functional screening, is discussed in the Methods chapter.

4.2 ORF identification rate

The primary objective of my ORF identification workflow is to maximise the discovery rate of ORF candidates actively undergoing translation. At the stage of building the translated ORF candidate library (after the conclusion of the translated ORF candidate screening step Figure 11), my candidate identification strategy is to deliberately tolerate a higher false-positive rate in the first stage. For fairness, the comparator tools (ORFquant and ribotricer) were run with their default parameters; I did not relax their internal thresholds to emulate this high-recall stance. This is done intentionally because candidates retained by BYU at discovery stage are subsequently filtered using orthogonal evidences. The replicate concordance across the three controls, reading frame periodicity, transcriptional support from matched RNA-seq; and a supervised random-forest model (Chapter 5.6) that integrates conservation, homology, RNA structure (MFE), ORF context/type and Kozak score will be applied to down sample false positive calls. This approach ensures minimal exclusion of potential translation events, thereby preserving the broadest possible candidate pool for subsequent analyses. As discussed in the previous chapter, my motivation for developing my own ORF identification workflow was to enhance the detection rate of ORFs in previously poorly annotated regions, particularly in untranslated regions.

Several Ribo-Seq-based ORF identification tools were already available. Therefore, to evaluate the performance of the proposed workflow (BYU), I benchmarked it against two state of the art algorithms, ORFquant and ribotricer. This comparative analysis was conducted using the Ribo-Seq and RNA-Seq data from the three control replicates of the mouse colon organoid experiment, which present a realistic use-case with moderate sequencing depth not specifically optimised for ORF discovery. The performance of each tool was assessed based on both the total number of identified ORFs (CDS and uORFs) and the consistency of these predictions across biological replicates.

4.2.1 Existing algorithms show low sensitivities and replicate agreements for uORFs

ORFquant

I first applied ORFquant, an algorithm that uses the triplet periodicity in Ribo-Seq data as a hallmark of active translation. After encountering and resolving initial technical hurdles with the tool's source code, I applied it

to the mouse organoid dataset. To evaluate its performance, I started with the less challenging task of identifying expressed CDS using data from the control groups. I processed the raw sequencing data, aligned the reads to the mouse genome GRCm39, and then analysed them using ORFquant. The results (Figure 20E) revealed an average of 5,400 CDSs were identified as translated, with only 1,546 CDSs shared among the three replicates. These findings did not align with the broader scientific understanding that the mouse genome encodes approximately 48,000 genes, of which about 50% (around 22,000 genes, translating to 120,000 transcripts) are protein-coding, and typically, 10,000 to 20,000 genes are actively expressed (Breschi, Gingeras, and Guigó 2017; Rudolph et al. 2016). ORFquant recovered only about 25% of the expected expressed CDSs (5,400/22,000). Additionally, the consistency of discovery across replicates was relatively low, with only 16% of CDSs found to be expressed among those three replicates.

To further analyse the degree of agreement among the replicates, Fleiss' kappa was calculated. This statistic measures agreement for categorical items (in this case, whether an ORF is detected or not), correcting for the agreement that would occur by chance (Fleiss and Cohen 1973). The coefficient is calculated as $\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$, where $\bar{P}$ is the observed agreement and $\bar{P}_e$ is the expected agreement by chance. The κ , as explained in the equation, is ranging from -1 and +1, and a higher κ value indicates that the algorithm is consistently identifying the same biological signals across replicates, rather than selecting random noise. The analysis yielded a κ of -0.134 indicating poor agreement among the three replicates. This negative κ suggests that the algorithm's assessments may not be consistent or reliable, calling into question the robustness of the ORFquant results. Furthermore, as shown in Figure 20F, the tool identified far fewer uORFs with even less agreement among replicates, contrary to the expectation based on literature suggesting that up to half of the mammalian genome features at least one uORF (McGillivray et al. 2018; Hong Zhang et al. 2021; Wethmar et al. 2014). One caveat here is that those uORF expected in the literatures could be silent untranslated uORF, however, the ORFquant identified uORF is far less than the expected, which may suggest a high false negative rate.

Those observations suggest this discordance is due to the ORFquant rather than intrinsic to the data. As the lack of agreement is strongest for uORFs, which is expected when triplet periodicity and the minimum coverage thresholds are applied to short features with inherently fewer informative codons. Meanwhile, ORFquant emphasises sustained periodicity and continuous footprint density along the ORF and assign reads within annotated transcript models, making calls sensitive to small sampling differences between replicates. The initial results did not meet my objective of discovering the majority of ORFs under translation, particularly in the 5' UTR. This discrepancy in performance encouraged further investigation into potential limitations affecting ORF discovery rates. I then hypothesised that, under its default settings, ORFquant might not be ideally suited for uORF detection as uORFs are short and frequently show weaker, less sustained periodicity than CDSs, so many genuine events are filtered inconsistently across replicates.

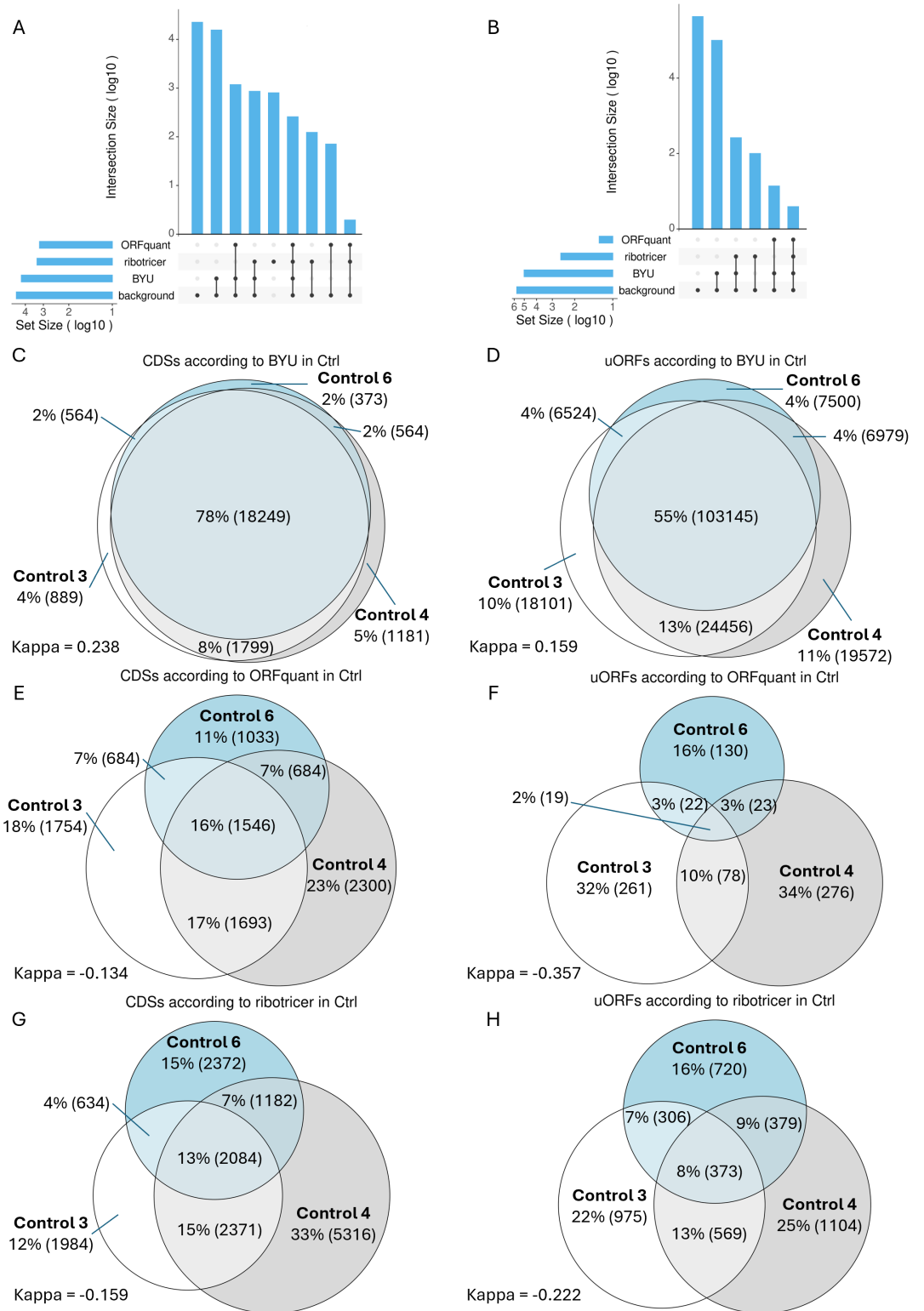


Figure 20: ORF identification rate comparisons in the mouse project. Comparison of ORF identification using the proposed BYU workflow with two state-of-the-art algorithms, ORFquant (Calviello, Hirsekorn, and Ohler 2020) and ribotricer (Choudhary, Li, and D. Smith 2020), based on low-depth Ribo-Seq and RNA-Seq data from three replicates (Control 3, 4, and 6) in mouse colon organoid experiments conducted by Dr Joseph Waldron (unpublished). Left panels illustrate CDS recovery rates, and right panels depict uORF discovery rates. A) and B) present upset plots comparing ORF discovery among the three different methods within CDS and 5' UTR regions respectively. Meanwhile, the background group represent CDSs and uORFs that identified using my workflow which without filtering on the Ribo-Seq threshold. C) and D) display Euler diagram showing the inter-replicate agreement on identified ORFs using BYU approach in the CDS and 5' UTR, respectively. E) and F) show similar inter-replicate agreement using ORFquant, whereas G) and H) utilise ribotricer. For all Euler diagrams, Fleiss' kappa value (-1 to 1) is calculated to statistically quantify the consistency of identification across the three replicates, where a higher value indicates better consistency. Collectively, these comparisons indicate that the BYU workflow achieves more consistent ORF identification across replicates, notably outperforming the other methods in detecting both CDS and uORFs, particularly when sequencing coverage is limited. This highlights the consistency and efficiency of the proposed BYU approach under suboptimal sequencing conditions.

Ribotricker

To test this hypothesis, I explored another ORF detection algorithm reputed for its ability to recover short active ORFs, known as ribotricker (Choudhary, Li, and D. Smith 2020). Ribotricker also operates on the principle of periodicity but introduces a novel calculation method focusing on the distribution of reads across the three frames of a codon. I input the same processed Ribo-Seq data into their algorithm to assess the CDS discovery performance. As illustrated in Figure 20G, ribotricker identified an average of 8,100 CDSs, with 2,084 CDSs consistently detected across the three replicates. Compared to the results from ORFquant, ribotricker exhibited a higher CDS discovery rate. However, the robustness across replicates remained relatively low with a κ of -0.159. Notably, in control group 4, ribotricker identified the highest number of CDSs (10,953), while its performance in the other two groups was more limited. Nonetheless, these figures still fall short of the expected number of detectable CDSs. The performance in discovering uORFs was predictably limited, with only 2,142 uORFs identified on average, displaying poor consistency across replicates (Figure 20H).

Limiting factors for periodicity-based methods

Understanding the factors contributing to the low performance of both tools is crucial. As mentioned, existing ORF detection heavily relies on periodicity, which can be obscured by the low signal to noise ratio associated with triplet periodicity, particularly for small ORFs such as uORFs. I hypothesised that low read depth might contribute to this inconsistency. To test this, I compared the read counts per base for uORFs discovered in only one replicate against those in the equivalent regions in other replicates. As illustrated in Figure 21, the consistently detected ('Intersection') uORFs exhibited the highest read depth. Conversely, in regions with low read counts, the algorithms become more hesitant to determine whether the reads represent noise or genuine signals of translation. This result supports my hypothesis that lower read depth contributes to the inconsistency of predictions. However, it also leads to a challenge: uORFs are not only typically less expressed, and therefore tend to have fewer reads covering them, but due to their short length, they also have a limited capacity for concurrent ribosome occupancy compared to a canonical CDS.

As the mouse sequencing data was from an experiment not specifically designed for ORF identification, the sequencing depth might not be adequate. To rule out variation introduced by the experiment design, I utilised a higher-depth human iPSC dataset (J. Chen et al. 2020). As illustrated in Figure 22, both algorithms demonstrated a higher ORF discovery rate for both CDS and uORFs. However, in terms of uORF detection, both algorithms still failed to deliver consistent predictions across replicates. This suggests that the performance of these ORF detection applications varies significantly with different experimental settings and that current tools are sensitive to the quality of input sequencing data, potentially limiting their reproducibility.

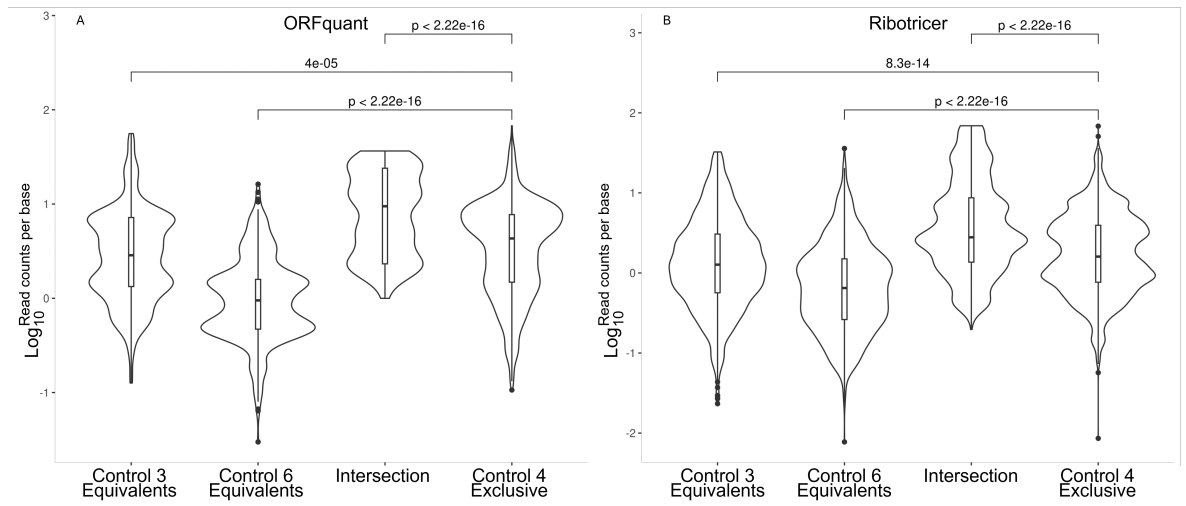


Figure 21: **Local read depth across positive and negative uORF samples.** Violin plots comparing Ribo-Seq read counts per base ($\text{log}_{10}()$) among uORF regions identified using ORFquant (A) and Ribotricer (B) across different groups. Specifically, groups include uORFs detected exclusively in 'control 4' ('Control 4 Exclusive'), uORFs consistently detected across all three replicates ('Intersection'), and corresponding genomic regions in 'control 3' and 'control 6' not classified as uORFs ('Control 3 Equivalents' and 'Control 6 Equivalents'). Statistical significance, determined by Wilcoxon rank-sum tests, is indicated for comparisons between groups. Higher read depth correlates with increased consistency of uORF detection, suggesting that low local read depth contributes significantly to the observed inconsistencies.

4.2.2 BYU achieves higher sensitivity and replicate consistency

Encountering challenges with existing ORF identification tools prompted me to develop my own workflow, currently named 'BYU'. The difficulty with existing algorithms arises primarily from their stringent candidate filtering requirements, which are not readily adaptable to diverse experimental setups. The design philosophy of my workflow is therefore guided by the principle of Occam's razor. Rather than beginning with a complex set of assumptions about what constitutes a translatable ORF (e.g. requiring strong triplet periodicity or specific sequence motifs), BYU starts with the simplest hypothesis: any transcribed region with ribosome occupancy above a statistically determined noise threshold is a potential candidate for translation.

This approach, which avoids multiplying initial assumptions, defines the workflow's more inclusive strategy as it initially generates a broader candidate dataset and subsequently constructs an informative feature matrix to systematically evaluate and challenge this simple starting hypothesis. This ensures that potential translation events are not prematurely discarded.

After structuring my workflow, I conducted a benchmark comparison to evaluate its reproducibility and ORF discovery rate. Piping the raw sequencing fastq files into BYU resulted in a significantly higher ORF discovery rate than the other tools (Figure 20A and B, Figure 22A and B). As illustrated in both mouse and human datasets (Figure 20C and D; Figure 22C and D), the number of identified CDSs and uORFs are close to the expected counts.

Meanwhile, a much higher inter-replicate agreement was achieved comparing to the other tools. As shown in Figure 20C,E and G, the κ of BYU in the CDS cohort is 0.238 which is the only positive value. As mentioned early in this section. Kappa is calculated to represent the proportion of agreement beyond what would be

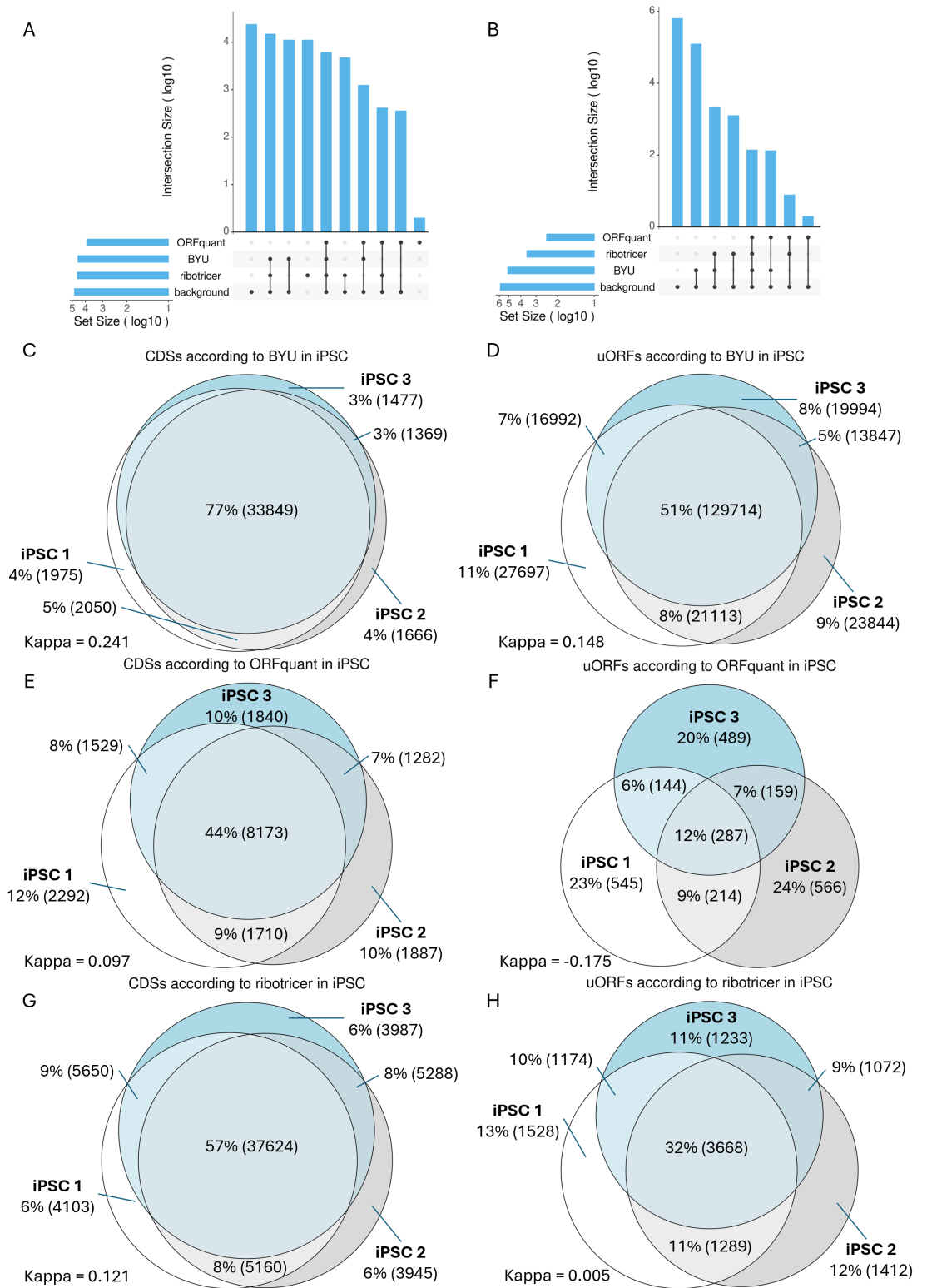


Figure 22: ORF identification rate comparisons in the human project. Comparison of ORF identification using the proposed BYU workflow with two state-of-the-art algorithms, ORFQuant (Calviello, Hirsekorn, and Ohler 2020) and ribotricer (Choudhary, Li, and D. Smith 2020), based on Ribo-Seq and RNA-Seq data from three replicates (iPSC 1, 2, and 3) in human induced pluripotent stem cell experiments conducted by Chen et al. (J. Chen et al. 2020). Left panels illustrate CDS recovery rates, and right panels depict uORF discovery rates. A) and B) present upset plots comparing ORF discovery among the three different methods within CDS and 5' UTR regions respectively. The background group represents CDSs and uORFs identified using the BYU workflow without filtering based on the Ribo-Seq read depth threshold. C) and D) display Euler diagram showing the inter-replicate agreement on identified ORFs using the BYU approach in the CDS and 5' UTR, respectively. E) and F) show similar inter-replicate agreement using ORFQuant, whereas G) and H) utilise ribotricer. For all Euler diagrams, Fleiss' kappa value (-1 to 1) is calculated to statistically quantify the consistency of identification across the three replicates, where a higher value indicates better consistency. Collectively, these comparisons indicate that the BYU workflow achieves more consistent ORF identification across replicates, outperforming other methods in detecting both CDS and uORFs.

1518 expected by chance ($\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$). Higher values indicate stronger agreement. The net κ of 0.238 is not high, as
1519 a value closer to 1 would indicate better agreement. However, the objective here is to compare BYU with the
1520 alternative established predictors, where performance was much worse, showing negative values even for the
1521 CDS group. A negative κ , by definition, indicates poor agreement which is lower than agreement by chance.

1522 As for the uORF cohort, all three algorithms performed less promisingly than in the CDS cohort, while BYU
1523 continued to outperform the other algorithms. This consistency benchmarking revealed that BYU performs
1524 moderately but consistently higher than existing methods. These results align with the current understanding
1525 of the scale of expressed genes and the prevalence of uORFs.

1526 4.2.3 Comparison against the background demonstrates effective noise filtration

1527 To rigorously evaluate the discriminatory power of the proposed workflow, it is necessary to benchmark it not
1528 only against other tools but also against a permissive baseline that represents the maximum potential search
1529 space. For this purpose, I generated a background dataset. This group represents the raw output of the pipeline
1530 without applying the Ribo-Seq read count threshold (the 5 read/STU elbow filter defined in Chapter 3).

1531 Essentially, the Background set retains all genomic segments that show evidence of transcription (RNA-Seq
1532 coverage) and minimal ribosome occupancy(Ribo-Seq coverage). In the mouse dataset, this resulted in a mas-
1533 sive initial pool of approximately 46,921 CDSs and 630,957 uORFs. This dataset serves as a high sensitivity,
1534 but low specificity control, effectively acting as a negative control for the filtering process. Since no enrich-
1535 ment for genuine translation signal has been performed, the Background group captures the full spectrum of
1536 transcriptional activity, including events like transcriptional leakage where ribosomes may transiently associate
1537 with mRNAs without productive elongation that are not filtered during the sequencing steps.

1538 The magnitude of this Background set (comprising a threefold excess of CDSs and an order of magnitude
1539 more uORFs compared to the filtered output) highlighted the prevalence of biological noises such as random
1540 ribosome and other RNA binding protein occupancy in raw profiling data. It demonstrates that detecting a
1541 ribosome footprint is necessary but insufficient evidence for functional translation. Consequently, the Back-
1542 ground set provides a critical baseline for calculating the false positive rate. By comparing the final filtered
1543 output to these massive background numbers, we can verify that the BYU algorithm does not indiscriminately
1544 label all detectable ORFs as translating. Instead, the workflow selectively retained only those candidates with
1545 reproducible, high-confidence signals while successfully discarded the vast majority of the noises present in
1546 the raw sequencing data.

1547 4.2.4 Comparative validation against gold standard datasets confirms higher sensitiv- 1548 ity

1549 While the internal consistency metrics (κ) and background comparisons demonstrate the workflow's precision
1550 relative to noise, distinguishing False Positives (FP) from True Positives (TP) definitively requires orthogonal

1551 data. Sequencing based predictions ideally shall be validated by physical evidence of the peptide product or
1552 the specific mechanics of translation initiation. In the absence of paired proteomics data generated specifically
1553 for this study, I searched for published datasets that could serve as gold standard. After evaluating available
1554 resources, as described in the previous gold standard section, I selected two studies that represent the state-of-
1555 the-art in initiation mapping and tissue specific deep proteome coverage in CDS. These are the Ingolia et al.
1556 (2011) ribosome profiling-based dataset (referred as the ‘Mouse Ref CDS’ set) and the Lu et al. (2022) mass
1557 spectrometry-based dataset (referred as the ‘Colon Ref CDS’ set). However, at the time of this research, no
1558 equivalent uORF dataset of similar scale could be identified.

1559 The Ingolia dataset is widely regarded as the foundational benchmark for the mouse translome. Derived
1560 from mouse embryonic stem cells, it applied Harringtonine treatment to map translation initiation sites with
1561 high precision. This dataset was chosen to assess the sensitivity of workflows in capturing the landscape of
1562 translation initiation events across the genome. However, ribosome occupancy does not always guarantee the
1563 production of a stable peptide. To address this, I selected the Lu et al. proteome atlas. Unlike earlier proteomic
1564 maps which relied on less sensitive Data-Dependent Acquisition (DDA), Lu et al. employed Data-Independent
1565 Acquisition (DIA-MS) to generate the deepest available quantitative proteome of mouse tissues. This allows
1566 for a rigorous test of whether the ORFs predicted by my computational workflow correlate with physically
1567 detectable proteins in a relevant tissue context.

1568 **Assessment of translation initiation potential (Whole Mouse CDS)**

1569 I initiated the validation process by aligning predictions with the ‘Mouse Ref CDS’ dataset derived from In-
1570 golia et al. (2011). This reference set comprises 10,230 canonical coding sequences empirically validated as
1571 translating in mouse embryonic stem cells (mESCs) by Ingolia et al. Unlike standard Cycloheximide treated
1572 samples, where reads are distributed across the coding regions, Harringtonine footprints provide map of Trans-
1573 lation Initiation Sites. Therefore, it serves as a comprehensive benchmark specifically for where translation
1574 initiation occurs across the mouse genome. Consequently, this comparison serves as a rigorous test of the
1575 workflow’s specificity. It assesses whether the ORFs identified by BYU correspond to genuine, biochemically
1576 active initiation events rather than stochastic ribosome associations or scanning complexes.

1577 The main objective is to determine that BYU workflow can selectively recover the examined initiation land-
1578 scape while excluding non-specific background noise. As illustrated in Figure 23A and B, the workflow suc-
1579 cessfully identified a significantly higher proportion of these validated TISs (Sensitivity: 0.632) compared to
1580 ORFquant (0.245) and Ribotricer (0.226), with a minimal and expected trade-off in precision, where BYU
1581 achieved 0.668, while ORFquant and Ribotricer achieved 0.715 and 0.654, respectively. However, it is im-
1582 portant to note that an increase in the True Positive rate drives a proportional increase in the False Positive
1583 rate. Furthermore, a horizontal comparison across the evaluated tools reveals that the ratio of True Positives
1584 to False Positives remains broadly consistent. This is reflected in the comparable precision scores among the
1585 algorithms. This high overlap with the ‘Mouse Ref CDS’ confirms that the candidates selected by the BYU

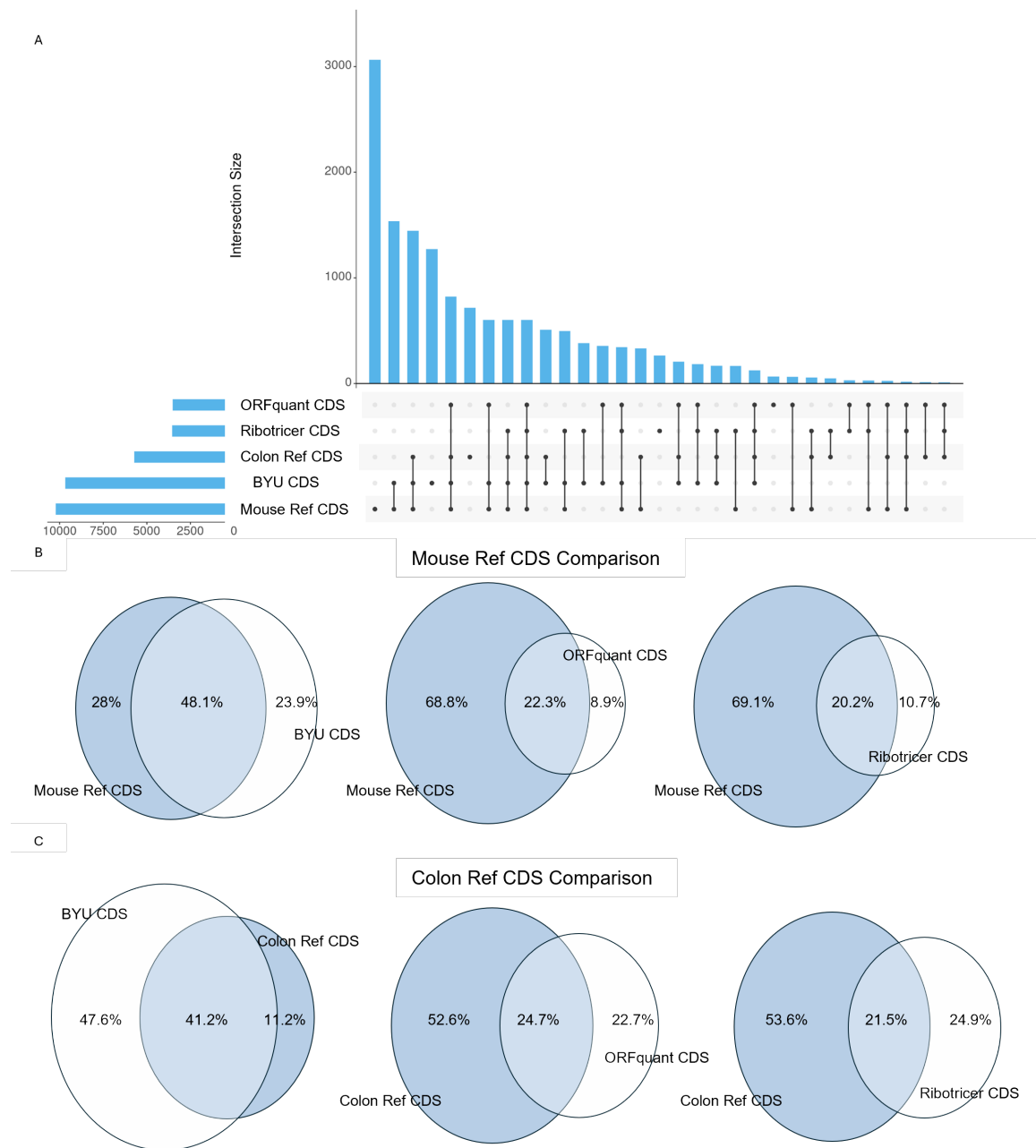


Figure 23: Mouse translated CDS comparison between computational prediction and panel of translating CDS. Comparison of CDS identification using the proposed BYU workflow, ORFquant, and ribotricer against two published ‘gold standard’ resources: a comprehensive mouse translated CDS panel (‘Mouse Ref CDS’) from Ingolia et al. (2011) and a colon-specific translating gene set (‘Colon Ref CDS’) from Lu et al. (2022). A) shows an upset plot visualising the global intersection of translating genes across the three computational methods and the two reference datasets. The vertical bars indicate the size of each intersection set; B) presents Venn diagrams quantifying the overlap between each computational method and the ‘Mouse Ref CDS’ dataset. The intersection region represents the method’s sensitivity (TP). The region unique to the reference dataset represents False Negatives, and the region unique to the prediction represents False Positives; C) provides equivalent overlap analyses using the colon-specific gene dataset (‘Colon Ref CDS’). These comparisons demonstrate that the BYU workflow (blue) achieves a higher overlap with gold standard data than competing tools, indicating superior sensitivity in distinguishing genuine translation events from background noise.

workflow are not random artefacts drawn from the initial Background pool. It is also important to note that the reference dataset is derived from undifferentiated embryonic stem cells, whereas the data analysed in this study originates from differentiated intestinal organoids. The unique candidates identified by BYU likely represent differentiation-specific genes and organoid markers that are transcriptionally silent in the pluripotent state of mESCs. The False Negatives (present in the Ingolia reference but missed by BYU) were found enriched for pluripotency markers and stem cell factors that are naturally downregulated or silenced in the differentiated organoid model.

Assessment of translated peptides in colon (Colon Ref CDS)

Following the validation of initiation mechanics, I sought to verify the specificity of the workflow regarding the final output of translation in colon tissue. To this end, I benchmarked the results against the ‘Colon Ref CDS’ from the Lu et al. (2022) proteome. This orthogonal reference consists of 5,730 proteins explicitly detected in mouse colon tissue using DIA-MS. By restricting the reference to the colon tissue only, which biologically approximates the intestinal organoid model, I aimed to test if the BYU workflow could correctly distinguish the tissue-specific proteome from the general genomic background therefore further testing the specificity of the model calling ORF translation.

Similar to the whole Mouse Ref CDS set analysis, this colon specific set demonstrated that the BYU workflow effectively captures the correct cohort of tissue-specific proteins, achieving the highest sensitivity (0.786) among the tested methods. However, the overall precision of all three algorithms was lower than that observed in the whole mouse gene sets due to the high FP rate for all the tools. It could be due to that mass spectrometry, despite recent advances, remains less sensitive to small, low-abundance, or highly hydrophobic peptides (such as microproteins) compared to sequencing-based methods therefore leading to a much higher FP discovery as many real TP cannot be correctly detected. Furthermore, MS performed on tissue samples cannot fully represent the organoid sample data used in this study. Despite this limitation, BYU maintained a similar middle range performance regarding precision among the three methods. This result discriminated genuine protein-coding potential from the transcriptional noise quantified in the background assessment, proving that the high detection numbers reflect true biological complexity rather than algorithmic inflation.

Gold standard	Algorithm	TP	FP	FN	Sensitivity	Precision
Mouse Ref CDS	BYU	48.1	23.9	28.0	0.632	0.668
	ORFquant	22.3	8.9	68.8	0.245	0.715
	ribotricer	20.2	10.7	69.1	0.226	0.654
Colon Ref CDS	BYU	41.2	47.6	11.2	0.786	0.464
	ORFquant	24.7	22.7	52.6	0.320	0.521
	ribotricer	21.5	24.9	53.6	0.286	0.463

Table 3: **Performance metrics for ORF identification methods.** Comparison of true positives (TP), false positives (FP), false negatives (FN), shown as percentages relative to the union gene set, alongside calculated sensitivity and precision values for three computational ORF identification methods (BYU, ORFquant, and ribotricer) evaluated against two reference datasets (Mouse Ref CDS and Colon Ref CDS). Sensitivity represents the proportion of correctly identified reference CDSs, while precision denotes the proportion of correct CDS identifications among all predictions.

The left control panel allows selection of species and genome build, submission of a gene list and optional upload of sample-specific Ribo-Seq counts. The main view presents several tabs corresponding to common tasks in this thesis: a generic uORFs table; homologous uORFs within species; a genome-wide view of homologous hits on an ideogram; differential translation analyses; and a correlation view between uORF ribosome occupancy and CDS changes.

All persistent data are stored in a single SQLite file to keep the system portable and installation-free. Using SQLite ensures that the exact objects analysed in the thesis can be delivered as a single file with the app. The portal calls the same R functions used in the batch workflow, so results shown interactively match those reported elsewhere.

The screenshot shows the 'uORF Viewer' interface. On the left, there's a sidebar with 'Select or search your species' (set to Mouse), a 'Gene list' (containing Vmm1r184 and Zfp273), 'Submit gene(s)', 'Select your genome build' (set to GRCm39), and an 'Upload your RPF counts file' section. The main area has tabs for 'uORFs', 'Homologous uORFs', 'Homologous seqs on ideogram', 'uORF DE analysis', and 'uORF CDS ribosome occupancy change correlation'. The 'uORFs' tab is active, displaying a table with columns: i_distance, uORF_length, weighted_conservation, weighted_consisg, weighted_cdsoverlap, complete_paralogs, paralogs_uorf_pept, paralogs_uorf, and paralogs_genome. The table shows 10 entries. At the bottom, there's a pagination bar showing 'Showing 1 to 10 of 360 entries' and a 'Download' button.

i_distance	uORF_length	weighted_conservation	weighted_consisg	weighted_cdsoverlap	complete_paralogs	paralogs_uorf_pept	paralogs_uorf	paralogs_genome
4400	270	0.06322962962962962	0	0	0	46	48	5
4558	69	0.199	1	0	0	15	24	3
4400	222	0.07422525252525252	0.6666666666666666	0	0	51	48	5
4438	21	0.01238095238095238	0	0	0			
4236	219	0.1242100456621005	0	0	0	29	40	2
4400	51	0.01468627450980392	0	0	0	9	24	2
4400	36	0.01813888888888889	0	0	0	9	24	2
4307	93	0.1891182795698925	0	0	0	54	24	2
4202	84	0.3224047619047619	1	0	0	9	16	2
4202	6	0.4243333333333334	1	0	0			

Figure 25: Portal interface.

4.4 Summary

Across mouse and human datasets, I evaluated the proposed ORF identification workflow by comparing it with existing tools, ORFquant and Ribotricer. In the mouse colon organoids experiment (moderate depth, not optimised for ORF discovery), the periodicity focused tools under identified both main ORFs and uORFs showing a weaker replicate concordance, particularly for short upstream ORFs. In the higher resolution human dataset, mORF recovery improved for all methods, yet uORF discovery by existing tools remained inconsistent across replicates. In contrast, BYU achieved substantially higher sensitivity while maintaining comparable precision, with better replicate consistency in uORFs.

In this chapter, I evaluated the proposed ORF identification workflow by comparing it with existing tools, ORFquant and ribotricer. I assessed the performance of these workflows using mouse sequencing data. For the mouse project, as the experiment was not specifically designed for ORF identification, the Ribo-Seq data

1654 was at a lower resolution. This limited the performance of both ORFquant and ribotricer in identifying both
1655 uORFs and main ORFs. Additionally, the consistency of existing tools in uORF discovery across replicates
1656 was low, further underscoring their limited effectiveness in uORF identification. In contrast, the proposed
1657 BYU workflow achieved an ORF identification rate closer to the theoretical ORF identification rate, suggesting
1658 a higher reliability. Furthermore, by comparison of the identification results with the gold standard datasets
1659 in the mouse study, the BYU approach achieved the highest sensitivity while maintaining a similar level of
1660 selectivity.

1661 These findings reflect a deliberate design choice: the first pass in BYU is high-recall and does not impose a
1662 strict triplet periodicity gate; candidates are then filtered and prioritised using orthogonal evidence. This staged
1663 approach increases sensitivity without simply lowering selectivity. The resulting dataset provides a relative
1664 robust foundation for the subsequent analyses, which are presented in the next chapter.

1665 Finally, to make the workflows' outputs usable beyond bioinformaticians, I implemented a lightweight R Shiny
1666 portal backed by SQLite. The portal exposes species selection, gene-list queries and interactive uORF tables
1667 and plots, enabling bench scientists to explore results without coding to run the pipeline. This aligns with one of
1668 my personal objectives as a bioinformatician: to make bioinformatics approachable and practically accessible
1669 to bench scientists.

Chapter 5

Systematic Characterisation of uORF Candidates: Sequence Pattern, Evolution Conservation, and Translation Regulatory

Following on Chapter 4, where the BYU workflow achieved a higher sensitivity with a comparable precision and stronger replicate consistency in untranslated regions, I then asked whether properties of the sequences and their contexts can be used to prioritise the resulting candidate ORFs and reduce false positives. The central idea is to decouple sensitive discovery from stricter curation. With a broad candidate set, I then evaluated orthogonal features that may indicate selective pressure or mechanistic plausibility and used those to refine the catalogue.

In this chapter I examined short sequence patterns in uORFs, codon-usage differences relative to matched CDSs, conservation landscapes around uORFs, the presence of homologous uORFs across genes, differential translation under MNK/mTOR perturbation, and finally use supervised models that integrate these features mentioned above. The BYU derived candidates used in following analysis were uORFs identified by the pipeline introduced in Chapters 3 and 4

5.1 Chaos game representation reveals distinct sequence patterns in cell cycle-related uORFs

As discussed in the Introduction, regulatory sequences and structures within 5' UTRs, such as uORFs, have been identified as potent regulators of gene expression. With the identification of uORF candidates, I hypothesised that genes which are regulated simultaneously within the same signalling cascade might exhibit similar sequence patterns within their uORFs. To test this hypothesis, it is crucial to identify a set of genes likely to be co-regulated. The mitotic cell cycle, a meticulously regulated process in eukaryotic species, consists of genes that are tightly synchronised, providing an ideal test set for this analysis (Dominguez et al. 2016). Gene annotations related to the cell cycle were sourced from the study by Daniel et al. (2014), which provides comprehensive gene expression profiles across various stages. In this dataset, genes are categorised into G1-S and G2-M phases based on their peak expression patterns. Table 8 displays a selection from this database. According to this paper, 1,180 genes exhibit significant peak expression, with 304 in the G1-S phase and 876 in the G2-M phase. Subsequently, I mapped the actively translated uORFs identified from the human iPSC dataset to this cell cycle gene set. It is also important to note a potential selection bias in this approach that the iPSC dataset represents a specific, predominantly asynchronous transcriptome, the landscape of translated uORFs may differ from that of a synchronised cell population. The translated uORFs that function may fall below detection thresholds in the baseline iPSC data. However, mapping against this dataset could establish a baseline pool of uORFs with proven translational capacity. Despite this limitation, the analysis revealed that 219 genes from the G1-S phase and 675 genes from the G2-M phase were found to harbour uORFs that is undertranslation. This resulted in the identification of 1,385 uORFs candidates in the G1-S phase and 4,383 uORFs candidates in the G2-M phase.

To investigate whether uORFs co-regulated within specific cell cycle phases share functional sequence patterns, sequence comparison was required. Traditional alignment-based approaches are generally ineffective for this purpose as uORFs are typically short, vary significantly in length, and often lack global linear homology. Forcing alignments on such highly divergent sequences frequently yields uninformative results and can obscure genuine compositional similarities. Then I employed an alignment-free sequence feature extraction approach, chaos game representation (CGR), to generate a feature matrix for further sequence pattern analysis. The CGR algorithm functions as a recurrent iterative system, utilised to produce a fractal representation of symbolic sequences. This concept is illustrated in Figure 6A. In the CGR algorithm, each nucleotide base (A, T, C, and G) is assigned to one of four corners of a square. The process starts at the centre of this square. Subsequently, each dot is placed halfway between its current position and the corner designated for the next base in the sequence. This procedure is repeated iteratively, with each new dot positioned halfway from its predecessor toward the corner corresponding to the next symbol in the sequence. Further details about CGR are provided in the Methods chapter. The feature matrix is generated by counting the hits within grids at a specified resolution (Figure 6B). This matrix represents the usage of 1-mer to k-mer, where 'k' is the resolution of the grid. It then

visualises these 1-mer to k-mer usages within a defined topological structure. For this analysis, I employed the ‘kaos’ package in R to calculate the CGR results, setting the resolution to 3. This setting defines the finest k-mer as a 3-mer, thereby generating an 8×8 feature matrix. While standard k-mer analysis effectively captures motif frequency, it inherently treats the sequence as a bag-of-words, thereby discarding information regarding the spatial distribution and sequential order of nucleotides. In contrast, CGR maps the sequence into a continuous 2D space, preserving global structural information and revealing fractal patterns. This approach enables the capture of higher order sequence properties that are obscured in simple frequency based counting.

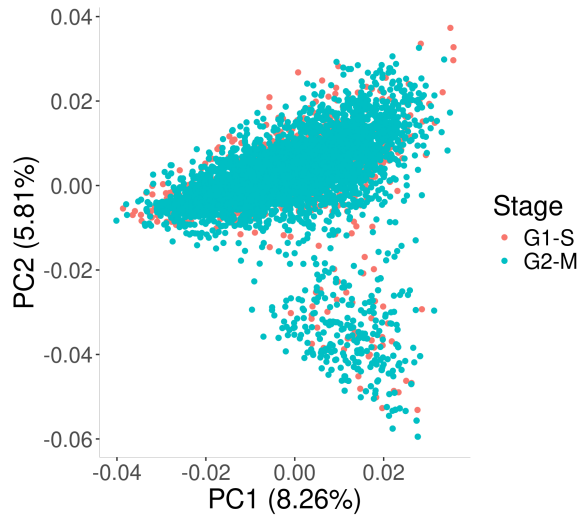


Figure 26: **Principal component analysis representation of the feature matrix.** The scatter plot depicts the PCA results of the uORF sequences feature matrices, with the first principal component (PC1) capturing 8.26% of the variance and the second principal component (PC2) capturing 5.81% of the variance. Each point represents a uORF sequence, colour coded by cell cycle stage: red for G1-S and teal for G2-M. The feature matrix was calculated using Chaos Game Representation (CGR).

Accuracy	Sensitivity	Specificity	Sample Size	Positive case
0.879	0.981	0.866	2770	G1-S stage

Table 4: **Cell cycle-related uORF sequence pattern discriminator performance** This table summarises the performance metrics (accuracy, sensitivity, specificity) of an SVM model trained to discriminate uORF sequences from genes expressed at different cell cycle stages. The training set consisted of balanced uORF sequence data from genes with peak expression in the G1-S and G2-M phases. The accuracy represents the percentage of correct predictions, sensitivity measures the proportion of actual positives that are correctly identified, and specificity measures the proportion of actual negatives that are correctly identified.

With the feature matrix now generated, it is possible to apply machine learning algorithms to discern the unique patterns within uORF sequences across the G1-S and G2-M gene sets. Initially, I employed the principal component analysis (PCA) to visualise the distribution of the 64-dimensional input feature matrices. PCA is fundamentally an orthogonal transformation that converts a set of observations of possibly correlated variables into a set of linearly uncorrelated variables known as principal components. Specifically, principal components can be viewed as a linear equation that includes a series of linear coefficients indicating the projection direction. As illustrated in Figure 26, no clear linear separation of the stages was observed. The first two principal components accounted for only 8.26% and 5.81% of the total variance, respectively. This suggests that the dataset is complex and high-dimensional, with significant information distributed across multiple components beyond the first two. Such complexity underlines the necessity for employing advanced machine learning

1737 algorithms to determine any potential patterns within the data.

1738 For this task, the chosen machine learning algorithm is the support vector machine (SVM), which is well-
 1739 regarded for its effectiveness in separate data in high-dimensional spaces. SVM operates by training a super-
 1740 vised binary discriminator which works by maximising the margin between data classes across all dimensions
 1741 and effectively drawing a hyperplane to separate the classes (Figure 7). While SVM inherently performs linear
 1742 discrimination, the application of kernel functions allows it to manage datasets that are not linearly separable,
 1743 enhancing its adaptability (Figure 8). More details on the general implementation of the SVM algorithm was
 1744 introduced in the Methods chapter. In this study, the input for the SVM consisted of 8×8 feature matrices
 1745 representing the uORF sequences within the G1-S and G2-M gene sets. To ensure a balanced dataset for the
 1746 SVM analysis, 1,385 uORFs from the G2-M stage were randomly selected from a total of 4,383, thus equating
 1747 the number of cases in the G1-S set. Subsequently, the combined total of 2,770 samples was divided into 70%
 1748 for training (1,916 samples) and 30% for testing (854 samples). The model was trained using the ‘ksvm’ func-
 1749 tion from the ‘kernlab’ package in R, employing a radial basis function (RBF) kernel. The penalty parameter
 1750 C was set to 1, and the kernel width parameter σ was set at 0.1, following extensive benchmarking of various
 1751 parameter combinations to optimise prediction accuracy. This SVM model achieved relatively high accuracy,
 1752 correctly labelling 88% of the testing set, with robust sensitivity and specificity scores (Table 4).

Accuracy	Sensitivity	Specificity	Sample Size	Positive case
0.489	0.482	0.497	5000	Group 1

Table 5: **Null model for sequence pattern discriminator.** This table summarises the performance metrics (accuracy, sensitivity, specificity) of a null SVM model trained using randomly selected CDS sequences from the same gene set. The null model serves as a baseline, verifying that the SVM model does not overfit the data and does not capture artificial sequence patterns.

1753 To rule out the possibility of the SVM settings overfitting the data and leading to pseudo-high accuracy, I also
 1754 trained a null model where 5,000 randomly selected CDSs in M-like transcripts were assigned to two groups
 1755 and split into training and testing sets in the same manner as the cell cycle stage model (Table 5). This model
 1756 was trained with identical SVM settings. The results from this null model showed an accuracy of around 49%,
 1757 as expected, suggesting that the high performance of the primary SVM model is not due to overfitting but
 1758 reflects genuine predictive capability as shown in Table 4. However, while the model effectively identifies
 1759 certain patterns within the sequences, it does not provide explanations of these patterns. As mentioned earlier
 1760 in this section, the feature matrix generated from the CGR represents k-mer frequency, with k set to 3 in this
 1761 analysis. Given that a 3-mer in genomic terms often corresponds to a codon, this may suggest that specific
 1762 codon usage patterns may be embedded within the sequences and potentially be picked up by the SVM. In
 1763 brief, the CGR SVM based sequence pattern analysis indicates that uORFs in G1-S and G2-M genes carry
 1764 sequence signals consistent with cell-cycle related regulation.

5.2 Codon analysis reveals phase-dependent AT3/GC3 biases in uORFs

Each amino acid is encoded by a set of 3-mer nucleotides known as codons. While 64 possible codon combinations encode 21 amino acids, there is redundancy in codon choice for the same amino acid, referred to as synonymous codons. The third nucleotide in a codon, often called the ‘wobble’ position, plays a crucial role due to its flexibility in base pairing (Fabbri et al. 2021). The code is structured such that substitutions at this position often preserve the amino acid identity (4-fold degenerate sets such as GCN for alanine), thereby contributing to codon redundancy. Although this is not universal for example, CAU and CAC encode histidine whereas CAA and CAG encode glutamine. This feature permits certain mismatches during translation, allowing a single tRNA to recognise multiple codons through wobble base pairing, although the efficiency of translating non-optimal codons by one tRNA may be reduced (Rudolph et al. 2016). Furthermore, synonymous codons are not used with equal frequencies, and different combinations of codons can influence the structure and stability of mRNA, as well as the availability of tRNA, thus affecting overall gene translation (Chaney et al. 2017). Given that the analysis of uORF sequence patterns suggested potential codon usage biases embedded within these sequences, and considering my research interest in exploring how uORFs might influence overall transcript expression, I have integrated codon usage analysis into the ORF characterisation workflow. To analyse codon usage differences, I have defined the following calculations:

1. Codon frequency within amino acid groups for each transcript:

For each transcript pair, consisting of a uORF and its corresponding main open reading frame (mORF also known as CDS, here denoted as mORF to avoid the conflict abbreviation *c* of Codon), compute the codon frequency within amino acid groups separately. Define $n_{m,i}(c)$ and $n_{u,i}(c)$ as the counts of codon c in the mORF and uORF of the i -th transcript, respectively. Also, let $\sum_{c' \in aa} n_{m,i}(c')$ and $\sum_{c' \in aa} n_{u,i}(c')$ be the total counts of all codons for the amino acid aa in the mORF and uORF of the i -th transcript. This normalization focuses on the relative use of codons per amino acid:

$$f_{m,i}(c) = \frac{n_{m,i}(c)}{\sum_{c' \in aa} n_{m,i}(c')}, \quad f_{u,i}(c) = \frac{n_{u,i}(c)}{\sum_{c' \in aa} n_{u,i}(c')} \quad (5.1)$$

For example, if c represents the codon CTT for Leucine, c' would iterate over the group of codons (CTT, CTC, CTA, CTG) to calculate the total occurrences of Leucine codons in the i -th transcript.

2. Compute codon frequency differences for each codon in each transcript:

Calculate $\Delta f_i(c)$ to represent the frequency differences for codon c between the uORF and mORF of the i -th transcript:

$$\Delta f_i(c) = f_{u,i}(c) - f_{m,i}(c) \quad (5.2)$$

3. Aggregate differences for each codon across all transcripts:

Sum the frequency differences $\Delta f_i(c)$ for each codon c across all transcripts to determine the overall codon usage difference for that specific codon $\Delta F(c)$, where T is the total number of transcripts analysed:

$$\Delta F(c) = \sum_{i=1}^T \Delta f_i(c) \quad (5.3)$$

As suggested by the SVM results discussed in the previous chapter, uORFs in G1-S and G2-M phases exhibit distinct, yet unidentified sequence patterns. This led me to hypothesise that codon usage bias might be a contributing factor. To explore this, I compared the codon usages between the two phases (Figure 27). The analysis revealed a pronounced pattern: codons with Adenine or Thymine at the third position ('AT3') were predominantly observed in G2-M genes, while G1-S uORFs showed a preference for codons with Guanine or Cytosine at the third position ('GC3'). This differential codon usage implies that the sequence patterns identified by SVM may be influenced by underlying codon biases. Prior studies have shown that GC3 codons can increase mRNA stability, leading to longer half-lives and potentially higher expression levels (Buhr et al. 2016; Hia et al. 2019). Given that uORFs can regulate the expression of main coding sequences (CDS) and exhibit specific sequence patterns potentially driven by codon preferences, these findings raise important questions. It remains unclear how codon biases within uORFs might influence the expression of the main coding region. Furthermore, it is yet to be determined whether the observed codon preferences apply to entire transcripts or if uORFs and main ORFs have different preferences.

I then calculated the aggregated codon usage differences $\Delta F(c)$ between identified uORFs and CDS for both G1-S and G2-M genes, visualising the results in histograms (Figure S6). The histograms indicate a preference for AT3 codons in CDS sequences, while uORFs were predominantly enriched for GC3 codons in both G1-S and G2-M genes, with a more pronounced bias observed in G2-M genes. However, it is important to justify this finding within the context of the mRNA global nucleotide architecture. Nucleotide compositions are not uniform across transcripts where it often exhibits a positional gradient, with 5' leader sequences have a higher GC density compared to the downstream coding regions (Gao et al. 2021). Therefore, the observed GC3 enrichment in uORFs could potentially reflect this local background bias (positional effect) rather than a specific selection pressure on codon choice. To distinguish between genuine translational selection and this underlying background variation, the influence of positional nucleotide content is analysed in the subsequent section. Meanwhile this also prompted me to investigate whether such uORF-CDS codon biases are widespread across all transcripts. Performing the same analysis on all uORFs and their corresponding CDS revealed a similar codon preference pattern (Figure 28). The prevalent use of GC3 codons in uORFs suggests a potential regulatory function, possibly modulating the pace of ribosomal translation to affect downstream gene expression. This aligns with existing research suggesting that GC3 codons contribute to more stable mRNA structures, potentially regulating translational initiation and elongation processes (Hia et al. 2019; Gillen, Giacomelli, et al. 2021; Gillen, Waldron, and Bushell 2021).

To determine if the codon usage differences between CDSs and uORFs are statistically significant, a Null model was generated. First, each CDS was repeatedly shuffled to create a pseudo-sequence set retaining the same nu-

cleotide composition, with 300 permutations per sequence. We then compared the codon usage in the original mORF with that of the permuted set (Figure 29). The Null model conformed to a normal distribution according to the Shapiro-Wilk test. As illustrated in Figure 29, when compared with the Null model, the codon biases in mORFs/uORFs mostly fall outside the 95% confidence interval, indicating a significant difference. By correlating codon usage patterns with further translational efficiency and homologous analysis within the context of uORFs, my workflow seeks the opportunity to uncover potential novel regulatory layers that contribute to the complexity of gene expression.

However, the observed differences in codon preference between uORFs and CDSs may be attributed to the general enrichment of GC content towards the 5' end of transcripts (L. Zhang et al. 2004; Gao et al. 2021). To determine whether the codon preferences are driven by codon optimisation rather than general GC enrichment, I visualised the average GC content variation across the entire mouse transcriptome. This was achieved by employing a sliding window technique, calculating local GC content ratios across various transcript regions using a 10 nt window with a 1 nt step size. Canonical CDSs are typically orders of magnitude longer than uORFs. Consequently, a direct length matched comparison is unfeasible as it would bias the analysis to only the shortest CDSs. To resolve this, I adopted a metagene scaling approach where each feature was resampled to a fixed relative width. Given the varying lengths of different transcript components and the random positioning of uORFs within the 5' UTR, I calculated the average GC content ratio for each component individually, resampling each to a fixed size: 1000 nt for CDS, 150 nt for the 5' UTR, and 50 nt each for uORFs and the non-coding regions surrounding uORFs within the UTR. This normalisation allows for the direct visualisation of GC density trends across regions, independent of their absolute physical lengths. The results (Figures 30A and B) indicate an increase in GC ratio towards the 5' end of transcripts. As shown in panel B, the entire 5' UTR exhibits a higher GC content compared to the CDS. However, dissecting the UTR further into uORF and 'truly non-coding' regions reveals that the non-coding regions have significantly higher GC content than the coding regions, as shown in panel A. Therefore, the enrichment of GC in 5' UTR region might lead by the non-coding region rather than the uORFs.

Additionally, to mitigate the impact of outliers on the observations, I made a series of intra-transcript comparisons of GC content ratios, comparing CDS to their corresponding uORFs, non-coding regions, and the entire 5' UTR. These comparisons (Figures 30C to E) reveal that many CDS and their associated uORFs share similar GC ratios, indicating a correlation. In contrast, the non-coding regions and the entire 5' UTR tend to be more GC-rich compared to the CDS. Further visualisation using violin plots in Figure 31 confirms that CDS and uORFs generally exhibit similar GC content, whereas non-coding regions are significantly more GC-rich. Put differently, within the 5' UTR, the uORF segments are, on average, less GC-rich than the surrounding 5' UTR background. These findings indicate that the GC3 codon preference observed in uORFs is not merely a consequence of global GC enrichment at the 5' UTR level, but reflects a codon-position-specific pattern.

5.2.1 Tumour-associated synonymous mutations in uORFs exhibit an AT3 codon bias

With recent research suggesting that synonymous mutations may drive tumorigenesis (Supek et al. 2014; Buhr et al. 2016), I explored whether changes in codon preference due to synonymous mutations in the main coding regions of cancer cells also occur in uORFs. To investigate this I then acquired the human mutation data from the cancer genome atlas pan-cancer analysis project (TCGA) which investigates molecular characterisation of over 20,000 primary cancer and matched normal samples spanning 33 cancer types (The Cancer Genome Atlas Research Network et al. 2013). All synonymous mutations from the 5' UTRs across all TCGA projects were compiled and mapped onto human uORF candidates identified from the iPSCs (J. Chen et al. 2020). Given that mutations in the 5' UTRs are not typically annotated as synonymous, I developed a pipeline to match and filter synonymous single nucleotide variants (SNVs) using the mutation annotation format (MAF), retrieved via the 'TCGAbiolinks' package in R. Through this pipeline, I identified 3,337 synonymous mutations within uORFs. Subsequently, I analysed these mutations to assess changes in codon usage preferences. The results were visualised using histograms (Figure 32) which align with the methodology used in prior comparisons between CDS and uORF codon usage. In these histograms, positive values indicate a codon preference in tumour cells, while negative values suggest a preference in normal samples. The analysis revealed an enrichment of AT3 codons in the uORFs of tumour cells. Such codon changes might influence the stability of the 5' UTR, potentially impacting the overall expression patterns within the tumour transcriptome.

This systematic analysis of codon usage within uORFs and corresponding main coding sequences (CDS) across various transcripts has elucidated notable differences in codon preferences, which are likely to influence mRNA stability and translational efficiency. The observed preference for GC3-rich codons in uORFs may indicate a regulatory role, potentially modulating the translational landscape and impacting gene expression. By employing a robust statistical framework to compare the observed codon usage against a null model, this codon research substantiates that the observed differences are statistically significant and not attributable to random variation. Meanwhile, the GC ratio analysis at the level of CDS, uORF level revealed that the codon preference is not due to the GC enrichment towards the 5' end of the transcript. Additionally, the examination of synonymous mutations in human cancer genomes highlights how codon usage biases could influence tumour biology. In summary, uORFs and their paired CDSs differ systematically in codon usage with a phase-dependent AT3/GC3 biases beyond the average 5' UTR GC content alone. Together, these findings underscore the importance of codon usage analysis in uncovering insights into gene regulation and expression patterns across diverse biological contexts and potentially in cancers.

5.3 Conservation landscape analysis highlights the evolutionary selection of functional uORFs

A uORF candidate conserved across different species is more likely to be genuine, as its sequence has been preserved through speciation and evolution (Zuccotti et al. 2020), therefore encoding functional peptides. To evaluate this, I utilised the PhastCons 35-way conservation score, which provides genome-wide, per-base conservation scores derived from 35 vertebrate species, including mouse, chimp, zebrafish, and dolphin (Siepel et al. 2005). I calculated the arithmetic mean of these scores across each putative uORF locus to determine its overall conservation. While a high conservation score may reflect a high potential for encoding functional peptides and provide insights into the evolutionary conservation of ORFs, setting a rigid threshold on the absolute conservation score to determine evolutionary selection could be overly simplistic. If a uORF exhibits a significantly higher conservation score than the non-coding sequences upstream and downstream of the locus, this could suggest that the uORF is under strong selective pressure (Shatskikh et al. 2020), further supporting its authenticity as a bona fide uORF. To capture this variation, I have termed it the ‘conservation landscape,’ focusing on identifying significant peaks that indicate evolutionary importance. Consequently, I have developed an ORF conservation landscape analysis algorithm to assess whether the identified ORF is surrounded by less conserved sequences. To clarify the operation of this algorithm (Algorithm 1), the corresponding pseudocode is provided.

In my algorithm, a critical step is the delineation of the surrounding regions for each uORF, including both upstream and downstream regions. The size of each region is set to half the length of the uORF plus an additional 100nt, as specified in lines 21 to 24 of my pseudocode (Algorithm 1). This analysis, conducted at the transcript level, accommodates potential overlaps due to alternative splicing in mammalian cells. To manage these overlaps, I developed the recursive function ‘`extend_boundaries(region, direction, chromosome)`’, detailed from line 1 to 15. This function accepts four parameters: `region`, which holds the genomic coordinates defining the upstream and downstream edges of the uORFs’ surrounding region; `direction`, specifying the direction for extension; `cds`, the set of all genomic ranges of CDSs; and `chromosome`, the specific chromosome genomic range. Given that CDS generally exhibit higher conservation scores, this function first checks for any overlap of the defined uORF surrounding regions with any CDSs at the same genomic coordinates. It then calculates the CDS-free coordinates as the `non_cds_region`. If this region is less than 100nt, an additional 100nt is recursively appended to the affected upstream or downstream sequence until the 100nt length threshold is met, while ensuring that extensions do not exceed chromosome boundaries. After establishing the surrounding regions, the algorithm compares the average conservation score between the uORF and its adjacent areas, as outlined from line 25 to 30. If the uORF exhibits higher conservation than its surroundings, it is classified as ‘conserved’; otherwise, it is marked as ‘not conserved’. However, as mentioned before, this method specifically detects purifying selection acting on the internal coding sequence, which marks the uORF producing functional peptides. Conversely, uORFs acting purely as regulatory roadblocks may exhibit low internal

Algorithm 1: Analyse uORF conservation landscape

```

Input:
uorfs = List (sets) of uORFs nucleotides with genomic coordinates.
cds = List (sets) of CDSs nucleotides with genomic coordinates
chromosome = List of sizes of each chromosome boundaries
Output:
landscape = ORF surrounding region conservation status
1 Function extend_boundaries(region, direction, cds, chromosome):
2   non_cds_region  $\leftarrow$  region  $\setminus$  cds
3   if non_cds_region  $\not\subseteq$  chromosome then
4     | non_cds_region.edgedirection  $\leftarrow$  chromosome.edgedirection
5     | return non_cds_region
6   end
7   if length non_cds_region < 100nt then
8     | if direction = 'upstream' then
9       | region.edgedirection  $\leftarrow$  region.edgedirection - 100nt
10    | else
11      | region.edgedirection  $\leftarrow$  region.edgedirection + 100nt
12    | end
13    | return extend_boundaries(region, direction, cds, chromosome)
14  end
15  return non_cds_region
16 end
17 begin
18   landscape = [ ]
19   foreach uorf in uorfs do
20     | half_uORF_length  $\leftarrow$  length uorf  $\div$  2
21     | upstream.edgeupstream  $\leftarrow$  uorf.edgeupstream - (half_uORF_length + 100nt)
22     | upstream.edgedownstream  $\leftarrow$  uorf.edgeupstream
23     | downstream.edgedownstream  $\leftarrow$  uorf.edgedownstream + (half_uORF_length + 100nt)
24     | downstream.edgedownstream  $\leftarrow$  uorf.edgedownstream
25     | upstream  $\leftarrow$  extend_boundaries(upstream, 'upstream', cds, chromosome)
26     | downstream  $\leftarrow$  extend_boundaries(downstream, 'downstream', cds, chromosome)
27     | if upstream  $\cup$  downstream =  $\emptyset$  then
28       | status  $\leftarrow$  'not available'
29     | end
30     | surround_conservation  $\leftarrow$  upstream.conservation + downstream.conservation
31     | if surround_conservation > uorf.conservation then
32       | status  $\leftarrow$  'not conserved'
33     | else
34       | status  $\leftarrow$  'conserved'
35     | end
36     | landscape.append({uorf, status})
37   end
38   return landscape
39 end

```

1927 consensus (low conservation score) yet remain functionally critical. For those regulatory uORFs, the major
 1928 evolutionary constraint could be restricted to the retention of the start and stop codons, with the intervening
 1929 sequence drifted.

1930 Among the 52,784 uORFs discovered in the control groups, 28,951 exhibited an average conservation score be-

low 0.5. Following the landscape analysis, approximately 39% of these were classified as relatively conserved. I have picked two examples of uORFs with relatively low conservation scores but are more conserved than the adjacent non-coding regions. As shown in Figure 33A, a peak in the conservation landscape has been identified within the uORF of the tumour necrosis factor receptor (TNFR)-associated factor 6 (TRAF6). TRAF6 plays a crucial role as a signalling adaptor protein, influencing various pathways, including mTORC1 signalling (Takahara et al. 2020). The overall conservation score of this uORF is 0.13, which is higher than that of the surrounding non-coding regions, with notable peaks at the initiation site. It is noticeable that among all parts of the uORF, the conservation remains at a low level, which may indicate a late stage of evolution. Subsequently, in the downstream differential expression analysis (Figure 38B), this uORF is shown to be up-regulated, while the expression of the CDS remains unchanged. Meanwhile, as illustrated in Figure 33B, the uORF of Ap1S3 displays a conservation score of 1 at the initiation site, whereas the termination region is poorly conserved, scoring 0. This divergent conservation profile may indicate that the less conserved segments were acquired more recently in phylogenetic development. Additionally, Figures 33C and D present uORFs in the same gene E2F1 that starkly contrast in conservation: a uORF with notable conservation and another with nearly no conservation, respectively, where unlike TRAF6 and Ap1S3, the surrounding non-coding regions of the latter are more conserved. These examples were selected to illustrate the range of patterns observed such as uORFs with low overall conservation but higher than flank signal, strong start-codon conservation with weak downstream signal, and contrasting uORFs within the same gene.

In summary, the conservation landscape analysis presented in this study provides a systematic framework to differentiate genuine peptides encoding uORFs from potential false positives by examining their conservation across multiple species. By integrating the PhastCons 35-way conservation scores, this method highlights the evolutionary significance of uORFs and enriches our comprehension of their potential functional implications in genetic expression and regulation. The algorithm developed for this analysis enables a quantitative assessment of the conservation status of uORFs relative to their surrounding non-coding regions. Notably, the identification of significant conservation peaks within certain uORFs, such as those in TRAF6 and Ap1S3, suggests their evolutionary and functional relevance. This analytical framework provides a valuable basis for further investigations into the regulatory mechanisms of uORFs and their potential roles in diverse biological processes.

5.4 Homologous sequence analysis identifies conserved uORFs in co-regulated gene families

Multiple studies have demonstrated that upstream open reading frames (uORFs) can regulate gene translation, primarily by inhibiting the initiation process in the main coding sequences (CDS) (Kumar, Srinivas, and Patankar 2015; I. P. Ivanov, G. Loughran, and Atkins 2008). More recent research has broadened our understanding, revealing that uORFs might also affect CDS translation by either encoding functional peptides or

Algorithm 2: Analyse uORF homologous sequence extended region similarity decay

```

Input:
uorfs = List (sets) of uORFs nucleotides with genomic coordinates
genome = Reference genome sequences
Output:
uorf_homologous_sequence_status = uORF homologous sequence extended region similarity
1 Function blast(region, genome, to, limit = 80%...):
2   sequence  $\leftarrow$  get_sequence(region, genome)
3   matching_regions  $\leftarrow$  match_sequence(sequence, to, limit ...)
4   return matching_regions
5 end
6 Function extend_blast(regions, direction = 'upstream', blast_count = NA, ext_size = 100nt,
   genome, ...):
7   if blast_count = NA then
8     | blast_count  $\leftarrow$  count(regions)
9   end
10  last_blast  $\leftarrow$  regions
11  if direction = 'upstream' then
12    | regions.edgedirection  $\leftarrow$  regions.edgedirection - ext_size
13  else
14    | regions.edgedirection  $\leftarrow$  regions.edgedirection + ext_size
15  end
16  if  $\exists$  regions.edge > genome.edge then
17    | return regions
18  end
19  extended_blast  $\leftarrow$  blast(regions.query, genome, regions, ...)
20  if count(extended_blast)  $\leq$  blast_count / 2 and direction = 'downstream' then
21    | return blast_count
22  else if count(extended_blast)  $\leq$  blast_count / 2 then
23    | return extend_blast(last_blast, 'downstream', last_blast, genome)
24  else if  $|\overline{\text{extended\_blast.edge}_{\text{direction}}} - \overline{\text{last\_blast.edge}_{\text{direction}}}| > \text{ext\_size}/2$  then
25    | return extend_blast(regions, direction, blast_count, genome)
26  else if direction = 'upstream' then
27    | return extend_blast(regions, 'downstream', blast_count, genome)
28  else
29    | return blast_count
30  end
31 end
32 begin
33   uorf_homologous_sequence_status = [ ]
34   foreach uorf in uorfs do
35     | uorf_blast  $\leftarrow$  blast(uorf, genome, genome)
36     | if uorf_blast  $\neq \emptyset$  then
37       | extended_blast  $\leftarrow$  extend_blast(uorf_blast, genome)
38       | if  $\overline{\text{length } \text{extended\_blast}} - \overline{\text{length } \text{regions.query}} > 500\text{nt}$  then
39         | | complete  $\leftarrow$  'incomplete'
40       | else
41         | | complete  $\leftarrow$  'complete'
42       | end
43       | uorf_homologous_sequence_status.append({uorf, extended_blast, complete})
44     | end
45   end
46   return uorf_homologous_sequence_status
47 end

```

creating secondary structures that influence ribosomal activity (McGillivray et al. 2018; Wu et al. 2020). For instance, some studies have identified short peptides from uORFs that interact with small molecules within cells, leading to ribosomal stalling directly at the uORF sites, thereby suppressing further translation of the CDS (Jayaram et al. 2021; J. Chen et al. 2020). Given that sequence similarity has been effectively used to infer homology and related functions among coding genes (Pearson 2013), it is plausible to use similar methods to uncover functional relationships between uORFs.

To achieve this, we utilised the basic local alignment search tool (BLAST) to align all candidate uORFs identified in the control groups of the mouse experiment to the mouse genome. The BLAST runs were executed in parallel using the ‘rBLAST’ R package as the interface within my workflow. I opted for BlastN, setting the word size to 12, restricting results to a minimum of 80% identity, and an expectation value (E-value) of 1. More details about the BLAST algorithm are available in the Methods section. This approach identified 23,395 uORFs that have at least 10 hits, each showing more than 90% sequence identity and an E-value lower than 0.01. The low E-Value accounts for the limited background probability of finding a match by random chance given the database size. An example is provided for uORF (Racgap1) showing homologous hits across the genome (Figure 34A). Racgap1 (Rac GTPase-activating protein 1), a component of the Centralspindlin complex, is implicated in several biological processes such as cytokinesis, cell proliferation, and differentiation (Xu et al. 2024), and is located on chromosome 15. The extensive genomic distribution of Racgap1’s homologous uORF indicates 26 genes encode this homologous sequence. While this recurrence could be explained by paralogy or representing a conserved RNA-binding protein interaction motif, a question was emerged that do these distinct loci share functional coherence? To determine if these uORF-containing genes converge on shared biological pathways rather than being randomly distributed, gene set enrichment analysis (GSEA) was conducted. The ‘clusterProfiler’ R package was applied, selecting the GO biological process as the pathway database with a q-value cutoff of 0.1. The results (Figure 34B) reveal that these potential homologous uORFs are predominantly enriched in biological processes integral to the cell cycle, chromatid segregation, and spindle assembly. Notably, many of these homologous uORFs, such as those in Pthr2 and Nek4 (Figure S4 and S5) are also identified to exhibit significant ribosome occupancy. Pthr2 and Nek4, both implicated as potential oncogenes, play crucial roles in cell cycle regulation (Nguyen et al. 2012; Corpuz, Ramos, and Matter 2020). The uORFs identified exhibit a homologous search distribution similar to that of the Racgap1 uORF.

Repetitive DNA is commonly found throughout the genome (Ruiz-Ruano et al. 2016; Treangen and Steven L. Salzberg 2012). Given this, I hypothesised that uORFs with more than 10 genome hits might belong to larger repetitive sequence structures. To investigate this, I developed a sequence-extending analysis algorithm. To clarify the operation of this algorithm, the corresponding pseudocode is provided (Algorithm 2). This algorithm, based on recursive build-ups as detailed as the function ‘extend_blast(regions, direction, blast_count, extend_size, genome, ...)’.

This function accepts five parameters as well as additional arguments to internal functions: `regions`, holding the genomic coordinates that define the upstream and downstream edges of all the BLAST hits, with `regions.query` specifically referencing the uORF as the BLAST query; `direction`, specifying the direction of the extension; `blast_count`, representing the initial count of

BLAST hits of the original uORF; `extend.size`, indicating the step size for each extension; and `genome`, which specifies the reference genome sequences used in the analysis. This function begins by storing the initial count of BLAST hits (line 7-10). It operates recursively: first, it extends all BLAST hit sequences by 100 nt in the specified direction (line 11-15), ensuring that extensions do not surpass the chromosome boundaries (line 16-18). Subsequently, the extended uORF (`regions.query`) is re-blasted against these extended sequences (line 19). The new BLAST results are then compared to the initial hit counts to assess similarity and count consistency (line 20-30). The process starts with an upstream extension and continues recursively until more than 50% of the sequences exhibit less than 80% similarity. Following this, the function similarly extends downstream until the same similarity threshold is met, concluding the recursive process. After the extension process, if a uORF extends more than 500 nucleotides, it is classified as an incomplete homologous sequence (line 36-44). This analysis revealed that only a small fraction (below 0.5%) of uORFs are part of larger repetitive sequences. This suggests that the vast majority of uORFs are distinct and complete homologous sequences, indicating that evolutionary pressure selectively preserves these uORF regions rather than arbitrarily maintaining large segments of sequence.

This analysis has adeptly leveraged homology to enhance our understanding of the potential functional roles of uORFs and their impact on gene regulation. By identifying extensive homologous sequences across the genome in specific gene families such as *Racgap1*, *Pth2*, and *Nek4*, this study highlights the potential for co-regulation and enrichment of biological processes related to cell cycle control and chromatin dynamics. These findings suggest that uORFs may serve regulatory functions beyond simple translation inhibition, opening new avenues for research into their roles in cellular homeostasis and disease. Furthermore, the sparse presence of uORFs within larger repetitive sequences suggests that evolutionary pressures have selectively preserved these elements for their critical regulatory capacities. Overall, integrating homologous sequence analysis into uORF research offers a preliminary framework for unravelling the evolutionary and functional complexities of these genomic elements.

5.5 Differential expression analysis

The Ribo-Seq and RNA-Seq data not only facilitated the prediction of ORFs but also enabled the exploration of variations in ORF expression across different experimental setups. Specifically, the Ribo-Seq data illustrate differences in ORF expression through variations in ribosome occupancy across ORF regions. Meanwhile, the RNA-Seq data provide insights into changes in transcription levels of each gene. Additionally, with the identification of translated uORFs and CDSs, comparisons of ribosome occupancy between uORFs and corresponding CDSs were conducted. This analysis is critical for determining how uORFs might influence or modulate the expression of the main coding regions. Consequently, I have performed differential expression (DE) analyses at both the transcript level using RNA-Seq data and the ORF level using Ribo-Seq data from the mouse experiments. Reads were aligned with STAR using the parameters listed in Methods Chapter Section 2.2.1, and counts were summarised at transcripts level for RNA-Seq and ORF level for Ribo-Seq prior to DE modelling.

2037 with DESeq2.

2038 Before conducting DE analysis, it is crucial to address any potential batch effects, which can arise from system-
 2039 atic technical variations such as differences in sample preparation times or sources. If left unaddressed, these
 2040 variations can significantly confound analyses by introducing biases, leading to false-positive or false-negative
 2041 findings. To appropriately deal with batch effects, it is essential to consider the experimental design and apply
 2042 correcting methods accordingly. I used sequencing data from the mouse experiments conducted by Dr Joseph
 2043 Waldron et al., as mentioned in Chapter 4.1.1. The experimental design comprised one control group and three
 2044 treatment groups, each consisting of three biological replicates derived from APC^{-/-} KRAS^{G12D} mouse colon
 2045 organoids. This specific genotype delivers a translational landscape characterised by elevated eIF2 α phospho-
 2046 rylation. This organoids simulates a basal activation of the Integrated Stress Response (ISR) node makes this
 2047 organoid model sensitive to perturbations in the translational machinery, thereby making it ideal for investi-
 2048 gation of translation regulatory mechanisms (Adomavicius et al. 2019). Initial PCA visualisation of the raw
 2049 RNA-Seq and Ribo-Seq data (Figure 35A and C, respectively) showed clustering by replicates, which could
 2050 obscure my analysis focus. To correct these batch effects, I used functions provided in the 'DESeq2' package,
 2051 specifying biological replicates as a batch factor in the design matrix. This approach allows DESeq2 to fit a
 2052 generalised linear model (GLM) that accounts for variance introduced by the experimental design, including
 2053 the replicates. More details about this model is covered in Section 2.2.3 in the Methods chapter. After the batch
 2054 effect correction (Figure 35B and D) the adjusted PCA plots indicated that this correction effectively empha-
 2055 sised biological signals pertinent to the experimental treatments, thereby facilitating more accurate downstream
 2056 analyses. After consultation with Dr Joseph Waldron, who performed the original experiments, we identified
 2057 that variations likely arose during organoid preparation. The successful batch-effect correction resulted in sam-
 2058 ples clustering distinctly according to treatment conditions, confirming the appropriateness of proceeding with
 2059 further downstream analysis.

2060 Subsequently, DE analysis was conducted to compare the three treatment groups against the control group
 2061 using both RNA-Seq and Ribo-Seq data at transcript level, using the GLM framework described in Method
 2062 Section 2.2.3. As illustrated in Figure 36, the group treated only with the MNK inhibitor (Panels A and D)
 2063 displayed no significant differentially expressed genes or uORFs. The results confirmed our understanding
 2064 that MNK inhibition alone is insufficient to perturb translation in this scenario, consistent with the observation
 2065 that eIF4E phosphorylation is not the rate limiting step for general synthesis under normal conditions (Knight
 2066 et al. 2021). In contrast, the group receiving the combined MNK+mTOR inhibitors (Panels C and F) showed
 2067 regulatory effects, identifying 4 DE genes and 14 DE uORFs.

2068 To understand the specificity of this gene regulation, it is important to fully review the basal translational
 2069 landscape of the APC^{-/-} KRAS^{G12D} mouse colon organoid model. As mentioned earlier, these organoids exhibit
 2070 constitutively elevated eIF2 α phosphorylation. The high p-eIF2 α inhibits eIF2B, thereby limiting the recycling
 2071 of the eIF2-TC (eIF2-GTP-tRNA_i^{Met} ternary complex) (Jackson, Hellen, and Pestova 2010; Knight et al. 2021).
 2072 Then it may promotes leaky scanning where ribosome bypass the primary TIS, thereby altering the translation
 2073 of transcripts harbouring regulatory uORFs (Dasgupta and Prensner 2024). Recent work has shown that treating

2074 *APC/KRAS* mutant mice with ISRIB (Integrated Stress Response InhiBitor), a molecule that uncouples p-
2075 eIF2 α from eIF2B inhibition, thereby restoring eIF2-TC levels, has reduced tumour proliferation similar to
2076 the MNK+Rapamycin combined treatment (Knight et al. 2021). The differentially expressed uORFs in the
2077 MNK+Rapamycin cohort raises the hypothesis that this combination may mimic the restoration of ternary
2078 complex dynamics or lower the level of cap-dependent translation thereby impacting potential uORF regulated
2079 mRNAs.

2080 Interestingly, the differential expression results also highlighted that the top-regulated genes differ between
2081 RNA-Seq and Ribo-Seq data, indicating a significant variety in translation efficiency across these genes. This
2082 observation prompted me to investigate the correlation between RNA-Seq and Ribo-Seq reads, particularly in
2083 the CDS and uORF regions. To this end, I plotted the RNA-Seq and Ribo-Seq reads for both CDS and uORFs
2084 (Figure 37). The results clearly show a strong correlation between RNA-Seq and Ribo-Seq reads in the CDS
2085 regions. However, for the uORFs, the correlation is less pronounced, likely due to a large proportion of uORFs
2086 exhibiting low expression rates. This finding further encouraged investigation into whether CDS expression
2087 correlates with uORF expression, aiming to understand the underlying mechanisms of translation regulation in
2088 these regions.

$$\text{Translation Efficiency} = \frac{\left(\frac{\text{Ribo-Seq counts}}{\text{ORF size}} \right)}{\left(\frac{\text{RNA-Seq counts}}{\text{Transcript length}} \right)} \quad (5.4)$$

2089 I then examined the correlation between changes in the translation efficiency of CDSs and their corresponding
2090 uORFs in the mTOR inhibitor and dual-targeted treatment groups compared to the control group. Translation
2091 efficiency is calculated by dividing the Ribo-Seq read count per ORF size by the corresponding RNA-Seq reads
2092 per transcript length as shown in this equation. Since uORFs and CDSs are on the same transcripts, the trans-
2093 lation efficiency essentially reflects the Ribo-Seq read counts per ORF size. Therefore, changes in translation
2094 efficiency are determined by the fold change in Ribo-Seq reads. As shown in Figure 38, for both treatment
2095 groups, it is clear that for the majority of the transcripts, the increased ribosome occupancy on uORFs will
2096 lead to a decrease in CDS expression. Each dot in this scatter plot represents a uORF, illustrating the ribosome
2097 occupancy change on itself and its downstream CDS upon different treatments. If uORF and CDS translation
2098 were independent, meaning the translation of uORFs had no impact on downstream CDS, the changes in ribo-
2099 some occupancy would align near the $y = x$ diagonal line. To clarify the relationship further, I included two
2100 reference lines at $y = x \pm 0.5$ to highlight where uORFs and their CDSs have at least a $2^{0.5}$ fold change differ-
2101 ence in expression. For those uORFs above the line of $y = x + 0.5$, it is considered that ribosome occupancy
2102 on the uORF is preferentially higher than on the CDS, potentially negatively regulating CDS expression due to
2103 the treatment; on the contrast, for the uORF that below the line $y = x - 0.5$, under the treatment, ribosomes are
2104 more likely to bind to the CDS under treatment, suggesting that uORFs may positively influence CDS expres-
2105 sion. I then performed gene set enrichment analysis on these diverged groups. As illustrated in Figure 39A,
2106 genes containing uORFs associated with positive downstream regulation were significantly enriched for bio-

logical processes related to development and cellular differentiation, including mesenchyme development and neural cell growth. Meanwhile, the right panel demonstrates that uORFs exerting a negative repressive effect on their CDS were predominantly associated with protein catabolic processes, intracellular transport, and stress responses. This distinct functional separation supports the hypothesis that uORF mediated regulation is not uniform but functionally based on the biological imperative. The enrichment of stress response and catabolic pathways within the negatively correlated group aligns with the repressive model of uORFs regulation (such as ATF4 uORF regulation), where uORFs act as a brake to rapidly suppress energy-intensive housekeeping functions under stress conditions (mTOR inhibition).

5.6 Machine learning

Now, equipped with all the features characterising the uORFs candidates collected through our ORF identification workflow, and with lists of wet-lab proofed ORFs with certain functionality such as those ORF Chen et al. discovered to be critical in cell survival (2020). I am now able to apply some machine learning algorithms to determine if these ORF groups exhibit distinguishable features.

5.6.1 Datasets and features for machine learning

To build and validate the machine learning models, this analysis relies on the datasets established in previous chapters. The primary ORF candidates and their associated features are drawn from the human iPSC dataset (J. Chen et al. 2020) as introduced in Chapter 4.1.2 and the mouse organoid dataset from Bushell lab as detailed in Chapter 4.1.1.

For model training and validation, especially for the task of identifying functionally significant ORFs, requiring ground truth labels. For this, I utilise the experimentally validated datasets introduced in Chapter 4.1.1 and 4.2.4 which are: 1) the whole mouse gene set, a broad Ribo-Seq-based set of 10,230 translated CDSs (Nicholas T. Ingolia, Lareau, and Weissman 2011); 2) the colon gene set, a tissue-specific, MS-based proteome of 5,730 proteins that provides biological specificity relevant to the mouse organoid model (Lu et al. 2022); 3) and the human functional ORF set, a collection of 570 non-canonical ORFs (including 230 uORFs) validated by CRISPR screening as essential for cell growth and viability (J. Chen et al. 2020).

The features used to describe each ORF candidate are those developed throughout Chapter 5 and 4. These are summarised in Table 6. This comprehensive feature matrix forms the input for the models described below.

5.6.2 Feature integration enables clustering of ORF types in UMAP

The underlying concept is to assess if an increase in informational detail can enhance the clarity of data clusters. For a proof of concept (Figure 41) I visualised the ORFs identified in the human dataset from Chen et al. using

Feature	Description
Conservation	Average per-base phastCons score across the ORF.
MFE	Minimum free energy ($\log_2$ -transformed) as calculated by ViennaRNA.
Kozak Score	Score indicating similarity to the consensus Kozak translation initiation sequence.
Phase Score	Ribosome footprint periodicity score .
Read Counts	Total Ribo-Seq reads aligned to the ORF ($\log_{10}$ -transformed).
Coverage	Proportion of the ORF covered by Ribo-Seq reads.
ORF Size	Length of the ORF in nucleotides ($\log_2$ -transformed).
Blast Hit Number	Number of homologous sequences identified by genome-wide BLAST.
Blast Hit Length	Average length of homologous sequences identified by BLAST.
CDS Size	Length of the corresponding main CDS ($\log_2$ -transformed).
CDS Frame	Reading frame of the main CDS relative to the transcript.
ORF Frame	Reading frame of the candidate ORF relative to the transcript.
RNA-Seq	RNA-Seq coverage for the corresponding transcript.

Table 6: **Features used for Machine Learning models.** This table lists all features engineered and extracted by the BYU workflow, which are used as inputs for the Random Forest and UMAP analyses.

UMAP, incorporating different levels of informational detail (2020). The initial UMAP visualisation (Panel A) was generated using a basic feature set, including ORF size, ORF reading frame, corresponding CDS reading frame, and Ribo-Seq read counts. These preliminary features, derived solely from gene annotations and ORF coordinates, are insufficient to fully capture the underlying data structure. By introducing additional features, such as phase score, conservation, MFE, blast hit number, blast average length, and Kozak score, in the second UMAP visualisation (panel B), the model is provided with more information to distinguish between ORFs. This increase in information leads to more distinct and well-defined groups in the second UMAP visualisation (panel B). When I labelled this UMAP with ORF types, a clear pattern emerged where uORFs, CDS (labelled as annotated), and overlapping uORF types each clustered into distinct groups. It is important to note that the features used for this clustering are distinct from the ORF-type labels used for colouring. The UMAP visualisation, therefore, demonstrated that these sequence and experimental features are sufficient to naturally separate ORFs into groups that align well with their known genomic classifications. This process exemplifies how gaining additional relevant information allows for more precise and structured data representations, enabling clearer insights.

5.6.3 Predictive modelling of translated CDSs

Next, continuing with the evaluation of translated CDS (chapter 4.2.4) using the ground truth dataset at workflow stage 2 (constructing the candidate ORF library), the feature matrix characterising ORFs has now been established. As previously described, these features can facilitate a more stringent filtering strategy to enhance the precision of translating ORF identification. Using two selected ground truth datasets, the whole mouse gene set and the colon-specific gene set, I trained two random forest models to identify and prioritise predictive features, subsequently assessing their performance on independent testing sets. Specifically, the ‘control 6’ replicate served as the basis for model training, with identified CDSs labelled as either ‘within gene set’ or ‘not within gene set’. To balance the dataset for training, an equal number of CDSs from each category (up to the

maximum available count in either group) was randomly selected, and subsequently divided into a training set (70%) and a testing set (30%). The random forest models were then trained using the training set and initially evaluated on the reserved testing set. Furthermore, to validate model generalisability, the trained models were evaluated against additional replicates within the control group (ctrl 4 & 3), as well as on a replicate from the MNK inhibition treatment group (EFT508). Since the task here is to identify whether CDSs belong within the ground truth dataset or not, true negatives can be identified as correctly labelled ‘not within gene set’ in the predictions. Therefore, besides the precision ($\frac{TN}{TN+FP}$), specificity ($\frac{TP}{TP+FP}$) and sensitivity ($\frac{TP}{TP+FN}$) can be calculated to assess the selectivity of the models (defined in Chapter 2.2.3).

The model performance metrics are summarised in Table 7. Both models achieved consistently high sensitivity and specificity across all testing groups, for both the whole mouse gene set and the colon-specific gene set. Moreover, as illustrated in the ROC plots (Figure 42A), both models demonstrated strong predictive performance not only in their respective split testing sets but also in independent tests across replicates within and outside their original treatment groups. Feature importance plots for these two models (Figure 43A) revealed that for the whole mouse gene set, ORF conservation emerged as the most significant predictive feature, closely followed by read counts. In contrast, the colon-specific gene set model placed greater emphasis on read counts as the primary discriminative feature. Furthermore, I trained two additional models based on the predictions from existing methods, ORFquant and ribotricer, following the same methodology as for the ground truth datasets. The models demonstrated comparable performance in predicting ribotricer results and slightly lower performance for ORFquant predictions. Feature importance analysis showed (Figure 43B), as expected, that both algorithms primarily relied on the phase score for translation determination.

Ground truth	Testing dataset	TP	FP	FN	TN	Sensitivity	Precision	Specificity
Whole mouse gene	ctrl6 test	1833	478	486	1788	0.790	0.793	0.789
	ctrl4	14844	2250	5572	7432	0.727	0.868	0.768
	ctrl3	14016	2130	5289	6851	0.726	0.868	0.763
	EFT508	13991	1942	5307	6945	0.725	0.878	0.781
Colon gene	ctrl6 test	3106	425	487	3221	0.864	0.880	0.883
	ctrl4	12200	4801	1940	11157	0.863	0.718	0.699
	ctrl3	11534	4925	1754	10073	0.868	0.701	0.672
	EFT508	11348	4466	1984	10387	0.851	0.718	0.699

Table 7: **Performance metrics for RF model evaluation.** This table summarises the performance metrics of a Random Forest (RF) model trained on a balanced training set from the ctrl6 dataset (70% of data). Model then was evaluated on ‘ctrl6 test’ (the remaining, unseen 30% of the ctrl6 dataset) to assess baseline performance, as well as on three independent datasets (ctrl4, ctrl3, and EFT508 where EFT508 is the treatment group representing the generalisation of the model). Model predictions were evaluated against various independent test datasets, comparing true positives (TP), false positives (FP), false negatives (FN), true negatives (TN), sensitivity, precision, and specificity across different ground truth reference datasets (Whole mouse gene and Colon gene).

5.6.4 Predictive modelling of functionally validated uORFs

I then attempted to use my feature matrix to characterise the uORFs identified in experiments validated by wet-lab techniques to see if the information gained is sufficient to distinguish uORFs with specific functions. To generate this data, I compiled the feature matrix for ORFs using a comprehensive set of attributes previously used for ORF visualisation (summarised in Table 6). These features were selected to capture various biological

2185 and sequence characteristics of the ORFs, focusing solely on uORFs this time. I labelled the ORFs based on
2186 their mention as functionally validated in the Chen et al. research paper (2020). Similar to the previous training
2187 procedure, I divided this labelled dataset into a training set (70%) and a testing set (30%). I trained a random
2188 forest model in R, setting the number of trees to 100, to classify the ORFs.

2189 To assess which features were most informative for this classification, I performed a feature importance analysis
2190 (Figure 40). This analysis evaluates how much the model's accuracy (Mean Decrease Accuracy) or node purity
2191 (Mean Decrease Gini) decreases when a given feature is excluded. This analysis revealed that 'Conservation',
2192 'MFE', and 'Kozak score' were among the most critical features for distinguishing these functionally validated
2193 uORFs from the general pool of candidates. This suggests that evolutionary conservation and features related to
2194 translational efficiency, Kozak context and structural stability might be the key determinants of uORF function.

2195 The model's predictive performance was evaluated using a ROC curve, achieving an AUC of 0.91 (Figure
2196 44A). Finally, I used UMAP to visualise the clustering of ORFs (Figure 44B), which demonstrated the model's
2197 capability to distinguish between ORFs with and without experimental validation.

2198 5.6.5 Summary

2199 This machine learning analysis completes the main argument of this workflow. The initial ORF detection
2200 stage, as detailed in Section 4.2 in Chapter 5, was designed for high sensitivity, ensuring a comprehensive list
2201 of candidates. As illustrated this strategy successfully identified a higher number of potential translation events,
2202 reducing false negatives, but at the expected cost of including more false positives than more conservative tools
2203 (ORFquant and ribotricer). To resolve this sensitivity vs specificity trade-off, the machine learning model
2204 functions as a high-dimensional integrator. As described in previous sections in this chapter, a rich feature
2205 matrix to characterise these candidates was built, capturing details of ORF candidates' conservation, codon
2206 usage, and expression. Then, this final section has now brought these elements together, applying machine
2207 learning as a powerful secondary filter. By using the feature matrix to train models, this step effectively curated
2208 the initial sensitive list, providing a potential method to distinguish genuine, functional True Positives from
2209 likely False Positives. This two-stage sensitive detection first, specific-curation latter approach forms the core
2210 of this thesis contribution to ORF identification.

2211 5.7 Conclusion

2212 In this chapter, subsequent downstream analyses of these ORF candidates were conducted. The analysis of
2213 uORF sequence patterns revealed distinct sequences associated with cell cycle regulation. A trained SVM
2214 model effectively distinguished uORFs in genes expressed during different cell cycle stages, based on their
2215 sequence features. This suggests the presence of codon preferences within different gene sets. Further codon
2216 analysis revealed significant biases towards AT-ending (AT3) codons in uORFs within G2-M genes. Addition-

2217 ally, analysis of GC content across transcripts and comparisons between codons in CDSs and uORFs indicated
2218 a significant bias towards AT3 codons in uORFs.

2219 The conservation landscape analysis employed PhastCons scores to reveal that many identified potential uORFs
2220 are selectively conserved compared to their surrounding non-coding regions, indicating that these sequences
2221 are under evolutionary pressure and supporting their functional relevance in gene regulation. Homologous
2222 sequence analysis using BLAST searches identified extensive homologous sequences for specific uORF can-
2223 didates, suggesting potential co-regulation among gene families involved in key cellular processes, such as
2224 cell cycle control. Furthermore, the homologous sequence extending analysis found that most uORFs can-
2225 didates are distinct and complete sequences rather than parts of larger repetitive elements, highlighting their
2226 evolutionary importance.

2227 Differential expression analysis demonstrated that MNK and mTOR inhibition significantly affected gene and
2228 potential uORF expression, with the dual inhibition strategy exerting the most pronounced effects. Finally, ma-
2229 chine learning techniques were used to classify ORFs candidates based on their feature matrices, distinguishing
2230 those with experimentally validated functions from others. Meanwhile, by training the model to learn from the
2231 ORFquant and ribotricer, it has been shown that indeed those algorithms are prioritising the phase score than
2232 the other features. And taking together, the feature-based BYU models generalised across replicates, recovered
2233 ground-truth CDSs and distinguished the validated uORFs, with conservation, read depth and Kozak context as
2234 leading predictors.

2235 Overall, this chapter underscores the complexity of uORF regulation in gene expression and the efficacy of
2236 the proposed workflow in ORF identification and characterisation. These findings reveal that uORF candidates
2237 may play significant roles in translational regulation, are subject to evolutionary conservation, and can influence
2238 gene expression through codon biases and sequence homology. And the take-home message from this chapter is
2239 that the proposed workflow substantially improves the sensitivity and specificity of potential ORF detection and
2240 offers powerful downstream tools for characterising translational regulation across diverse biological contexts.

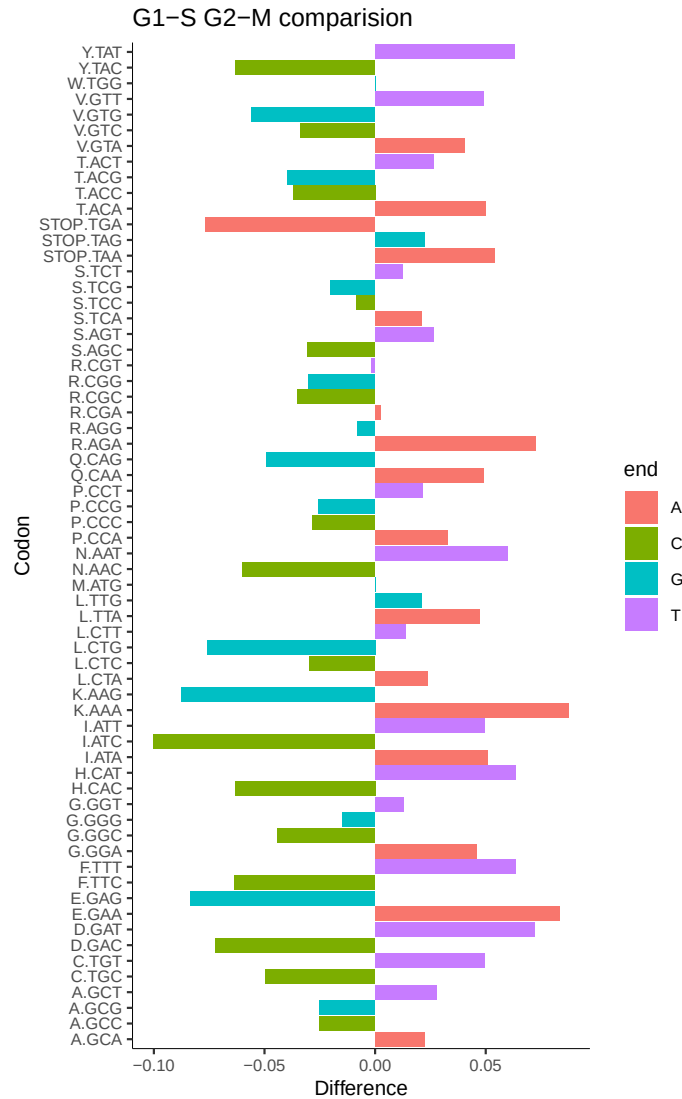


Figure 27: **Codon usage differences between G2-M and G1-S uORFs.** The codon frequency differences were determined by calculating the differences in codon occupancy for each encoded amino acid across all transcripts. A positive value indicates that the codon is favoured in the G2-M set, while a negative value suggests it is preferred in the G1-S. The colour coding represents the codon end, as illustrated in the figure legend. The difference scale measures the degree to which the relative frequency of a codon is $x\%$ higher or lower in the G1-S compared to the G2-M set. This plot suggests that uORFs in G2-M genes have a higher preference for AT3 codons compared to those in G1-S genes.

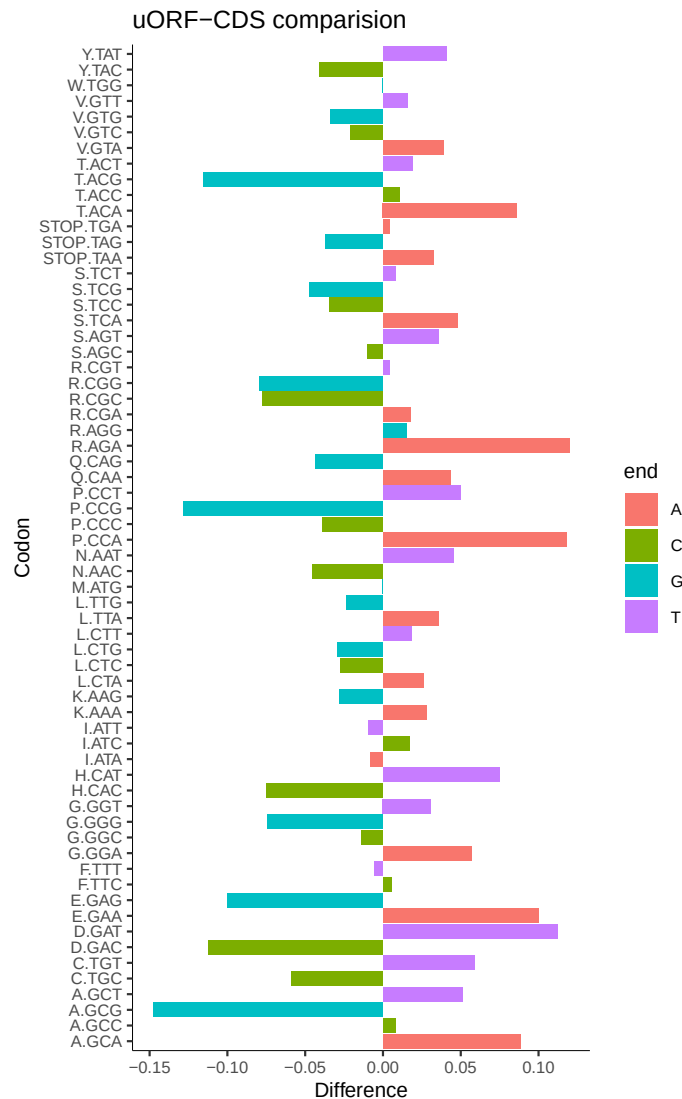


Figure 28: **Codon usage differences between uORFs and CDSs.** The codon frequency differences were determined by calculating the differences in codon occupancy for each encoded amino acid across all transcripts. A positive value indicates that the codon is favoured in the main ORF, while a negative value suggests it is preferred in the uORF. The colour coding represents the codon end, as illustrated in the figure legend. The difference scale measures the degree to which the relative frequency of a codon is x% higher or lower in the main ORF compared to the uORF. This plot suggests that uORFs have a higher preference for GC3 codons compared to their corresponding main CDS.

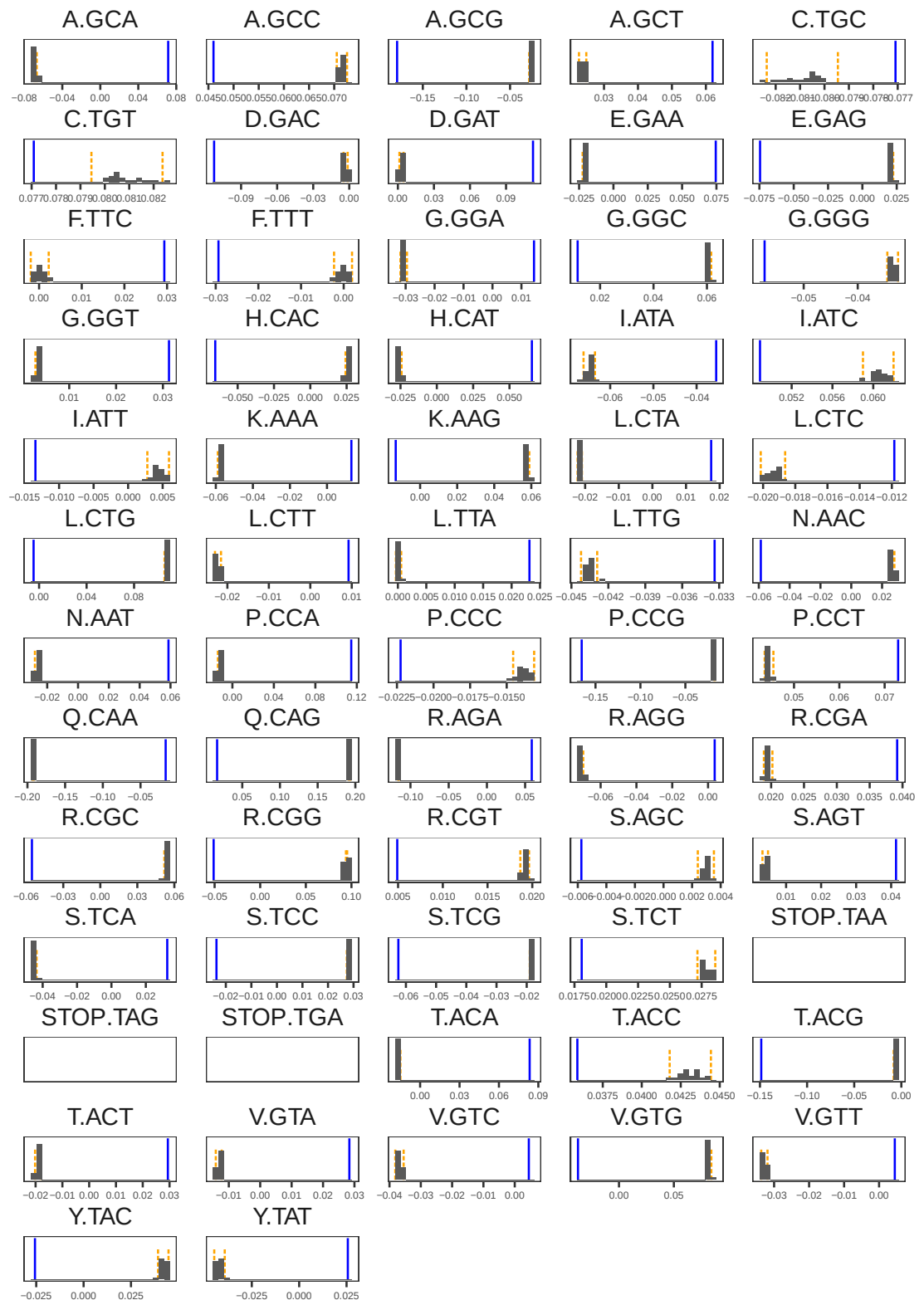


Figure 29: Null model for the codon usage comparison. The null model was created by shuffling the original CDS sequence 300 times, ensuring the same nucleotide frequencies were maintained. This model was then used to assess the codon usage in uORFs. In the corresponding codon graphs, an orange dashed line represents the 95% confidence interval, while a blue line shows the codon usage bias. The X-axis details changes in codon usage, the Y-axis quantifies the frequency of these biases, and each graph's title specifies the codon and its encoded amino acid. This analysis suggests that the codon usage preferences in uORFs are significant, rather than occurring by chance.

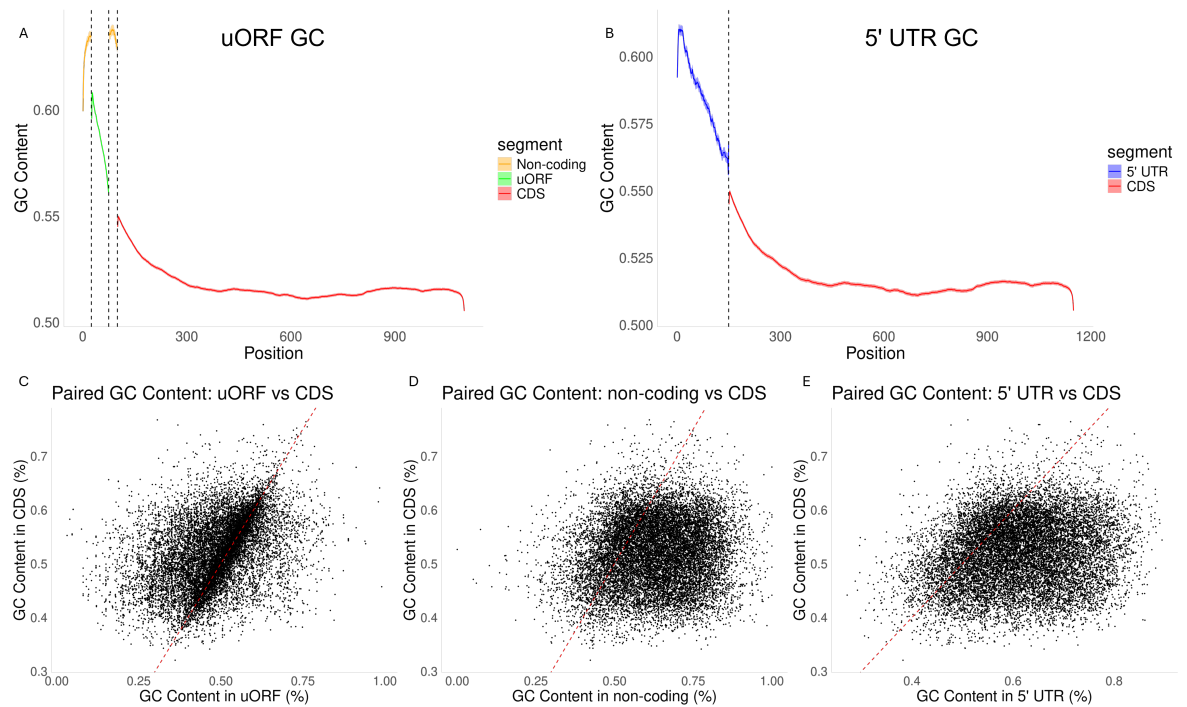


Figure 30: GC content trends across mouse transcriptome and within transcripts. Average GC content variations are illustrated across mouse transcripts, analysed using a sliding window approach. In A and B, the averaged GC content ratio change is at the level of the entire transcriptome, calculated using a sliding window approach. For coding sequences (CDS), a 10 nt window was applied with a 1 nt step, and the results were resampled to 1000 nt for presentation. The same sliding window settings were used for 5' UTRs, resampled to 150 nt, and for uORFs and surrounding non-coding regions, resampled to 50 nt each. The shaded areas represent the 95% confidence interval. The A is showing the uORF and its flanking non coding region GC content trend while B is showing the whole 5' UTR GC level trend across the position. C to E focus on the paired components within the same transcripts, comparing how GC content may vary within individual transcripts, providing insights into intra-transcript GC content dynamics. The analyses reveal that GC content varies significantly across different transcript regions, particularly between coding and non-coding regions. The GC content of uORFs and the main CDS regions are largely correlated, while the non-coding regions in the 5' UTRs are more GC-rich.

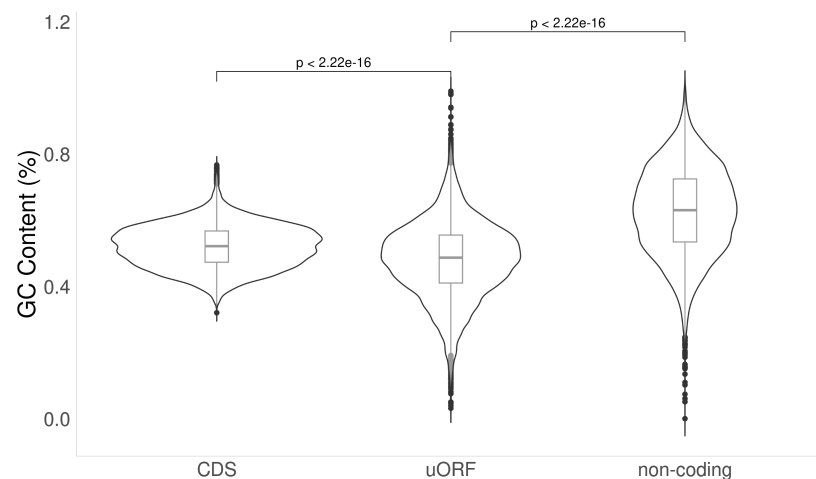


Figure 31: Violin plot of GC content distribution. Distribution of GC content within CDS, uORFs, and non-coding regions of the 5' UTR, as derived from the data presented in Figure 30. GC content of non-coding regions within the 5' UTR is significantly higher compared to CDS and uORFs, indicating that the overall elevated GC levels in the 5' UTR primarily originate from non-coding sequences rather than uORFs.

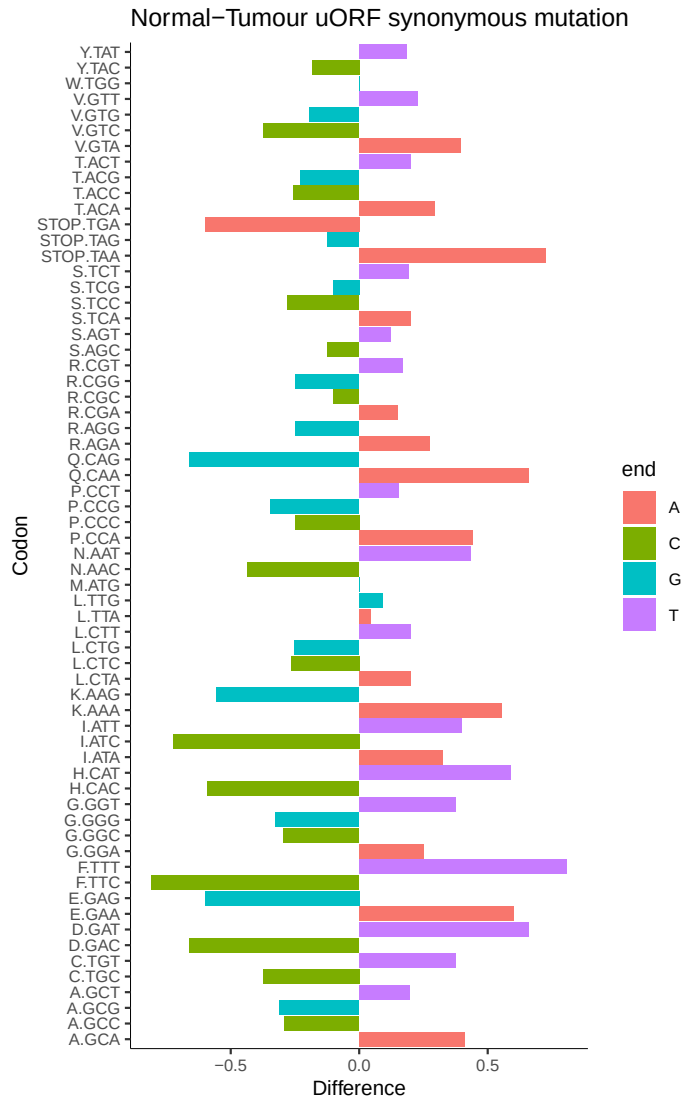


Figure 32: **Synonymous codon mutations in all TCGA projects.** There are 3337 synonymous mutations in 5' UTR among all TCGA projects. Codon usage differences were calculated by counting the synonymous mutations identified in a pre-built uORF library, using data from human induced pluripotent stem cells (iPSCs) reported by Chen et al. (2020) and annotated according to TCGA. A positive value indicates that the codon is favoured in the tumour transcripts, while a negative value suggests it is preferred in normal samples. The colour coding represents the codon end, as illustrated in the figure legend. This plot suggests that uORFs in tumour samples have a higher preference for AT3 codons compared to normal samples.

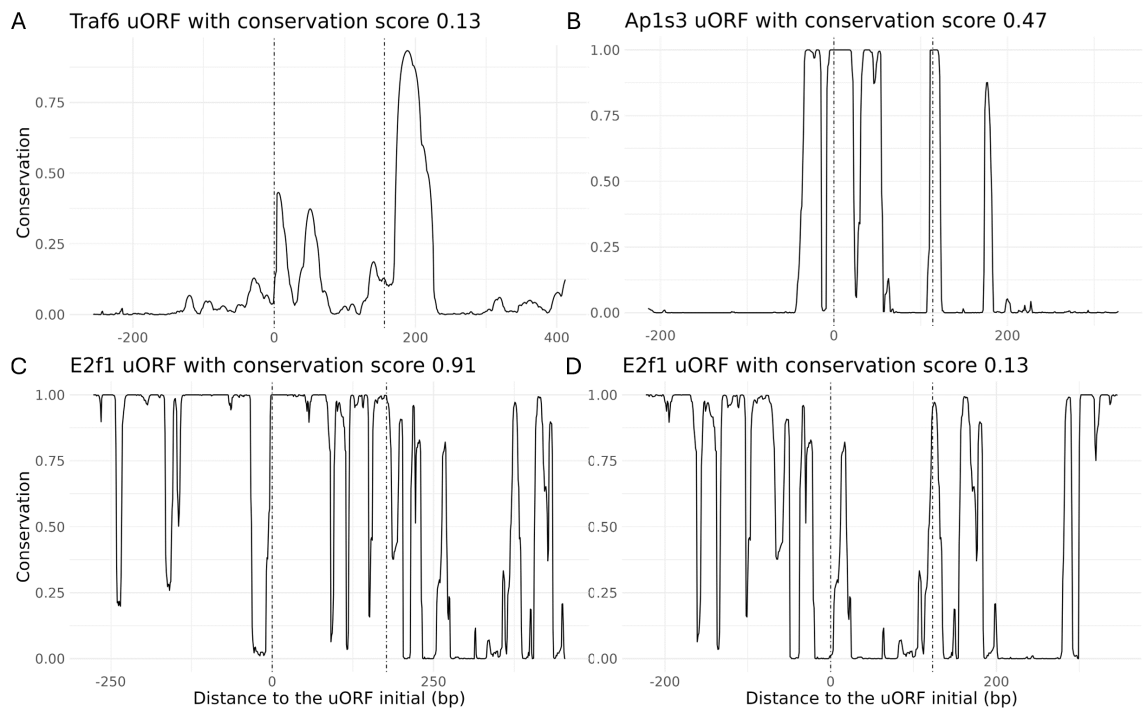


Figure 33: **Conservation landscape of uORFs.** In each panel, the X-axis represents the nucleotide position in the transcript relative to the start of the uORF, and the Y-axis displays the PhastCons 35-way conservation score (Siepel et al. 2005). The ORF region is demarcated by dashed lines. A) and B) feature uORFs with low conservation scores (< 0.5) that are still significantly higher than those of the surrounding non-coding regions. Specifically, A) shows a uORF of Traf6 with a conservation score of 0.13, and B) presents a uORF in Ap1S3 with an average conservation score of 0.47. C) illustrates a uORF with a high conservation score, serving as a positive example, while D) shows a uORF with a low conservation score that has not been identified as significantly conserved; both uORFs are located in E2f1.

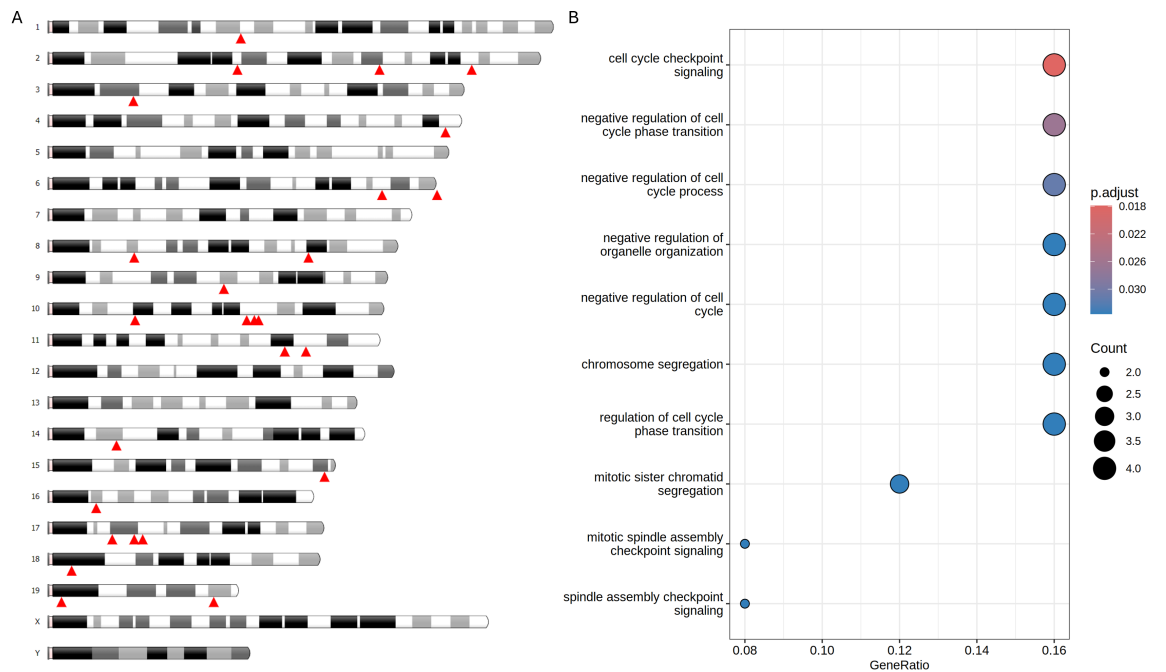


Figure 34: **Homologous hits of an uORF (Racgap1) on mouse chromosome ideogram with corresponding GSEA.** Representation of homologous sequences of a uORF from the Racgap1 gene across the mouse genome (A), and associated Gene Ontology enrichment analysis (B). In A, red triangles indicate genomic regions sharing greater than 90% sequence identity with the Racgap1 uORF. B presents the significantly enriched GO biological processes among the genes harbouring these homologous uORF sequences.

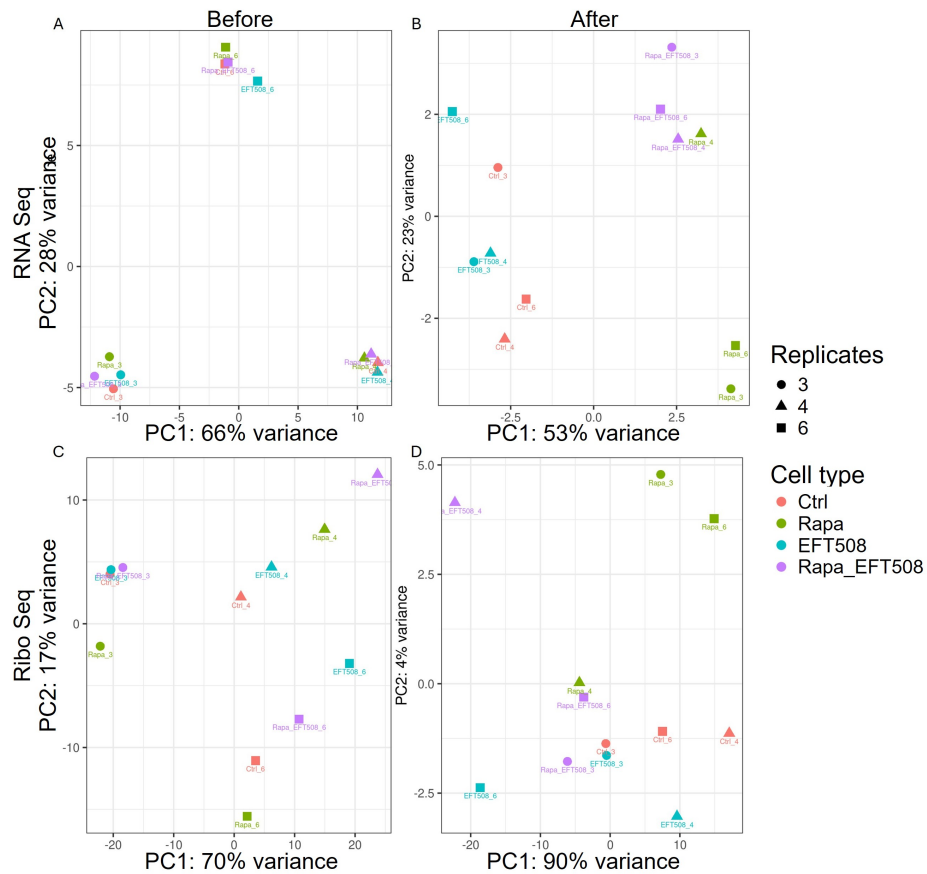


Figure 35: **Batch effect visualisation of RNA-Seq and Ribo-Seq data.** Principal Component Analysis (PCA) illustrating batch effects within RNA-Seq (top row) and Ribo-Seq (bottom row) datasets. A) and C) display the original un-normalised data, revealing substantial clustering driven by batch effects related to replicates. B) and D) show the data post-normalisation, processed using the ‘DESeq’ package in R to mitigate batch effects. Different cell types are differentiated by distinct colours, and replication is denoted by various shapes.

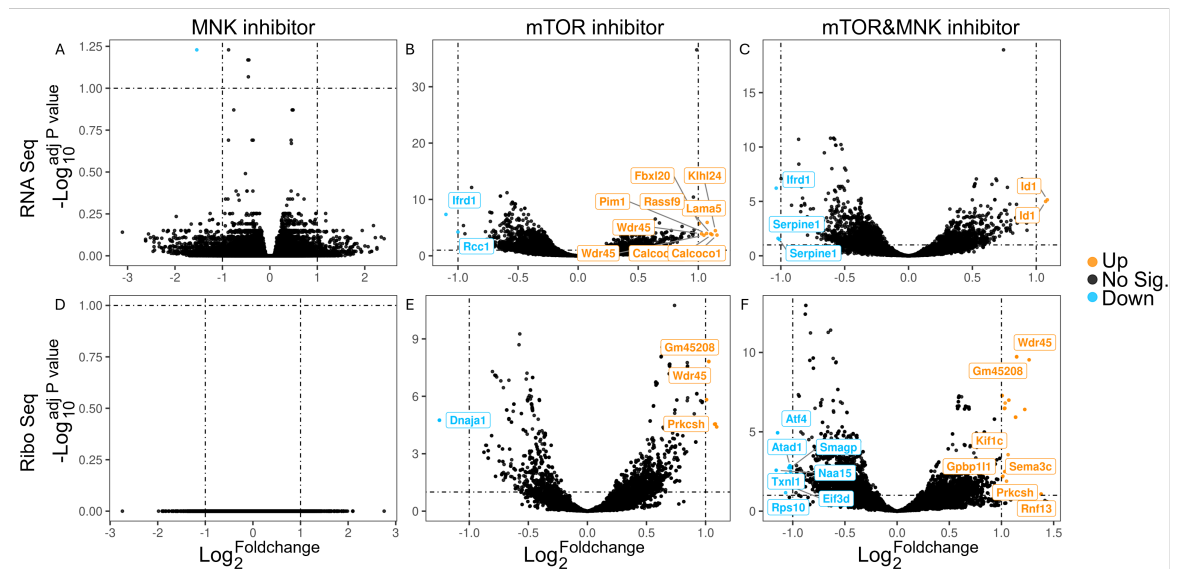


Figure 36: **Differential expression analysis** Volcano plots illustrating differential gene expression based on RNA-Seq (top row: A–C) and differential uORF translation based on Ribo-Seq data (bottom row: D–F) across three experimental conditions: MNK inhibition with eFT508 (A, D), mTOR inhibition with rapamycin (B, E), and combined mTOR and MNK inhibition (C, F). The X-axis represents the $\log_2$ (fold change) of RNA-Seq levels, while the Y-axis displays the $\log_{10}$ (adjusted P-value). Each dot represents a transcript; black indicates no significant change, blue signifies down-regulation, and orange indicates up-regulation. The significance threshold is set at 1 for $\log_{10}$ (adjusted P-value) and ± 1 for $\log_2$ (fold change).

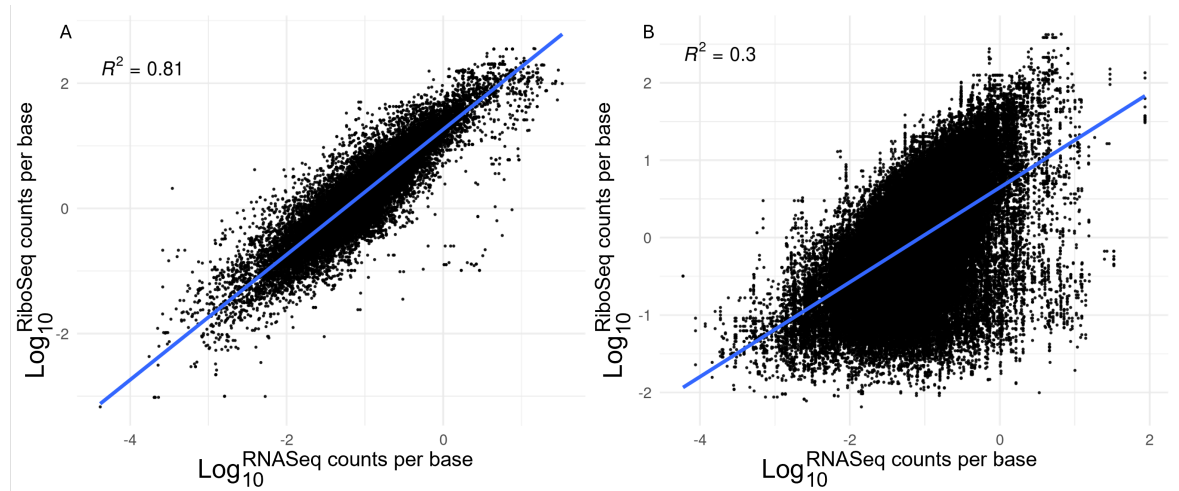


Figure 37: **Correlation between RNA-Seq and Ribo-Seq reads in CDS and uORFs.** Scatter plots illustrating the correlation between RNA-Seq (x-axis) and Ribo-Seq (y-axis) counts per base for CDS and uORF. Both axes are presented as $\log_{10}()$ transformed read counts per base. A, which represents the CDS regions, shows a strong correlation between RNA-Seq and Ribo-Seq reads, indicating consistent translation efficiency. Conversely, B shows the uORFs, where the correlation is less apparent, suggesting variability in translation efficiency in these regions.

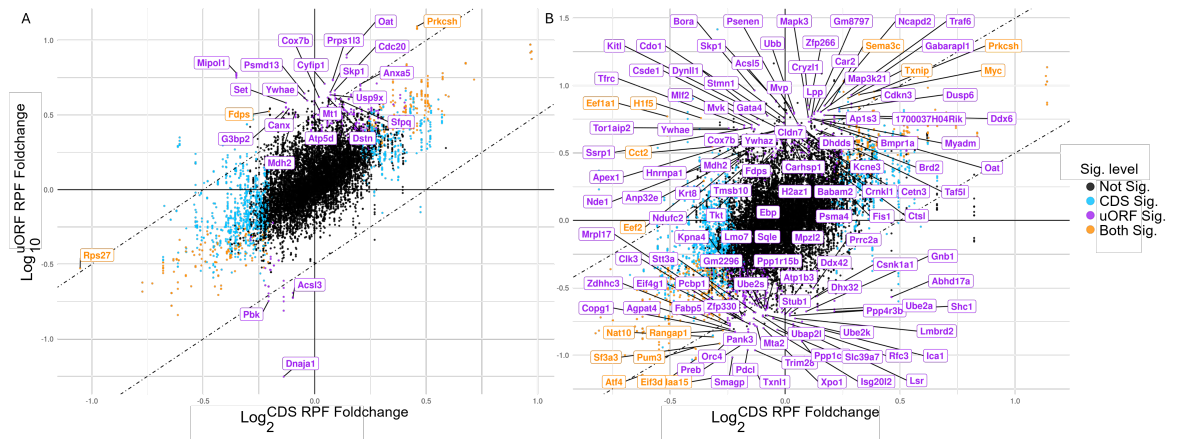


Figure 38: **Correlations between uORF and CDS ribosome occupancy changes.** The X-axis represents the $\log_2$ (fold change) in CDS ribosome occupancy for corresponding experimental setups, while the Y-axis shows the $\log_2$ (fold change) in uORF ribosome counts. The colour of each dot indicates the significance of differential expression, with an adjusted P-value less than 0.1: black dots represent non-significant changes in both CDS and uORF; orange indicates significant changes in ribosome binding rates for both CDS and uORF; blue represents significant changes only in CDS; and purple denotes significant changes only in uORF. Two diagonal lines, $y = x \pm 0.5$, set a rough boundary where uORF and CDS ribosome occupancy changes differ significantly. Gene names are labelled to highlight instances where both uORF and CDS ribosome occupancy rates are significantly altered, and where the changes in uORF occupancy significantly differ from changes in ribosome occupancy in CDS. A) shows differential expression (DE) between the rapamycin treatment group and the control group; B) illustrates DE between the combined eFT508 and rapamycin treatment group and the control group. Overall, both panels demonstrate a general positive correlation between uORF and CDS ribosome occupancy changes, reflecting global translational shifts in response to the inhibitors. However, a subset of uORFs, such as ATF4 and MYC, exhibits an anti-correlation to their corresponding CDS translation levels (located in the top-left and bottom-right). This non positive correlation is indicative of uORFs potentially acting as regulators of downstream main ORF translation under these specific treatment conditions.

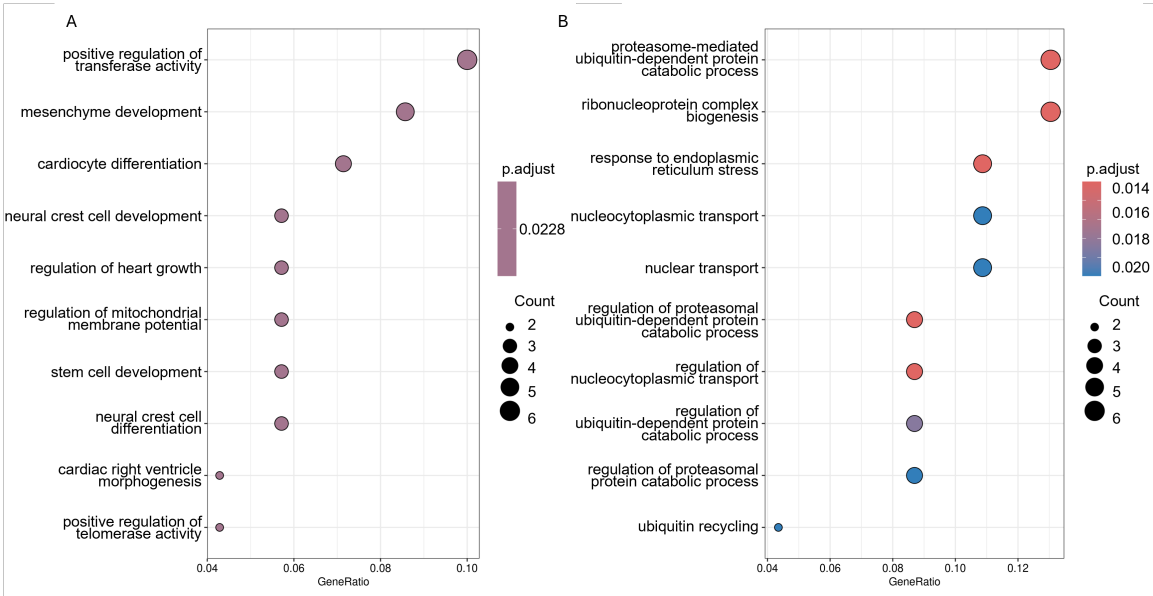


Figure 39: **Gene set enrichment analysis of genes where its uORFs have potential regulatory effects on CDS** Dot plots representing enriched Gene Ontology biological processes for genes with uORFs that are predicted to have regulatory effects on downstream CDS expression, identified through differential translation efficiency analysis comparing the combined eFT508 and rapamycin treatment group to the control group. A) shows enriched biological processes for uORFs that potentially up-regulate their CDS.; B) displays enriched processes for uORFs that downregulate their CDS.

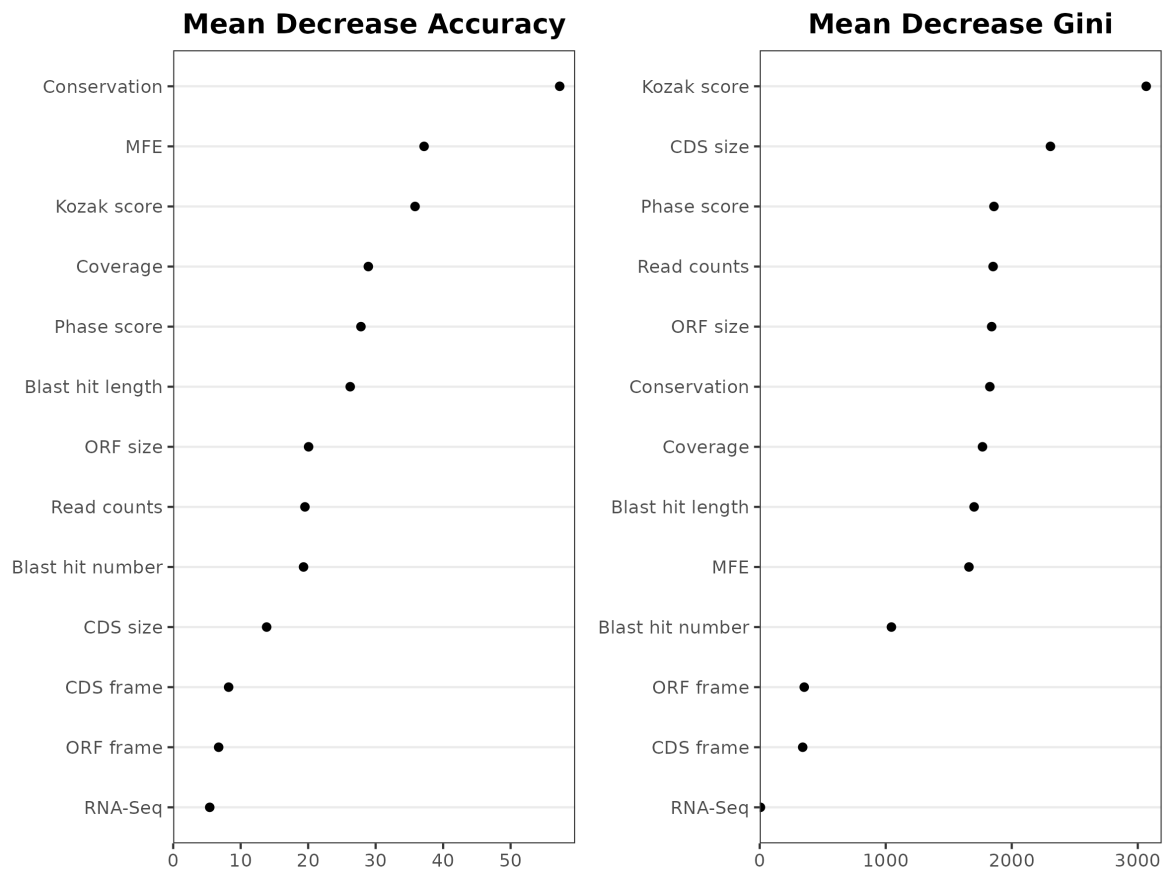


Figure 40: **Feature importance rankings in random forest models for predicting cell viability.** Feature importance analysis identifying predictive factors in random forest models trained for predicting of whether the ORFs significantly impact cell viability and proliferation. Feature rankings were determined based on mean decrease in accuracy (MDA) and mean decrease in Gini impurity (MDG) as used in the R `randomForest` package. MDA evaluates the mean decrease in classifier accuracy when the values of a feature are randomly permuted where a higher value indicate that feature is more critical for maintaining model predictive performance. MDG is a dimensionless measure representing the total decrease in node impurity (Gini index) of how each variable splits data into classes. Higher values suggest that the feature plays a substantial role in defining class boundaries.

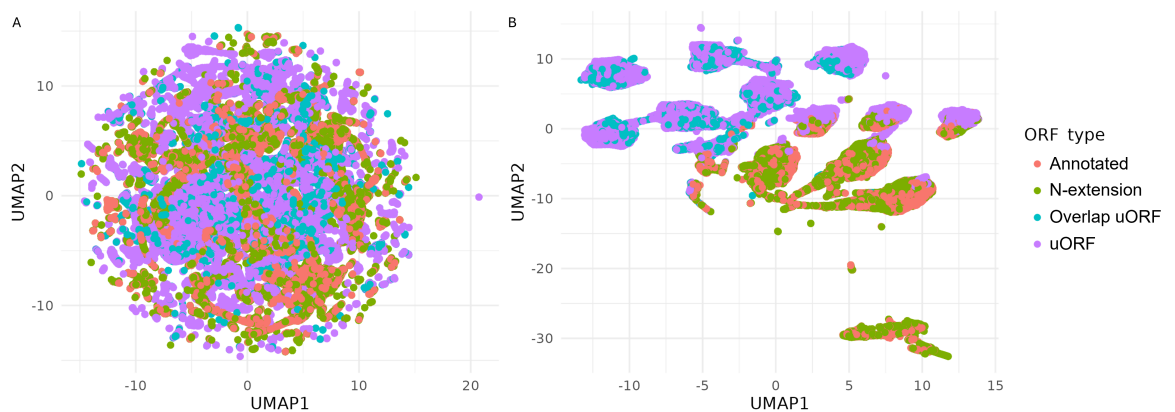


Figure 41: **Impact of information gain on cluster clarity.** UMAP visualisations demonstrating the improvement in clustering of ORFs by incorporating additional features in the ORF identification workflow. UMAP embeddings of ORFs using two feature sets. A) (restricted set): ORF size, ORF reading frame, corresponding CDS reading frame, and Ribo-Seq read counts. B) (full set): ORF size, ORF reading frame, corresponding CDS reading frame, Ribo-Seq read counts, phase score, conservation, minimum free energy (MFE), BLAST hit number, BLAST mean alignment length, and Kozak score. Colours indicate ORF type (uORF, annotated CDS, overlapping uORF) and were not used as embedding features.

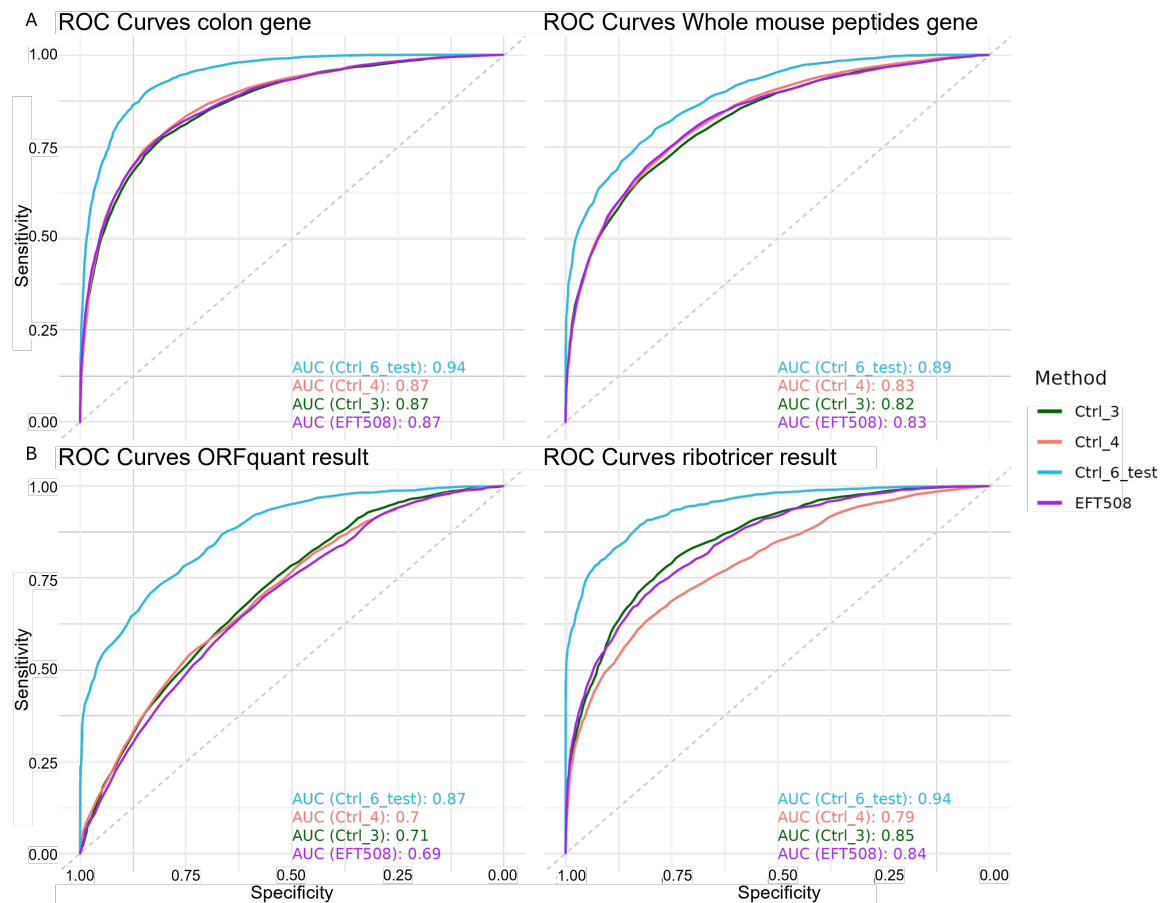


Figure 42: ROC analysis of Random Forest models discriminating translated CDSs. ROC curve analysis assessing the performance of random forest models trained to discriminate translated CDSs based on ground truth datasets (A) and other algorithms CDS predictions (B). Models were trained using the ‘ctrl 6’ dataset, balancing panel categories (‘within gene set’ vs ‘not within gene set’) and dividing the data into training (70%) and testing (30%) subsets. Model then was evaluated on ‘ctrl6 test’ (the remaining, unseen 30% of the ctrl6 dataset) to assess baseline performance, as well as on three independent datasets (ctrl4, ctrl3, and EFT508 where EFT508 is the treatment group representing the generalisation of the model). Models trained on the colon-specific gene dataset (upper left), the whole mouse gene dataset (upper right) ORFquant prediction (lower left) and ribotricer prediction (lower right) demonstrated consistent predictive performance across various replicates, as indicated by consistently high area under the ROC curve (AUC) values. This analysis highlights the generalisability and precision of random forest models in accurately distinguishing translating CDSs from non-translating candidates across different biological and experimental conditions.

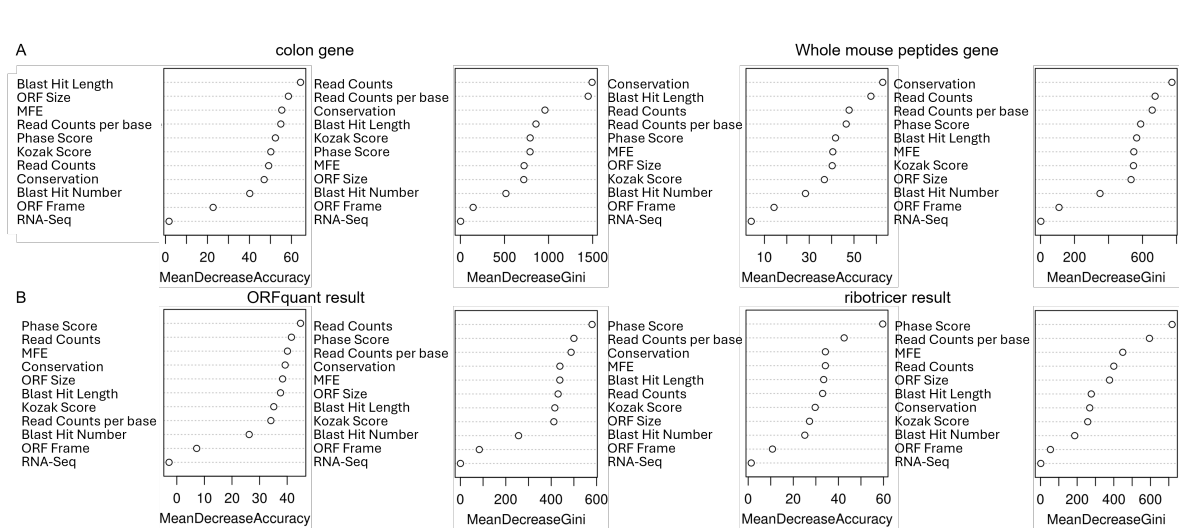


Figure 43: Feature importance rankings in random forest models for predicting translating CDSs. Feature importance analysis identifying predictive factors in random forest models trained for CDS discrimination. Feature rankings were determined based on mean decrease in accuracy (MDA) and mean decrease in Gini impurity (MDG) as used in the R `randomForest` package. MDA evaluates the mean decrease in classifier accuracy when the values of a feature are randomly permuted where a higher value indicate that feature is more critical for maintaining model predictive performance. MDG is a dimensionless measure representing the total decrease in node impurity (Gini index) of how each variable splits data into classes. Higher values suggest that the feature plays a substantial role in defining class boundaries. A) Feature importance for models trained using colon-specific gene and whole mouse gene ground truth datasets. For the colon-specific dataset, read counts emerged as the most influential feature, whereas for the whole mouse gene set, ORF conservation was the primary predictive factor. B) Feature importance for models trained to predict outcomes from established algorithms, ORFquant and ribotricer. Both models emphasised phase score as the dominant predictive feature.

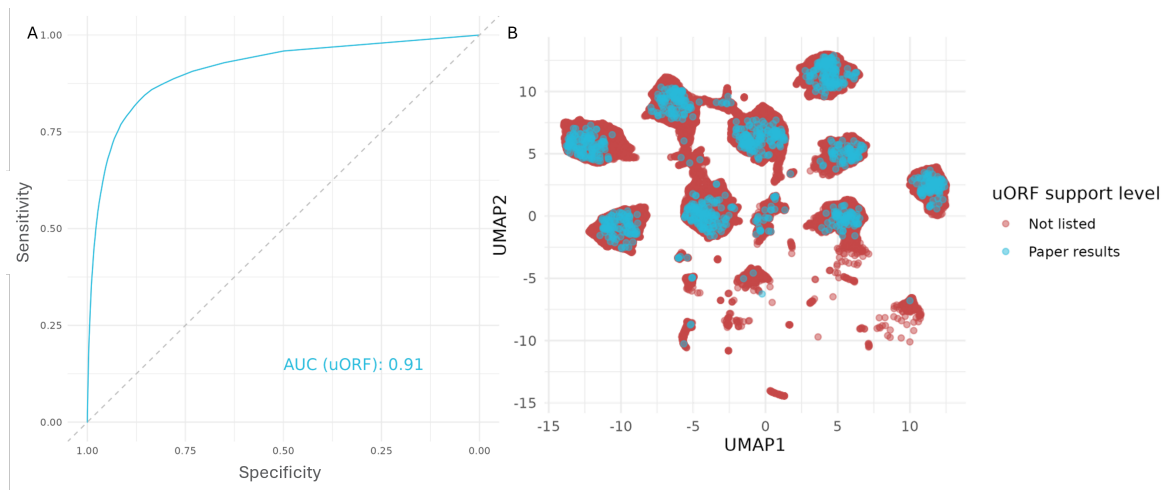


Figure 44: Random forest model performance and UMAP visualisation of ORF classification. A random forest model was trained on an ORF feature matrix that includes ORF size, ORF reading frame, corresponding CDS reading frame, Ribo-Seq read counts, phase score, conservation, MFE, blast hit number, blast average length, and Kozak score. The labels were based on whether the ORFs significantly impact cell viability and proliferation as validated by Chen et al. (2020). A) shows the model's ROC curve with an area under the ROC curve (AUC) values of 0.91, indicating high predictive accuracy. B) visualises the UMAP clustering of ORFs. ORFs validated by the Chen et al. study are labelled blue (Paper results) and those not shown are labelled red (Not listed).

Chapter 6

Discussion

The untranslated regions (UTRs) of transcripts are critical post-transcriptional regulatory hubs. Traditionally considered non-coding, these regions are now known to harbour a wide array of regulatory elements. Advances in sequencing technology, particularly ribosome profiling, have revealed that translation events frequently occur within UTRs, challenging the conventional view of a strictly monocistronic eukaryotic mRNA and sparking intense interest in the UTR's protein coding capacity. Upstream open reading frames, in particular, are predicted to be prevalent in mammalian transcripts, yet a comprehensive, experimentally validated map of the UTR translome has remained elusive.

This thesis confronted this challenge by developing a novel, data-driven workflow to improve the identification and characterisation of translated ORFs, with a specific focus on uORFs. The project evolved from an initial exploration of regulatory patterns using existing annotations to the construction of a de novo prediction pipeline. This was motivated by the discovery that state-of-the-art tools, such as ORFquant and ribotricer, demonstrated inconsistent performance, especially with lower-depth sequencing data and for the detection of short, non-canonical uORFs. My research culminated in a novel modular workflow, BYU, that integrates standardised data processing, multi-layered filtering, and comprehensive feature engineering to generate a list of ORFs candidates with translation potentials.

The application of this workflow to mouse and human datasets has yielded some biological insights into those identified uORF candidates. Initially, this compositional profiling analyses indicated that uORF sequences are not random, but could carry cell cycle linked regulatory signals as evidenced by alignment-free features separate uORFs associated with different cell cycle phases. Those identified uORFs could be peptides encoding sequence or part of larger motifs of RBP binding sites which has paired in frame start and stop codon. Then this sequence architecture findings inspired the codon usage analysis. uORF candidates displayed cell cycle phase dependent AT3/GC3 bias that differ from their corresponding downstream CDSs. This finding is consistent with literature that this codon usage preference may contribute to mRNA structures stability, and potentially

regulating translational initiation and elongation processes (Hia et al. 2019; Gillen, Waldron, and Bushell 2021). Beyond physicochemical properties, evolutionary analysis demonstrated that conservation within these regions is frequently high, in many cases exceeding that of the surrounding 5' leader sequence. This evidence of purifying selection then extend to the investigation of functional analysis where recurrent, homologous uORFs were identified across genes sharing common biological pathways, suggesting a potential mechanism of translational control. Then these static features were tested in a dynamic scenario. Disrupting the MNK/mTOR signalling caused a shift in ribosome occupancy between uORFs and their downstream CDSs, indicating that these identified uORF may participate in a responsive gating of downstream translation under stress. And in the end, an ML model was trained on these features to prioritise ORFs with experimental support, reinforcing the view that a defined subset of those identified uORFs candidates could act as regulators of CDS translation.

6.1 Overview of the research

The initial aim of this project was relatively open-ended, with the goal of “use machine learning to investigate regulatory patterns within the UTR and to ask how they help coordinate genome-wide expression programmes”. Over four years, the project evolved to centre on the identification and characterisation of uORFs, divided into several specific tasks.

During the first year, existing uORF annotations from human and mouse genomes, sourced from the uORFDB online database (Wethmar et al. 2014), and published literature (Hong Zhang et al. 2021) served as the primary research subjects. Two hypotheses were developed: first, that gene sets with distinct expression periodicity would have co-regulated uORFs with similar sequence patterns, given uORFs' regulatory roles over downstream CDS; second, that mutations within uORFs in tumour cells, similar to those in main coding regions, could influence codon preference. To explore the first hypothesis, cell cycle-related gene sets, specifically G1-S and G2-M-like groups, were selected due to their highly periodic expression patterns (Dominguez et al. 2016; Santos, Wernersson, and Jensen 2015). An SVM model was trained to distinguish uORF sequence patterns between these groups, with sequence patterns extracted using chaos game representation. The SVM successfully differentiated between uORFs in each group, indicating embedded patterns within the sequences. For the second hypothesis, codon analysis using TCGA data revealed common patterns of GC3 to AT3 transitions within the identified uORFs in tumours, consistent with observations in the main coding regions. This initial research established a framework for analysing uORF sequences and codon patterns from annotated coordinates.

In the second year, the project aimed to understand how uORFs might regulate downstream CDS translation and further investigate the sequence patterns detected by the SVM model, searching for more features to fine-characterise uORFs. Experimental data were sought to examine the effects of treatment on uORF versus CDS translation and to identify reliable uORF annotations. Unpublished data from Dr Joseph Waldron's project, generated in the Bushell lab, provided a key data source. This study explored the involvement of mTOR and MNK signalling pathways in regulating translation initiation in an APC^{-/-} KRAS^{G12D} mouse colon organoid

2300 model where eIF2 α is phosphorylated. The analysis assessed whether MNK and mTOR inhibitors differen-
2301 tially affected uORF and CDS translation. Workflow updates included expression analysis using Ribo-Seq and
2302 RNA-Seq data, revealing subsets of uORFs regulated differently from corresponding CDSs, suggesting reg-
2303 ulatory roles over CDS translation. Additionally, workflows for inter-species conservation and intra-species
2304 sequence homology analysis were developed, accompanied by an SQLite database to store uORF coordinates
2305 and annotations, and an R Shiny data portal for data access and expression analysis.

2306 The third year focused on re-evaluating ORF annotations have been applied in this project, discovering that
2307 many were predominantly AUG-initiated despite evidence suggesting that non-AUG initiation sites are com-
2308 mon in uORFs (J. Chen et al. 2020; Andreev, Gary Loughran, et al. 2022). Various ORF detection algorithms
2309 were reviewed, including ORFquant and Ribotricer, but their predictions were inconsistent across replicates
2310 and did not meet expected discovery rates. Consequently, a new ORF discovery method was developed to
2311 accurately capture uORFs. This workflow standardised raw data processing and included additional ORF se-
2312 quence annotations, such as molecular thermal data and Kozak sequence scoring. The resulting workflow,
2313 temporarily named BYU, demonstrated improved inter-replicate agreement and produced predictions aligned
2314 with expected ORF numbers.

2315 In the final year, the workflow was expanded into a comprehensive ORF identification and analysis toolbox,
2316 integrating developments from previous years. The latest version generated a feature matrix for each ORF
2317 candidates, incorporating inter- and intraspecies similarity, minimum free energy, and Kozak scores. Machine
2318 learning algorithms were then applied to group uORFs into biologically meaningful clusters. The toolbox
2319 represents the major output of this research, with all results presented in the previous chapters derived from its
2320 latest iteration. These results were structured around identification, characterisation, machine learning-based
2321 clustering, and downstream analysis.

2322 The primary objective of this thesis is to demonstrate the development of a novel and scalable workflow for
2323 ORF candidates detection and characterisation using Ribo-Seq and RNA-Seq data, highlighting its capability
2324 to reveal biological insights within experimental datasets.

2325 **6.2 Improved workflow for uORF detection: from translation to func-** 2326 **tion**

2327 The major outcome of this research was the development of a more consistent and scalable workflow for the de-
2328 tection and characterisation of ORFs candidates, as detailed in Chapter 3. My workflow reframes the challenge
2329 of uORF identification into two sequential questions: 1) can we reliably detect a translating ribosome produc-
2330 ing a peptide, distinguishing genuine signal from noise? And 2) are these translating ribosomes biologically
2331 meaningful? My workflow tries to address these two questions in two stages.

2332 The first stage, detection, is designed to address the performance limitations of existing tools. As shown in

Chapter 4, tools that rely heavily on triplet periodicity struggle to consistently identify uORFs, particularly in datasets with moderate or low sequencing depth. This is because accurate periodicity requires precise P-site identification and high read coverage, conditions that are often not met for short, sparsely translated uORFs (Ahmed et al. 2019; Erhard et al. 2018; Z. Xiao et al. 2018). The BYU workflow’s detection stage therefore prioritises recall, using a multi-layered filtering approach based on fundamental evidence of transcription and ribosome occupancy to first capture all potential translation events.

Once a comprehensive library of translated candidates is established, the workflow’s second stage addresses the question of function. This directly engages with a philosophical point in the field: does a uORF have to be functional to be called a uORF? My approach allows this question to be addressed empirically by gathering secondary evidence of biological relevance (e.g., conservation, Kozak context, homology). This framework allows for a more nuanced definition of terms: we can define a ‘translated uORF’ as one that passes the sensitive detection filters of stage one, and a ‘potentially functional uORF’ as a translated uORF that is further supported by features from stage two that suggest it is under selective pressure. It must be acknowledged that sequence conservation within the 5’ leader could theoretically arise from non-coding regulatory elements, such as RBP motifs, rather than peptide coding potential. But the initial stage of this workflow explicitly evaluates the necessity of ORF architecture. Although satisfying these structural parameters is not sufficient to prove functionality in isolation, it imposes a restrictive filter. Therefore, by enforcing the necessary condition of ORF structure, the probability that a conserved candidate is a misclassified, non-ORF regulatory motif is significantly minimised. A more detailed exploration of the rationale behind this two-stage design is provided in the following section.

6.2.1 Rationale and philosophical underpinnings of the BYU workflow

The design of the BYU workflow’s detection stage is guided by the principle of Occam’s razor, as justified in Chapter 3. It begins with the simplest hypothesis that requires the fewest assumptions: a region with sufficient evidence of both transcription (from RNA-Seq) and ribosome occupancy (from Ribo-Seq) is a candidate for translation. This workflow, unlike existing ORF detection applications, has the raw sequencing data processing pipeline integrated to ensure not only standardised but also high-quality sequence mapping, thereby reducing the false positive rate of potential artificial ORF identification. Furthermore, the ORF type naming in existing databases and identification tools is inconsistent, which complicates comparisons (Santos, Wernersson, and Jensen 2015; Martinez et al. 2020; Choudhary, Li, and D. Smith 2020; Calviello, Hirsekorn, and Ohler 2020). To address this, my workflow adopts the unified naming scheme proposed by Mudge et al. to standardise ORF typing (Mudge et al. 2022).

This workflow aimed to address the performance limitations of existing ORF-finding tools on uORFs. As demonstrated in Chapter 4, both Ribotricer and ORFquant struggled to produce consistent ORF predictions. Specifically, for uORFs, as shown in Figure 20, only 2% and 8% of uORFs were consistently identified across three replicates using ORFquant and Ribotricer, respectively. This suggests that the main discriminator, the

periodicity identifier, may not be suitable for uORF identification. To accurately calculate periodicity, two conditions must be met: a) accurate P-site identification and b) high local coverage of reads.

Due to the nature of Ribo-Seq, which sequences RPF fragments of 26 to 34 nt (Q. Wang and Mao 2023), the P-site is not located at the beginning of the sequenced reads and an offset is required for correct alignment. It is believed that different Ribo-Seq read lengths correspond to different offset values due to variations in ribosome conformations (Ahmed et al. 2019; Nicholas T Ingolia et al. 2012). Fundamentally, the P-site offset depends on the physical size and footprint of the ribosome protecting the mRNA. However, because these global offset values are typically determined empirically using the abundant footprints aligned to main coding regions, they may not be perfectly applicable to uORFs, as stated by Ahmed et al. (2019). Consequently, the offset applied to uORFs may not recover the P-site as accurately as it does in most CDS regions. As for local read coverage, uORF regions are less covered than CDS regions due to a lower overall translation rate. Given the systematic sequencing error rate, the signal to noise ratio is higher in uORF regions, leading to less accurate triplet periodicity (Erhard et al. 2018; Z. Xiao et al. 2018).

Since the two prerequisites for accurate periodicity determination in the uORF region cannot be easily satisfied, a new strategy is needed. This led me to think of what is the fundamental evidence provided by ribosome profiling. While the presence of a ribosome on an mRNA is a strong indicator, is it sufficient to claim translation? The goal, therefore, is to find additional evidence that the Ribo-Seq read is truly ribosome-protected or to identify signs that the Ribo-Seq-covered region is highly likely to be protein coding.

This then prompted a re-evaluation of the concept of a gene. While a gene is a unit of heredity that produces functional molecules (Orgogozo, Peluffo, and Morizot 2016), advances in sequencing have blurred the lines of the classic monocistronic model of eukaryotic mRNA (Bruce Alberts ... [and others] 1989). With the discovery that over half of human genes contain additional coding regions like uORFs (McGillivray et al. 2018; Wethmar et al. 2014; Patraquim et al. 2020), it becomes necessary to borrow from the methods used to determine the biotype of transcripts before the advent of Ribo-Seq.

According to the Ensembl gene annotation system (2016), transcripts are classified as protein-coding based on a combination of criteria, including ORF length, experimental evidence from proteomics, cross-species conservation, and known protein domains. This multi-evidence approach, often validated by manual curation (Frankish et al. 2019), inspired the second stage of my workflow.

While periodicity is an effective discriminator, its high false-negative rate for uORFs is a major drawback. Developers must choose between high precision at the cost of recall, or higher recall at the cost of precision. Most tools adopt the former strategy. However, I believe maximising the initial recognition of candidates is crucial; a potential ORF can be scrutinised with further evidence, but an ORF prematurely rejected is lost from all downstream analysis. Therefore, the BYU workflow uses the STU approach to first generate a broad library of candidates with high sensitivity, as demonstrated in Chapter 4 (Table 3). Subsequently, this library is annotated with a rich feature matrix (Kozak score, conservation, etc.) to allow for a more sophisticated, data-driven curation, leading to more confident and consistent predictions that align with expected expression

2404 rates (Tung et al. 2020; Frankish et al. 2019).

2405 **6.3 Key findings in downstream analysis**

2406 The downstream analysis of the identified potential translated uORFs yielded several significant findings that
2407 contribute to our understanding of uORFs and their roles in gene regulation. Taken together, the integrated
2408 downstream analyses show that those identified uORFs modulate downstream initiation and CDS translation
2409 (Chapter 5.5), exhibit characteristic codon-usage and GC3 biases (Chapter 5.2), display conservation and re-
2410 current 5' leader motifs consistent with regulon like control (Chapter 5.3, 5.4), and encode sequence pattern
2411 signals that can distinguishing cell state status (Chapter 5.1).

2412 **6.3.1 Sequence patterns emerging in uORF of cell cycle related genes**

2413 In the sequence pattern analysis, I applied the chaos game representation (CGR), a symbolic sequence repre-
2414 sentation method based on an iterated function system (Thind and Sinha 2023). CGR was originally developed
2415 in the field of fractal theory to explore and visualise self-similarity and fractal geometry. It can transform
2416 complex, long sequences such as DNA, RNA, and protein sequences into visual and mathematical patterns,
2417 revealing underlying structures, patterns, and motifs that are otherwise difficult to detect (Löchel et al. 2020).
2418 CGR provides a 2D fractal representation of a 1D sequence, and the key information embedded in the 2D
2419 fractal map is the distribution of repetitive units of a defined size.

2420 The sequence pattern analysis of cell cycle-related genes revealed that those uORFs candidates within G1-S
2421 and G2-M genes have distinct sequence patterns which is smaller repetitive sequences within each uORF. The
2422 CGR/SVM analysis shows that the small profile sequence composition within uORFs could carry sufficient
2423 signal to distinguish G1-S from G2-M sets with balanced performance well above a shuffled baseline (table
2424 4). Because these features are not related to read depth and ORF length, the separation suggests a genuine
2425 sequence pattern linked to cell-cycle state rather than the coverage artefacts.

2426 This result may imply that the functionality of those potential uORF in cell cycle related gene are encoded di-
2427 rectly within their sequence composition. Those uORFs could act as cis-regulatory elements on the ribosomes.
2428 This aligned well with the studies suggesting that specific internal sequence motifs could modulate ribosome
2429 behaviour during different translation phases (Plotkin and Kudla 2011; Bentele et al. 2013; Gunišová et al.
2430 2018). These sequences may function by encoding certain amino acid motifs, such as proline rich stretches that
2431 slow the formation of the peptide bond or interfere with termination, or clusters of positively charged residues
2432 that are known to interact negatively with the ribosomal exit tunnel, thus inducing stalling or pausing (Chiba
2433 et al. 2011; Gunišová et al. 2018).

6.3.2 Codon basis in uORFs of different cell cycle genes and between uORF and CDS

As mentioned in the Chapter 5, the sequence pattern analysis results prompted me to explore the nature of the repetitive units being detected. I hypothesised that there might be codon bias when comparing G1-S and G2-M uORFs, and the results confirmed a preference for AT-ending codons in G2-M uORFs. As discussed in the Introduction, codon bias can reflect mRNA stability, translation efficiency, and even drive mutations in tumour cells (Supek et al. 2014; Suntsova et al. 2019; Buhr et al. 2016). This analysis revealed that, in this discovery cohort of uORFs, G2-M uORFs exhibit a preference for AT-ending codons compared to the G1-S cohort. In the context of mammalian translation, studies have found that GC-ending codons are generally more optimal due to the high abundant tRNAs, whereas AT-ending codons are non-optimal which are decoded more slowly due to lower tRNA availability or weaker codon and anti-codon pairing (Peterson et al. 2020; Bentele et al. 2013; Frenkel-Morgenstern et al. 2012). This preference for the slow AT-ending codon in G2-M uORFs provides a potential mechanistic explanation for the repressive function proposed in the previous section. By encoding with non-optimal codons, these uORFs could induce ribosomal slowing or stalling during elongation.

Beyond phase-specific regulation, the global comparison between uORFs and their surrounding regions provided potential evidence of purifying selection. This revealed that while the non-coding portions of the 5' UTR are GC-rich, the uORF candidates themselves display a more balanced GC content. If these identified uORFs were merely random sequences that occur by chance within 5' UTR, they would not distinguish themselves from the local nucleotide background. This observation supports the view that uORFs may be subject to evolutionary constraints similar to those acting on canonical coding sequences. As reviewed by Plotkin and Kudla (2011) non-random codon usage in CDS reflects selective pressures on translational efficiency, the identification of certain codon bias in those identified potentially translated uORFs may suggest that they are preserved to support specific translational dynamics required for their regulatory function.

6.3.3 Homologous uORF identified

Motivated by the hypothesis that uORFs can act as shared regulatory elements across co-regulated genes, I searched all detected mouse uORFs candidates for sequence homology across the genome (Chapter 5.4). Briefly, each uORF sequence was queried against the mouse genome using nucleotide BLAST, and putative homologues were retained and clustered then downstream analyses assessed whether homologous uORFs were enriched within functionally related gene sets.

This analysis revealed families of genes that share nearly identical uORF sequences. Notably, several cell-cycle related genes, including *Racgap1*, *Pthr2* and *Nek4*, harbour the same uORF sequence. In combination with the codon-level and sequence-pattern analyses presented earlier, these findings support the view that subsets of genes involved in common biological processes can carry a conserved uORF element. And such a high conservation could mark them as potential functional uORFs.

Taken together, the presence of potential homologous uORFs across functionally related genes is consistent

with an RNA translational regulon architecture proposed by Keene (Keene 2007). In this framework, the cell may not regulate these genes individually, instead, by equipping them with identical uORF modules, the translational machinery can coordinately control the entire cohort.

6.3.4 uORFs as potential regulators of CDS translation

Differential expression analysis of Ribo-Seq data revealed the discordance in ribosome distribution following treatments. If translational control was uniform across the transcript, the fold-changes of uORFs and their downstream CDSs would be correlated approximate according to the line $y = x$. As shown in Figure 38, treatment targeting the mTOR and MNK signalling pathways resulted in regulatory uncoupling where ribosome occupancy was preferentially maintained or elevated on uORF candidates, while their corresponding downstream CDSs exhibited repressed translation. This observation aligns with other studies suggesting that uORFs negatively regulate the translation of downstream coding regions (Fan et al. 2018; Araujo et al. 2012; Kumar, Srinivas, and Patankar 2015).

6.4 Conclusion

This thesis establishes a new conceptual framework and an accompanying computational workflow that fundamentally advance our ability to map and interpret the translome. By separating the challenge of ORF identification into two distinct stages, the sensitive as detection of translation and the data-driven inference of function, this work provides a more noise resistant solution to the limitations of existing methods, particularly for the accurate discovery of non-canonical uORFs in experimentally realistic datasets. The resulting BYU workflow contributes to the field, offering an alternative but scalable, consistent approach to identify potential ORFs for exploring the hidden proteome.

The application of this workflow provided direct evidence for the complex and sophisticated regulatory roles of uORFs. The case studies on mouse and human data moved beyond simple identification to uncover tangible biological mechanisms. The discovery of distinct codon usage patterns in cell cycle-related uORFs, for example, situates these elements within the wider context of ‘codon optimality’, suggesting they act as key instruments in the temporal regulation of protein synthesis. Furthermore, the identification of potential homologous uORF sequences shared across functionally related genes provides strong evidence for the existence of ‘translational regulons’, a novel mechanism for coordinating the expression of entire gene networks post-transcriptionally. These findings collectively demonstrate that many uORFs are not translational noise but are instead critical regulatory elements that mediate the gene specific outcomes of master signalling pathways.

While this work provides a significant step forward, it also may guide the path for future research. A recent publication of an expanded GENCODE reference catalogue of translated non-canonical open reading frames (such as uORFs) also highlighted the difficulty of the identification (Chothani et al. 2026). They established a

2500 high-confidence 'Primary set' of 10,127 ncORFs, of which 66.2% are uORFs (uORF and overlap-uORF 17).
2501 Meanwhile, they have removed previous restrictive filters, allowing the inclusion of ncORFs of 16 codons or
2502 fewer and those initiated by non-AUG start codons thereby generating a more comprehensive set. They argued
2503 that the high frequency of short and non-AUG ncORFs could be driven by genuine biological selections, espe-
2504 cially in translational regulation in 5' UTR, rather than technical artefacts. However, the catalogue still relies
2505 on mainly translation signature scores such as reading frame uniformity which will penalise longer uORFs
2506 and those overlapping with other reading frames. This limitation may suggest the potential utility of the BYU
2507 workflow presented here, which offers a framework for functional inference. Looking ahead, the workflow's
2508 current relies mostly on existing gene annotations highlights the ultimate need for true de novo assembly of
2509 the translome, and the machine learning models (even AI models) for inferring function will evolve more
2510 powerful. The main achievement of this work has been the in-depth characterisation of protein-coding uORFs;
2511 the next step is to integrate this knowledge with the other regulatory layers encoded in UTRs, such as RNA
2512 secondary structures and protein binding sites. The emergence of large language models capable of decoding
2513 the regulatory grammar of non-coding regions (Benegas, Batra, and Song 2023; Hwang et al. 2024; Chu et al.
2514 2024) presents a particularly exciting opportunity. By combining the high-quality catalogue of potential trans-
2515 lated uORFs candidates produced by this workflow with the predictive power of these advanced models, the
2516 field is poised to achieve a truly comprehensive understanding of the regulatory landscape encoded within our
2517 genes, moving far beyond the canonical coding sequences.

References

- Adomavicius, Tomas, Margherita Guaita, Yu Zhou, Martin D. Jennings, Zakia Latif, Alan M. Roseman, and Graham D. Pavitt (2019). “The structural basis of translational control by eIF2 phosphorylation”. In: *Nature Communications* 10.1, p. 2136. DOI: 10.1038/s41467-019-10167-3.
- Ahmed, Nabeel, Pietro Sormanni, Prajwal Ciryam, Michele Vendruscolo, Christopher M. Dobson, and Edward P. O’Brien (2019). “Identifying A- and P-site locations on ribosome-protected mRNA fragments using Integer Programming”. In: *Scientific Reports* 9.1, p. 6256. DOI: 10.1038/s41598-019-42348-x.
- Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley, Julio Fernandez Banet, Konstantinos Billis, Carlos García Girón, Thibaut Hourlier, Kevin Howe, Andreas Kähäri, Felix Kokocinski, Fergal J. Martin, Daniel N. Murphy, Rishi Nag, Magali Ruffier, Michael Schuster, Y. Amy Tang, Jan-Hinnerk Vogel, Simon White, Amonida Zadissa, Paul Flicek, and Stephen M. J. Searle (2016). “The Ensembl gene annotation system”. In: *Database* 2016, baw093. DOI: 10.1093/database/baw093.
- 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 (2015). “A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance”. In: *Cell* 160.4, pp. 595–606. 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 (2022). “Non-AUG translation initiation in mammals”. In: *Genome Biology* 23.1, p. 111. DOI: 10.1186/s13059-022-02674-2.
- Andreev, Dmitry E., Patrick B. F. O’Connor, Gary Loughran, Sergey E. Dmitriev, Pavel V. Baranov, and Ivan N. Shatsky (2017). “Insights into the mechanisms of eukaryotic translation gained with ribosome profiling”. In: *Nucleic Acids Research* 45.2, pp. 513–526. DOI: 10.1093/nar/gkw1190.
- Andrews, Shea J. and Joseph A. Rothnagel (2014). “Emerging evidence for functional peptides encoded by short open reading frames”. In: *Nature Reviews Genetics* 15.3, pp. 193–204. DOI: 10.1038/nrg3520.
- Araujo, Patricia R., Kihoon Yoon, Daijin Ko, Andrew D. Smith, Mei Qiao, Uthra Suresh, Suzanne C. Burns, and Luiz O. F. Penalva (2012). “Before It Gets Started: Regulating Translation at the 5’ UTR”. In: *Comparative and Functional Genomics* 2012, pp. 1–8. DOI: 10.1155/2012/475731.

- Aspden, Julie L, Ying Chen Eyre-Walker, Rose J Phillips, Unum Amin, Muhammad Ali S Mumtaz, Michele Brocard, and Juan-Pablo Couso (2014). “Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq”. In: *eLife* 3, e03528. DOI: 10.7554/eLife.03528.
- Audic, Yann and Rebecca S. Hartley (2004). “Post-transcriptional regulation in cancer”. In: *Biology of the Cell* 96.7, pp. 479–498. DOI: 10.1016/j.biolcel.2004.05.002.
- Bae, Bongmin and Pedro Miura (2020). “Emerging Roles for 3’ UTRs in Neurons”. In: *International Journal of Molecular Sciences* 21.10, p. 3413. DOI: 10.3390/ijms21103413.
- Benegas, Gonzalo, Sanjit Singh Batra, and Yun S. Song (2023). “DNA language models are powerful predictors of genome-wide variant effects”. In: *Proceedings of the National Academy of Sciences* 120.44, e2311219120. DOI: 10.1073/pnas.2311219120.
- Bentele, Kajetan, Paul Saffert, Robert Rauscher, Zoya Ignatova, and Nils Blüthgen (2013). “Efficient translation initiation dictates codon usage at gene start”. In: *Molecular Systems Biology* 9.1, p. 675. DOI: 10.1038/msb.2013.32.
- Bohlen, Jonathan, Kai Fenzl, Günter Kramer, Bernd Bukau, and Aurelio A. Teleman (2020). “Selective 40S Footprinting Reveals Cap-Tethered Ribosome Scanning in Human Cells”. In: *Molecular Cell* 79.4, 561–574.e5. DOI: 10.1016/j.molcel.2020.06.005.
- Bohlen, Jonathan, Liza Harbrecht, Saioa Blanco, Katharina Clemm Von Hohenberg, Kai Fenzl, Günter Kramer, Bernd Bukau, and Aurelio A. Teleman (2020). “DENR promotes translation reinitiation via ribosome recycling to drive expression of oncogenes including ATF4”. In: *Nature Communications* 11.1, p. 4676. DOI: 10.1038/s41467-020-18452-2.
- Breiman, Leo (2001). “Random Forests”. In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.
- Breschi, Alessandra, Thomas R. Gingeras, and Roderic Guigó (2017). “Comparative transcriptomics in human and mouse”. In: *Nature Reviews Genetics* 18.7, pp. 425–440. DOI: 10.1038/nrg.2017.19.
- Bruce Alberts ... [and others] (1989). *Molecular biology of the cell*. Second edition. New York : Garland Pub., [1989].
- Buhr, Florian, Sujata Jha, Michael Thommen, Joerg Mittelstaet, Felicitas Kutz, Harald Schwalbe, Marina V. Rodnina, and Anton A. Komar (2016). “Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations”. In: *Molecular Cell* 61.3, pp. 341–351. DOI: 10.1016/j.molcel.2016.01.008.
- Bushnell, Brian, Jonathan Rood, and Esther Singer (2017). “BBMerge – Accurate paired shotgun read merging via overlap”. In: *PLOS ONE* 12.10. Ed. by Patrick Jon Biggs, e0185056. DOI: 10.1371/journal.pone.0185056.
- Calviello, Lorenzo, Antje Hirsekorn, and Uwe Ohler (2020). “Quantification of translation uncovers the functions of the alternative transcriptome”. In: *Nature Structural & Molecular Biology* 27.8, pp. 717–725. DOI: 10.1038/s41594-020-0450-4.
- Calviello, Lorenzo, Neelanjan Mukherjee, Emanuel Wyler, Henrik Zaubler, Antje Hirsekorn, Matthias Selbach, Markus Landthaler, Benedikt Obermayer, and Uwe Ohler (2016). “Detecting actively translated open read-

- ing frames in ribosome profiling data”. In: *Nature Methods* 13.2, pp. 165–170. DOI: 10.1038/nmeth.3688.
- Calviello, Lorenzo and Uwe Ohler (2017). “Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome”. In: *Trends in Genetics* 33.10, pp. 728–744. DOI: 10.1016/j.tig.2017.08.003.
- Calvo, Sarah E., David J. Pagliarini, and Vamsi K. Mootha (2009). “Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans”. In: *Proceedings of the National Academy of Sciences* 106.18, pp. 7507–7512. DOI: 10.1073/pnas.0810916106.
- Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden (2009). “BLAST+: architecture and applications”. In: *BMC Bioinformatics* 10.1, p. 421. DOI: 10.1186/1471-2105-10-421.
- Chaney, Julie L., Aaron Steele, Rory Carmichael, Anabel Rodriguez, Alicia T. Specht, Kim Ngo, Jun Li, Scott Emrich, and Patricia L. Clark (2017). “Widespread position-specific conservation of synonymous rare codons within coding sequences”. In: *PLOS Computational Biology* 13.5. Ed. by Claus O. Wilke, e1005531. DOI: 10.1371/journal.pcbi.1005531.
- Chen, Hung-Hsi and Woan-Yuh Tarn (2019). “uORF-mediated translational control: recently elucidated mechanisms and implications in cancer”. In: *RNA Biology* 16.10, pp. 1327–1338. DOI: 10.1080/15476286.2019.1632634.
- Chen, Jin, Andreas-David Brunner, J. Zachery Cogan, James K. Nuñez, Alexander P. Fields, Britt Adamson, Daniel N. Itzhak, Jason Y. Li, Matthias Mann, Manuel D. Leonetti, and Jonathan S. Weissman (2020). “Pervasive functional translation of noncanonical human open reading frames”. In: *Science* 367.6482, pp. 1140–1146. DOI: 10.1126/science.aay0262.
- Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu (2018). “fastp: an ultra-fast all-in-one FASTQ preprocessor”. In: *Bioinformatics* 34.17, pp. i884–i890. DOI: 10.1093/bioinformatics/bty560.
- Chiba, Shinobu, Takashi Kanamori, Takuya Ueda, Yoshinori Akiyama, Kit Pogliano, and Koreaki Ito (2011). “Recruitment of a species-specific translational arrest module to monitor different cellular processes”. In: *Proceedings of the National Academy of Sciences* 108.15, pp. 6073–6078. DOI: 10.1073/pnas.1018343108.
- Chothani, Sonia, Jorge Ruiz-Orera, Jack A S Tierney, Michal I Swirski, Hakon Tjeldnes, Leron W Kok, Jim Clauwaert, Eric W Deutsch, M Mar Alba, Julie L Aspden, Pavel V Baranov, Ariel Alejandro Bazzini, Elspeth A Bruford, Marie A Brunet, Tristan Cardon, Anne-Ruxandra Carvunis, Claudio Casola, Jyoti Sharma Choudhary, Kellie Dean, Pouya Faridi, Ivo Fierro-Monti, Isabelle Fournier, Adam Frankish, Mark Gerstein, Norbert Hubner, Yunzhe Jiang, Manolis Kellis, Thomas F Martinez, Gerben Menschaert, Pengyu Ni, Sandra Orchard, Xavier Roucou, Joel Rozowsky, Michel Salzet, Mauro Siragusa, Sarah Slavoff, Nicola Ternette, Juan Antonio Vizcaino, Aaron Wacholder, Wei Wu, Zhi Xie, Yucheng T Yang, Robert L Moritz, Eivind Valen, Jonathan Mudge, Sebastiaan van Heesch, John R Prensner, and Owen J L Rackham (2026).

- 2618 “An expanded reference catalog of translated open reading frames for biomedical research”. In: *Nucleic*
 2619 *Acids Research* 54.6, gkag234. DOI: 10.1093/nar/gkag234.
- 2620 Choudhary, Saket, Wenzheng Li, and Andrew D. Smith (2020). “Accurate detection of short and long active
 2621 ORFs using Ribo-seq data”. In: *Bioinformatics* 36.7. Ed. by Inanc Birol, pp. 2053–2059. DOI: 10.1093/
 2622 bioinformatics/btz878.
- 2623 Chu, Yanyi, Dan Yu, Yupeng Li, Kaixuan Huang, Yue Shen, Le Cong, Jason Zhang, and Mengdi Wang (2024).
 2624 “A 5’ UTR language model for decoding untranslated regions of mRNA and function predictions”. In:
 2625 *Nature Machine Intelligence* 6.4, pp. 449–460. DOI: 10.1038/s42256-024-00823-9.
- 2626 Chukka, Prakash Amruth Raj, Stacey D. Wetmore, and Nehal Thakor (2021). “Established and Emerging
 2627 Regulatory Roles of Eukaryotic Translation Initiation Factor 5B (eIF5B)”. In: *Frontiers in Genetics* 12,
 2628 p. 737433. DOI: 10.3389/fgene.2021.737433.
- 2629 Clote, P. (2005). “Structural RNA has lower folding energy than random RNA of the same dinucleotide fre-
 2630 quency”. In: *RNA* 11.5, pp. 578–591. DOI: 10.1261/rna.7220505.
- 2631 Corpuz, Austin D., Joe W. Ramos, and Michelle L. Matter (2020). “PTRH2: an adhesion regulated molecular
 2632 switch at the nexus of life, death, and differentiation”. In: *Cell Death Discovery* 6.1, p. 124. DOI: 10.1038/
 2633 s41420-020-00357-0.
- 2634 Cunningham, Fiona et al. (2022). “Ensembl 2022”. In: *Nucleic Acids Research* 50 (D1), pp. D988–D995. DOI:
 2635 10.1093/nar/gkab1049.
- 2636 Dasgupta, Anwesha and John R Prensner (2024). “Upstream open reading frames: new players in the landscape
 2637 of cancer gene regulation”. In: *NAR Cancer* 6.2, zcae023. DOI: 10.1093/narcan/zcae023.
- 2638 Davuluri, R. V. (2000). “CART Classification of Human 5’ UTR Sequences”. In: *Genome Research* 10.11,
 2639 pp. 1807–1816. DOI: 10.1101/gr.GR-1460R.
- 2640 Diaz de Arce, Alexander J, William L Noderer, and Clifford L Wang (2018). “Complete motif analysis of
 2641 sequence requirements for translation initiation at non-AUG start codons”. In: *Nucleic Acids Research*
 2642 46.2, pp. 985–994. DOI: 10.1093/nar/gkx1114.
- 2643 Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut,
 2644 Mark Chaisson, and Thomas R. Gingeras (2013). “STAR: ultrafast universal RNA-seq aligner”. In: *Bioin-*
 2645 *formatics* 29.1, pp. 15–21. DOI: 10.1093/bioinformatics/bts635.
- 2646 Doheny, Daniel, Sara G. Manore, Grace L. Wong, and Hui-Wen Lo (2020). “Hedgehog Signaling and Truncated
 2647 GLI1 in Cancer”. In: *Cells* 9.9, p. 2114. DOI: 10.3390/cells9092114.
- 2648 Domingos, Pedro (1999). “The Role of Occam’s Razor in Knowledge Discovery”. In: *Data Mining and Knowl-*
 2649 *edge Discovery* 3, pp. 409–425. DOI: 10.1023/A:1009868929893.
- 2650 Dominguez, Daniel, Yi-Hsuan Tsai, Nicholas Gomez, Deepak Kumar Jha, Ian Davis, and Zefeng Wang (2016).
 2651 “A high-resolution transcriptome map of cell cycle reveals novel connections between periodic genes and
 2652 cancer”. In: *Cell Research* 26.8, pp. 946–962. DOI: 10.1038/cr.2016.84.
- 2653 Donato, Luigi, Concetta Scimone, Carmela Rinaldi, Rosalia D’Angelo, and Antonina Sidoti (2021). “New
 2654 evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated com-

- parison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies”. In: *Neural Computing and Applications* 33.22, pp. 15669–15692. DOI: 10.1007/s00521-021-06188-z.
- Dutton-Regester, Ken, Jared J. Gartner, Rafi Emmanuel, Nouar Qutob, Michael A. Davies, Jeffrey E. Gershenwald, William Robinson, Steven Robinson, Steven A. Rosenberg, Richard A. Scolyer, Graham J. Mann, John F. Thompson, Nicholas K. Hayward, and Yardena Samuels (2014). “A highly recurrent RPS27 5’UTR mutation in melanoma”. In: *Oncotarget* 5.10, pp. 2912–2917. DOI: 10.18632/oncotarget.2048.
- Elfakess, Rofa, Hadar Sinvani, Ora Haimov, Yuri Svitkin, Nahum Sonenberg, and Rivka Dikstein (2011). “Unique translation initiation of mRNAs-containing TISU element”. In: *Nucleic Acids Research* 39.17, pp. 7598–7609. DOI: 10.1093/nar/gkr484.
- Erhard, Florian, Anne Halenius, Cosima Zimmermann, Anne L’Hernault, Daniel J Kowalewski, Michael P Weekes, Stefan Stevanovic, Ralf Zimmer, and Lars Dölken (2018). “Improved Ribo-seq enables identification of cryptic translation events”. In: *Nature Methods* 15.5, pp. 363–366. DOI: 10.1038/nmeth.4631.
- Fabbri, Lucilla, Alina Chakraborty, Caroline Robert, and Stéphan Vagner (2021). “The plasticity of mRNA translation during cancer progression and therapy resistance”. In: *Nature Reviews Cancer* 21.9, pp. 558–577. DOI: 10.1038/s41568-021-00380-y.
- Fan, Zirong, Jian Zheng, Yixue Xue, Xiaobai Liu, Di Wang, Chunqing Yang, Jun Ma, Libo Liu, Xuelei Ruan, Zhenhua Wang, and Yunhui Liu (2018). “NR2C2-uORF targeting UCA1-miR-627-5p-NR2C2 feedback loop to regulate the malignant behaviors of glioma cells”. In: *Cell Death & Disease* 9.12, p. 1165. DOI: 10.1038/s41419-018-1149-x.
- Faure, Guilhem, Aleksey Y. Ogurtsov, Svetlana A. Shabalina, and Eugene V. Koonin (2016). “Role of mRNA structure in the control of protein folding”. In: *Nucleic Acids Research* 44.22, pp. 10898–10911. DOI: 10.1093/nar/gkw671.
- Fleiss, Joseph L. and Jacob Cohen (1973). “The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability”. In: *Educational and Psychological Measurement* 33.3, pp. 613–619. DOI: 10.1177/001316447303300309.
- Frankish, Adam et al. (2019). “GENCODE reference annotation for the human and mouse genomes”. In: *Nucleic Acids Research* 47 (D1), pp. D766–D773. DOI: 10.1093/nar/gky955.
- Frenkel-Morgenstern, Milana, Tamar Danon, Thomas Christian, Takao Igarashi, Lydia Cohen, Ya-Ming Hou, and Lars Juhl Jensen (2012). “Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels”. In: *Molecular Systems Biology* 8.1, p. 572. DOI: 10.1038/msb.2012.3.
- Gao, Na L., Zilong He, Qianhui Zhu, Puzi Jiang, Songnian Hu, and Wei-Hua Chen (2021). “Selection for Cheaper Amino Acids Drives Nucleotide Usage at the Start of Translation in Eukaryotic Genes”. In: *Genomics, Proteomics & Bioinformatics* 19.6, pp. 949–957. DOI: 10.1016/j.gpb.2021.03.002.
- Gelhausen, Rick, Teresa Müller, Sarah L Svensson, Omer S Alkhnbashi, Cynthia M Sharma, Florian Eggenhofer, and Rolf Backofen (2022). “RiboReport - benchmarking tools for ribosome profiling-based identification of open reading frames in bacteria”. In: *Briefings in Bioinformatics* 23.2, bbab549. DOI: 10.1093/bib/bbab549.

- Gillen, Sarah L., Chiara Giacomelli, Kelly Hodge, Sara Zanivan, Martin Bushell, and Ania Wilczynska (2021). “Differential regulation of mRNA fate by the human Ccr4-Not complex is driven by coding sequence composition and mRNA localization”. In: *Genome Biology* 22.1, p. 284. DOI: 10.1186/s13059-021-02494-w.
- Gillen, Sarah L., Joseph A. Waldron, and Martin Bushell (2021). “Codon optimality in cancer”. In: *Oncogene* 40.45, pp. 6309–6320. DOI: 10.1038/s41388-021-02022-x.
- Grossman, Robert L., Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt (2016). “Toward a Shared Vision for Cancer Genomic Data”. In: *New England Journal of Medicine* 375.12, pp. 1109–1112. DOI: 10.1056/NEJMp1607591.
- Gruda, Dritjon (2024). “Three ways ChatGPT helps me in my academic writing”. In: *Nature*, pp. d41586–024–01042–3. DOI: 10.1038/d41586-024-01042-3.
- Gunišová, Stanislava, Vladislava Hronová, Mahabub Pasha Mohammad, Alan G Hinnebusch, and Leoš Shivaya Valášek (2018). “Please do not recycle! Translation reinitiation in microbes and higher eukaryotes”. In: *FEMS Microbiology Reviews* 42.2, pp. 165–192. DOI: 10.1093/femsre/fux059.
- Haimov, Ora, Hadar Sinvani, and Rivka Dikstein (2015). “Cap-dependent, scanning-free translation initiation mechanisms”. In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1849.11, pp. 1313–1318. DOI: 10.1016/j.bbagrm.2015.09.006.
- Hia, Fabian, Sheng Fan Yang, Yuichi Shichino, Masanori Yoshinaga, Yasuhiro Murakawa, Alexis Vandenbon, Akira Fukao, Toshinobu Fujiwara, Markus Landthaler, Tohru Natsume, Shungo Adachi, Shintaro Iwasaki, and Osamu Takeuchi (2019). “Codon bias confers stability to human mRNAs”. In: *EMBO reports* 20.11, e48220. DOI: 10.15252/embr.201948220.
- Hinnebusch, A. G., I. P. Ivanov, and N. Sonenberg (2016). “Translational control by 5'-untranslated regions of eukaryotic mRNAs”. In: *Science* 352.6292, pp. 1413–1416. DOI: 10.1126/science.aad9868.
- Huang, Lamei, Zhixing Guo, Fang Wang, and Liwu Fu (2021). “KRAS mutation: from undruggable to drug-gable in cancer”. In: *Signal Transduction and Targeted Therapy* 6.1, p. 386. DOI: 10.1038/s41392-021-00780-4.
- Huang, Zixia and Emma C. Teeling (2017). “ExUTR: a novel pipeline for large-scale prediction of 3'-UTR sequences from NGS data”. In: *BMC Genomics* 18.1, p. 847. DOI: 10.1186/s12864-017-4241-1.
- Hwang, Yunha, Andre L. Cornman, Elizabeth H. Kellogg, Sergey Ovchinnikov, and Peter R. Girguis (2024). “Genomic language model predicts protein co-regulation and function”. In: *Nature Communications* 15.1, p. 2880. DOI: 10.1038/s41467-024-46947-9.
- Inc., PerkinElmer (2019). *NEXTflex-Small-RNA-Seq-v3_19.01*.
- Ingolia, Nicholas T, Gloria A Brar, Silvia Rouskin, Anna M McGeachy, and Jonathan S Weissman (2012). “The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments”. In: *Nature Protocols* 7.8, pp. 1534–1550. DOI: 10.1038/nprot.2012.086.
- Ingolia, Nicholas T. (2016). “Ribosome Footprint Profiling of Translation throughout the Genome”. In: *Cell* 165.1, pp. 22–33. DOI: 10.1016/j.cell.2016.02.066.

- Ingolia, Nicholas T., Liana F. Lareau, and Jonathan S. Weissman (2011). “Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes”. In: *Cell* 147.4, pp. 789–802. DOI: 10.1016/j.cell.2011.10.002.
- Ivanov, I. P., G. Loughran, and J. F. Atkins (2008). “uORFs with unusual translational start codons autoregulate expression of eukaryotic ornithine decarboxylase homologs”. In: *Proceedings of the National Academy of Sciences* 105.29, pp. 10079–10084. DOI: 10.1073/pnas.0801590105.
- Ivanov, Ivaylo P., Andrew E. Firth, Audrey M. Michel, John F. Atkins, and Pavel V. Baranov (2011). “Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences”. In: *Nucleic Acids Research* 39.10, pp. 4220–4234. DOI: 10.1093/nar/gkr007.
- Jackson, Richard J., Christopher U. T. Hellen, and Tatyana V. Pestova (2010). “The mechanism of eukaryotic translation initiation and principles of its regulation”. In: *Nature Reviews Molecular Cell Biology* 11.2, pp. 113–127. DOI: 10.1038/nrm2838.
- Jayaram, Divya Ram, Sigal Frost, Chanan Argov, Vijayasteltar Belsamma Liju, Nikhil Ponnoor Anto, Amitha Muraleedharan, Assaf Ben-Ari, Rose Sinay, Ilan Smoly, Ofra Novoplansky, Noah Isakov, Debra Toiber, Chen Keasar, Moshe Elkabets, Esti Yeger-Lotem, and Etta Livneh (2021). “Unraveling the hidden role of a uORF-encoded peptide as a kinase inhibitor of PKCs”. In: *Proceedings of the National Academy of Sciences* 118.40, e2018899118. DOI: 10.1073/pnas.2018899118.
- Ji, Zhe (2018). “RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling”. In: *Current Protocols in Molecular Biology* 124.1, e67. DOI: 10.1002/cpmb.67.
- Jin, Xiaoping, Eileen Turcott, Silvia Englehardt, Gregory J. Mize, and David R. Morris (2003). “The Two Upstream Open Reading Frames of Oncogene mdm2 Have Different Translational Regulatory Properties”. In: *Journal of Biological Chemistry* 278.28, pp. 25716–25721. DOI: 10.1074/jbc.M300316200.
- Kearse, Michael G. and Jeremy E. Wilusz (2017). “Non-AUG translation: a new start for protein synthesis in eukaryotes”. In: *Genes & Development* 31.17, pp. 1717–1731. DOI: 10.1101/gad.305250.117.
- Keene, Jack D. (2007). “RNA regulons: coordination of post-transcriptional events”. In: *Nature Reviews Genetics* 8.7, pp. 533–543. DOI: 10.1038/nrg2111.
- Kim, Daehwan, Ben Langmead, and Steven L. Salzberg (2015). “HISAT: a fast spliced aligner with low memory requirements”. In: *Nature Methods* 12.4, pp. 357–360. DOI: 10.1038/nmeth.3317.
- Knight, John R.P., Constantinos Alexandrou, George L. Skalka, Nikola Vlahov, Kathryn Pennel, Leah Officer, Ana Teodosio, Georgios Kanellos, David M. Gay, Sebastian May-Wilson, Ewan M. Smith, Arafath K. Najumudeen, Kathryn Gilroy, Rachel A. Ridgway, Dustin J. Flanagan, Rachael C.L. Smith, Laura McDonald, Craig MacKay, Anne Cheasty, Kerri McArthur, Emma Stanway, Joshua D. Leach, Rene Jackstadt, Joseph A. Waldron, Andrew D. Campbell, Georgios Vlachogiannis, Nicola Valeri, Kevin M. Haigis, Nahum Sonnenberg, Christopher G. Proud, Neil P. Jones, Martin E. Swarbrick, Heather J. McKinnon, William J. Faller, John Le Quesne, Joanne Edwards, Anne E. Willis, Martin Bushell, and Owen J. Sansom (2021). “MNK Inhibition Sensitizes *KRAS* -Mutant Colorectal Cancer to mTORC1 Inhibition by Reducing eIF4E Phos-

- phorylation and c-MYC Expression". In: *Cancer Discovery* 11.5, pp. 1228–1247. DOI: 10.1158/2159-8290.CD-20-0652.
- Kosciuczuk, Ewa M., Diana Saleiro, and Leonidas C. Plataniias (2017). "Dual targeting of eIF4E by blocking MNK and mTOR pathways in leukemia". In: *Cytokine* 89, pp. 116–121. DOI: 10.1016/j.cyto.2016.01.024.
- Kozak, M. (2001). "Constraints on reinitiation of translation in mammals". In: *Nucleic Acids Research* 29.24, pp. 5226–5232. DOI: 10.1093/nar/29.24.5226.
- Kumar, Mayank, Vivek Srinivas, and Swati Patankar (2015). "Upstream AUGs and upstream ORFs can regulate the downstream ORF in *Plasmodium falciparum*". In: *Malaria Journal* 14.1, p. 512. DOI: 10.1186/s12936-015-1040-5.
- Lehninger, Albert L., David Lee Nelson, and Michael M. Cox (2005). *Lehninger principles of biochemistry*. 4th ed. New York: W.H. Freeman. 1 p. ISBN: 978-0-7167-4339-2.
- Leppek, Kathrin, Rhiju Das, and Maria Barna (2018). "Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them". In: *Nature Reviews Molecular Cell Biology* 19.3, pp. 158–174. DOI: 10.1038/nrm.2017.103.
- Löchel, Hannah F, Dominic Eger, Theodor Sperlea, and Dominik Heider (2020). "Deep learning on chaos game representation for proteins". In: *Bioinformatics* 36.1. Ed. by Jonathan Wren, pp. 272–279. DOI: 10.1093/bioinformatics/btz493.
- Lorenz, Ronny, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker (2011). "ViennaRNA Package 2.0". In: *Algorithms for Molecular Biology* 6.1, p. 26. DOI: 10.1186/1748-7188-6-26.
- Lu, Tian, Liujia Qian, Yuting Xie, Qiushi Zhang, Wei Liu, Weigang Ge, Yi Zhu, Lijia Ma, Cheng Zhang, and Tiannan Guo (2022). "Tissue-Characteristic Expression of Mouse Proteome". In: *Molecular & Cellular Proteomics* 21.10, p. 100408. DOI: 10.1016/j.mcpro.2022.100408.
- Magny, Emile G., Jose Ignacio Pueyo, Frances M.G. Pearl, Miguel Angel Cespedes, Jeremy E. Niven, Sarah A. Bishop, and Juan Pablo Couso (2013). "Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames". In: *Science* 341.6150, pp. 1116–1120. DOI: 10.1126/science.1238802.
- Manning, Kassie S. and Thomas A. Cooper (2017). "The roles of RNA processing in translating genotype to phenotype". In: *Nature Reviews Molecular Cell Biology* 18.2, pp. 102–114. DOI: 10.1038/nrm.2016.139.
- Martinez, Thomas F., Qian Chu, Cynthia Donaldson, Dan Tan, Maxim N. Shokhirev, and Alan Saghatelian (2020). "Accurate annotation of human protein-coding small open reading frames". In: *Nature Chemical Biology* 16.4, pp. 458–468. DOI: 10.1038/s41589-019-0425-0.
- Matoulikova, Eva, Eva Michalova, Borivoj Vojtesek, and Roman Hrstka (2012). "The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells." In: *RNA Biology* 9.5, pp. 563–576. DOI: 10.4161/rna.20231.

- McGillivray, Patrick, Russell Ault, Mayur Pawashe, Robert Kitchen, Suganthi Balasubramanian, and Mark Gerstein (2018). “A comprehensive catalog of predicted functional upstream open reading frames in humans”. In: *Nucleic Acids Research* 46.7, pp. 3326–3338. DOI: 10.1093/nar/gky188.
- McGlinicy, Nicholas J. and Nicholas T. Ingolia (2017). “Transcriptome-wide measurement of translation by ribosome profiling”. In: *Methods* 126, pp. 112–129. DOI: 10.1016/j.ymeth.2017.05.028.
- Michel, Audrey M., Kingshuk Roy Choudhury, Andrew E. Firth, Nicholas T. Ingolia, John F. Atkins, and Pavel V. Baranov (2012). “Observation of dually decoded regions of the human genome using ribosome profiling data”. In: *Genome Research* 22.11, pp. 2219–2229. DOI: 10.1101/gr.133249.111.
- Mockenhaupt, Stefan and Eugene V. Makeyev (2015). “Non-coding functions of alternative pre-mRNA splicing in development”. In: *Seminars in Cell & Developmental Biology* 47-48, pp. 32–39. DOI: 10.1016/j.semcd.2015.10.018.
- Mohammad, Mahabub Pasha, Anna Smirnova, Stanislava Gunišová, and Leoš Shivaya Valášek (2021). “eIF4G is retained on ribosomes elongating and terminating on short upstream ORFs to control reinitiation in yeast”. In: *Nucleic Acids Research* 49.15, pp. 8743–8756. DOI: 10.1093/nar/gkab652.
- Morales, Joannella, Shashikant Pujar, Jane E. Loveland, Alex Astashyn, Ruth Bennett, Andrew Berry, Eric Cox, Claire Davidson, Olga Ermolaeva, Catherine M. Farrell, Reham Fatima, Laurent Gil, Tamara Goldfarb, Jose M. Gonzalez, Diana Haddad, Matthew Hardy, Toby Hunt, John Jackson, Vinita S. Joardar, Michael Kay, Vamsi K. Kodali, Kelly M. McGarvey, Aoife McMahon, Jonathan M. Mudge, Daniel N. Murphy, Michael R. Murphy, Bhanu Rajput, Sanjida H. Rangwala, Lillian D. Riddick, Françoise Thibaud-Nissen, Glen Threadgold, Anjana R. Vatsan, Craig Wallin, David Webb, Paul Flicek, Ewan Birney, Kim D. Pruitt, Adam Frankish, Fiona Cunningham, and Terence D. Murphy (2022). “A joint NCBI and EMBL-EBI transcript set for clinical genomics and research”. In: *Nature* 604.7905, pp. 310–315. DOI: 10.1038/s41586-022-04558-8.
- 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 (2022). “Standardized annotation of translated open reading frames”. In: *Nature Biotechnology* 40.7, pp. 994–999. DOI: 10.1038/s41587-022-01369-0.
- Munzarová, Vanda, Josef Pánek, Stanislava Gunišová, István Dányi, Béla Szamecz, and Leoš Shivaya Valášek (2011). “Translation Reinitiation Relies on the Interaction between eIF3a/TIF32 and Progressively Folded cis-Acting mRNA Elements Preceding Short uORFs”. In: *PLoS Genetics* 7.7. Ed. by Hideko Kaji, e1002137. DOI: 10.1371/journal.pgen.1002137.

- Nguyen, Christine L., Richard Possemato, Erica L. Bauerlein, Anyong Xie, Ralph Scully, and William C. Hahn (2012). “Nek4 Regulates Entry into Replicative Senescence and the Response to DNA Damage in Human Fibroblasts”. In: *Molecular and Cellular Biology* 32.19, pp. 3963–3977. DOI: 10.1128/MCB.00436-12.
- Orgogozo, V., A.E. Peluffo, and B. Morizot (2016). “The “Mendelian Gene” and the “Molecular Gene””. In: *Current Topics in Developmental Biology*. Vol. 119. Elsevier, pp. 1–26. ISBN: 978-0-12-417194-7. DOI: 10.1016/bs.ctdb.2016.03.002.
- Palaniswamy, Ramesh, Stephan Teglund, Matthias Lauth, Peter G Zaphiropoulos, and Takashi Shimokawa (2010). “Genetic variations regulate alternative splicing in the 5’ untranslated regions of the mouse glioma-associated oncogene 1, Gli1”. In: *BMC Molecular Biology* 11.1, p. 32. DOI: 10.1186/1471-2199-11-32.
- Parra, Columba de la, Amanda Ernlund, Amandine Alard, Kelly Ruggles, Beatrix Ueberheide, and Robert J. Schneider (2018). “A widespread alternate form of cap-dependent mRNA translation initiation”. In: *Nature Communications* 9.1, p. 3068. DOI: 10.1038/s41467-018-05539-0.
- Patraquim, Pedro, Muhammad Ali Shahzad Mumtaz, José Ignacio Pueyo, Julie Louise Aspdén, and Juan-Pablo Couso (2020). “Developmental regulation of canonical and small ORF translation from mRNAs”. In: *Genome Biology* 21.1, p. 128. DOI: 10.1186/s13059-020-02011-5.
- Pearson, William R. (2013). “An Introduction to Sequence Similarity (“Homology”) Searching”. In: *Current Protocols in Bioinformatics* 42.1. DOI: 10.1002/0471250953.bi0301s42.
- Peterson, Jackson, Siqi Li, Erin Kaltenbrun, Ozgun Erdogan, and Christopher M. Counter (2020). “Expression of transgenes enriched in rare codons is enhanced by the MAPK pathway”. In: *Scientific Reports* 10.1, p. 22166. DOI: 10.1038/s41598-020-78453-5.
- Plotkin, Joshua B. and Grzegorz Kudla (2011). “Synonymous but not the same: the causes and consequences of codon bias”. In: *Nature Reviews Genetics* 12.1, pp. 32–42. DOI: 10.1038/nrg2899.
- Proud, Christopher G. (2019). “Phosphorylation and Signal Transduction Pathways in Translational Control”. In: *Cold Spring Harbor Perspectives in Biology* 11.7, a033050. DOI: 10.1101/cshperspect.a033050.
- Qing, Guoliang, Bo Li, Annette Vu, Nicolas Skuli, Zandra E. Walton, Xueyuan Liu, Patrick A. Mayes, David R. Wise, Craig B. Thompson, John M. Maris, Michael D. Hogarty, and M. Celeste Simon (2012). “ATF4 Regulates MYC-Mediated Neuroblastoma Cell Death upon Glutamine Deprivation”. In: *Cancer Cell* 22.5, pp. 631–644. DOI: 10.1016/j.ccr.2012.09.021.
- Rahmani, Fatemeh, Maureen Hummel, Jolanda Schuurmans, Anika Wiese-Klinkenberg, Sjef Smeekens, and Johannes Hanson (2009). “Sucrose Control of Translation Mediated by an Upstream Open Reading Frame-Encoded Peptide”. In: *Plant Physiology* 150.3, pp. 1356–1367. DOI: 10.1104/pp.109.136036.
- Rudolph, Konrad L. M., Bianca M. Schmitt, Diego Villar, Robert J. White, John C. Marioni, Claudia Kutter, and Duncan T. Odom (2016). “Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States”. In: *PLOS Genetics* 12.5. Ed. by Nicolas Galtier, e1006024. DOI: 10.1371/journal.pgen.1006024.

- 2871 Ruiz-Ruano, Francisco J., María Dolores López-León, Josefa Cabrero, and Juan Pedro M. Camacho (2016).
 2872 “High-throughput analysis of the satellitome illuminates satellite DNA evolution”. In: *Scientific Reports*
 2873 6.1, p. 28333. DOI: 10.1038/srep28333.
- 2874 Russell, Paul J, Jacob A Slivka, Elaina P Boyle, Arthur H M Burghes, and Michael G Kears (2023). “Transla-
 2875 tion reinitiation after uORFs does not fully protect mRNAs from nonsense-mediated decay”. In: *RNA*.
- 2876 Santos, Alberto, Rasmus Wernersson, and Lars Juhl Jensen (2015). “Cyclebase 3.0: a multi-organism database
 2877 on cell-cycle regulation and phenotypes”. In: *Nucleic Acids Research* 43 (D1), pp. D1140–D1144. DOI:
 2878 10.1093/nar/gku1092.
- 2879 Saxton, Robert A. and David M. Sabatini (2017). “mTOR Signaling in Growth, Metabolism, and Disease”. In:
 2880 *Cell* 168.6, pp. 960–976. DOI: 10.1016/j.cell.2017.02.004.
- 2881 Schulz, Julia, Nancy Mah, Martin Neuenschwander, Tabea Kischka, Richard Ratei, Peter M. Schlag, Esmeralda
 2882 Castaños-Vélez, Iduna Fichtner, Per-Ulf Tunn, Carsten Denkert, Oliver Klaas, Wolfgang E. Berdel, Jens P.
 2883 von Kries, Wojciech Makalowski, Miguel A. Andrade-Navarro, Achim Leutz, and Klaus Wethmar (2018).
 2884 “Loss-of-function uORF mutations in human malignancies”. In: *Scientific Reports* 8.1, p. 2395. DOI: 10.
 2885 1038/s41598-018-19201-8.
- 2886 Schuster, Samantha L. and Andrew C. Hsieh (2019). “The Untranslated Regions of mRNAs in Cancer”. In:
 2887 *Trends in Cancer* 5.4, pp. 245–262. DOI: 10.1016/j.trecan.2019.02.011.
- 2888 Shatskikh, Aleksei S., Alexei A. Kotov, Vladimir E. Adashev, Sergei S. Bazylev, and Ludmila V. Olenina
 2889 (2020). “Functional Significance of Satellite DNAs: Insights From Drosophila”. In: *Frontiers in Cell and*
 2890 *Developmental Biology* 8, p. 312. DOI: 10.3389/fcell.2020.00312.
- 2891 Siepel, Adam, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram
 2892 Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson,
 2893 Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler (2005). “Evolutionarily conserved
 2894 elements in vertebrate, insect, worm, and yeast genomes”. In: *Genome Research* 15.8, pp. 1034–1050. DOI:
 2895 10.1101/gr.3715005.
- 2896 Starck, S. R., J. C. Tsai, K. Chen, M. Shodiya, L. Wang, K. Yahiro, M. Martins-Green, N. Shastri, and P. Walter
 2897 (2016). “Translation from the 5’ untranslated region shapes the integrated stress response”. In: *Science*
 2898 351.6272, aad3867–aad3867. DOI: 10.1126/science.aad3867.
- 2899 Steri, Maristella, M. Laura Idda, Michael B. Whalen, and Valeria Orrù (2018). “Genetic variants in mRNA
 2900 untranslated regions”. In: *Wiley Interdisciplinary Reviews: RNA* 9.4, e1474. DOI: 10.1002/wrna.1474.
- 2901 Suntsova, Maria, Nurshat Gaifullin, Daria Allina, Alexey Reshetun, Xinmin Li, Larisa Mendeleeva, Vadim
 2902 Surin, Anna Sergeeva, Pavel Spirin, Vladimir Prassolov, Alexander Morgan, Andrew Garazha, Maxim
 2903 Sorokin, and Anton Buzdin (2019). “Atlas of RNA sequencing profiles for normal human tissues”. In:
 2904 *Scientific Data* 6.1, p. 36. DOI: 10.1038/s41597-019-0043-4.
- 2905 Supek, Fran, Belén Miñana, Juan Valcárcel, Toni Gabaldón, and Ben Lehner (2014). “Synonymous Mutations
 2906 Frequently Act as Driver Mutations in Human Cancers”. In: *Cell* 156.6, pp. 1324–1335. DOI: 10.1016/j.
 2907 cell.2014.01.051.

- 2908 Szamecz, Béla, Edit Rutkai, Lucie Cuchalová, Vanda Munzarová, Anna Herrmannová, Klaus H. Nielsen,
 2909 Laxminarayana Burela, Alan G. Hinnebusch, and Leoš Valášek (2008). “eIF3a cooperates with sequences
 2910 5 of uORF1 to promote resumption of scanning by post-termination ribosomes for reinitiation on *GCN4*
 2911 mRNA”. In: *Genes & Development* 22.17, pp. 2414–2425. DOI: 10.1101/gad.480508.
- 2912 Takahara, Terunao, Yuna Amemiya, Risa Sugiyama, Masatoshi Maki, and Hideki Shibata (2020). “Amino acid-
 2913 dependent control of mTORC1 signaling: a variety of regulatory modes”. In: *Journal of Biomedical Science*
 2914 27.1, p. 87. DOI: 10.1186/s12929-020-00679-2.
- 2915 Taliaferro, J. Matthew, Nicole J. Lambert, Peter H. Sudmant, Daniel Dominguez, Jason J. Merkin, Maria S.
 2916 Alexis, Cassandra A. Bazile, and Christopher B. Burge (2016). “RNA Sequence Context Effects Measured
 2917 In Vitro Predict In Vivo Protein Binding and Regulation”. In: *Molecular Cell* 64.2, pp. 294–306. DOI:
 2918 10.1016/j.molcel.2016.08.035.
- 2919 Telenti, Amalio, Levi C. T. Pierce, William H. Biggs, Julia di Iulio, Emily H. M. Wong, Martin M. Fabani,
 2920 Ewen F. Kirkness, Ahmed Moustafa, Naisha Shah, Chao Xie, Suzanne C. Brewerton, Nadeem Bulsara,
 2921 Chad Garner, Gary Metzker, Efren Sandoval, Brad A. Perkins, Franz J. Och, Yaron Turpaz, and J. Craig
 2922 Venter (2016). “Deep sequencing of 10,000 human genomes”. In: *Proceedings of the National Academy of*
 2923 *Sciences* 113.42, pp. 11901–11906. DOI: 10.1073/pnas.1613365113.
- 2924 The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R
 2925 Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart (2013).
 2926 “The Cancer Genome Atlas Pan-Cancer analysis project”. In: *Nature Genetics* 45.10, pp. 1113–1120. DOI:
 2927 10.1038/ng.2764.
- 2928 Thind, Amarinder Singh and Somdatta Sinha (2023). “Using Chaos-Game-Representation for Analysing the
 2929 SARS-CoV-2 Lineages, Newly Emerging Strains and Recombinants”. In: *Current Genomics* 24.3, pp. 187–
 2930 195. DOI: 10.2174/0113892029264990231013112156.
- 2931 Thomson, D.J. (1982). “Spectrum estimation and harmonic analysis”. In: *Proceedings of the IEEE* 70.9,
 2932 pp. 1055–1096. DOI: 10.1109/PROC.1982.12433.
- 2933 Treangen, Todd J. and Steven L. Salzberg (2012). “Repetitive DNA and next-generation sequencing: computa-
 2934 tional challenges and solutions”. In: *Nature Reviews Genetics* 13.1, pp. 36–46. DOI: 10.1038/nrg3117.
- 2935 Tung, Kuo-Feng, Chao-Yu Pan, Chao-Hsin Chen, and Wen-chang Lin (2020). “Top-ranked expressed gene
 2936 transcripts of human protein-coding genes investigated with GTEx dataset”. In: *Scientific Reports* 10.1,
 2937 p. 16245. DOI: 10.1038/s41598-020-73081-5.
- 2938 Vatansever, Sezen, Burak Erman, and Zeynep H. Gümüş (2019). “Oncogenic G12D mutation alters local con-
 2939 formations and dynamics of K-Ras”. In: *Scientific Reports* 9.1, p. 11730. DOI: 10.1038/s41598-019-
 2940 48029-z.
- 2941 Vejnar, Charles E., Mario Abdel Messih, Carter M. Takacs, Valeria Yartseva, Panos Oikonomou, Romain Chris-
 2942 tiano, Marlon Stoeckius, Stephanie Lau, Miler T. Lee, Jean-Denis Beaudoin, Damir Musaev, Hiba Darwich-
 2943 Codore, Tobias C. Walther, Saeed Tavazoie, Daniel Cifuentes, and Antonio J. Giraldez (2019). “Genome
 2944 wide analysis of 3’ UTR sequence elements and proteins regulating mRNA stability during maternal-to-

- zygotic transition in zebrafish”. In: *Genome Research* 29.7, pp. 1100–1114. DOI: 10.1101/gr.245159.118.
- Wagner, Susan, Jonathan Bohlen, Anna Herrmannová, Jan Jelínek, Thomas Preiss, Leoš Shivaya Valášek, and Aurelio A. Teleman (2022). “Selective footprinting of 40S and 80S ribosome subpopulations (Sel-TCP-seq) to study translation and its control”. In: *Nature Protocols* 17.10, pp. 2139–2187. DOI: 10.1038/s41596-022-00708-4.
- Wagner, Susan, Anna Herrmannová, Vladislava Hronová, Stanislava Gunišová, Neelam D. Sen, Ross D. Hannan, Alan G. Hinnebusch, Nikolay E. Shirokikh, Thomas Preiss, and Leoš Shivaya Valášek (2020). “Selective Translation Complex Profiling Reveals Staged Initiation and Co-translational Assembly of Initiation Factor Complexes”. In: *Molecular Cell* 79.4, 546–560.e7. DOI: 10.1016/j.molcel.2020.06.004.
- Wang, Qiuyi and Yuanhui Mao (2023). “Principles, challenges, and advances in ribosome profiling: from bulk to low-input and single-cell analysis”. In: *Advanced Biotechnology* 1.4, p. 6. DOI: 10.1007/s44307-023-00006-4.
- Wethmar, Klaus, Adriano Barbosa-Silva, Miguel A. Andrade-Navarro, and Achim Leutz (2014). “uORFdb—a comprehensive literature database on eukaryotic uORF biology”. In: *Nucleic Acids Research* 42 (D1), pp. D60–D67. DOI: 10.1093/nar/gkt952.
- Wongfieng, Wipaporn, Amonrat Jumnainsong, Yaovalux Chamgramol, Banchob Sriipa, and Chanvit Leelayuwat (2017). “5’-UTR and 3’-UTR Regulation of MICB Expression in Human Cancer Cells by Novel microRNAs”. In: *Genes* 8.9, p. 213. DOI: 10.3390/genes8090213.
- Wu, Qiushuang, Matthew Wright, Madelaine M Gogol, William D Bradford, Ning Zhang, and Ariel A Bazzini (2020). “Translation of small downstream ORFs enhances translation of canonical main open reading frames”. In: *The EMBO Journal* 39.17. DOI: 10.15252/embj.2020104763.
- Xiao, Wenjing, Yang Sun, Jinpeng Xu, Na Zhang, and Lina Dong (2022). “uORF-Mediated Translational Regulation of ATF4 Serves as an Evolutionarily Conserved Mechanism Contributing to Non-Small-Cell Lung Cancer (NSCLC) and Stress Response”. In: *Journal of Molecular Evolution* 90.5, pp. 375–388. DOI: 10.1007/s00239-022-10068-y.
- Xiao, Zhengtao, Rongyao Huang, Xudong Xing, Yuling Chen, Haiteng Deng, and Xuerui Yang (2018). “De novo annotation and characterization of the translome with ribosome profiling data”. In: *Nucleic Acids Research* 46.10, e61–e61. DOI: 10.1093/nar/gky179.
- Xu, Zhiyang, Shaohang Wu, Jiahua Tu, Mingyang Wang, Weicheng Liang, Jiangdong Cheng, Jun Guan, and Jianxin Xu (2024). “RACGAP1 promotes lung cancer cell proliferation through the PI3K/AKT signaling pathway”. In: *Scientific Reports* 14.1, p. 8694. DOI: 10.1038/s41598-024-58539-0.
- Zhang, Hanyu, Che-Lun Hung, Meiyuan Liu, Xiaoye Hu, and Yi-Yang Lin (2019). “NCNet: Deep Learning Network Models for Predicting Function of Non-coding DNA”. In: *Frontiers in Genetics* 10, p. 432. DOI: 10.3389/fgene.2019.00432.

- 2980 Zhang, Hong, Yirong Wang, Xinkai Wu, Xiaolu Tang, Changcheng Wu, and Jian Lu (2021). “Determinants
2981 of genome-wide distribution and evolution of uORFs in eukaryotes”. In: *Nature Communications* 12.1,
2982 p. 1076. DOI: 10.1038/s41467-021-21394-y.
- 2983 Zhang, Lingang, Simon Kasif, Charles R. Cantor, and Natalia E. Broude (2004). “GC/AT-content spikes as
2984 genomic punctuation marks”. In: *Proceedings of the National Academy of Sciences* 101.48, pp. 16855–
2985 16860. DOI: 10.1073/pnas.0407821101.
- 2986 Zhao, Boxuan Simen, Ian A. Roundtree, and Chuan He (2017). “Post-transcriptional gene regulation by mRNA
2987 modifications”. In: *Nature Reviews Molecular Cell Biology* 18.1, pp. 31–42. DOI: 10.1038/nrm.2016.
2988 132.
- 2989 Ziemann, Mark, Antony Kaspi, and Assam El-Osta (2016). “Evaluation of microRNA alignment techniques”.
2990 In: *RNA* 22.8, pp. 1120–1138. DOI: 10.1261/rna.055509.115.
- 2991 Zuccotti, Paola, Daniele Peroni, Valentina Potrich, Alessandro Quattrone, and Erik Dassi (2020). “Hypercon-
2992 served Elements in Human 5’UTRs Shape Essential Post-transcriptional Regulatory Networks”. In: *Fron-
2993 tiers in Molecular Biosciences* 7, p. 220. DOI: 10.3389/fmolb.2020.00220.

Appendices

Gene ID	Gene Name	Periodic Rank	Stage
ENSG00000101412	E2F1	1.476	G1-S
ENSG00000105173	CCNE1	1.999	G1-S
ENSG00000175305	CCNE2	1.749	G1-S
ENSG00000077514	POLD3	1.995	G1-S
ENSG00000197724	PHF2	2.229	G1-S
ENSG00000160336	ZNF761	2.435	G1-S
ENSG00000075218	GTSE1	1.953	G2-M
ENSG00000121957	GPSM2	1.855	G2-M
ENSG00000162063	CCNF	0.945	G2-M
ENSG00000134222	PSRC1	1.015	G2-M
ENSG00000146648	EGFR	2.432	G2-M
...	...	...	...

Table 8: **Gene information and periodic stage.** This table lists the Ensembl gene IDs, with the periodic rank indicating the level of periodicity—lower ranks denote higher periodicity. The ‘Stage’ column identifies the cell cycle phase during which peak expression occurs.

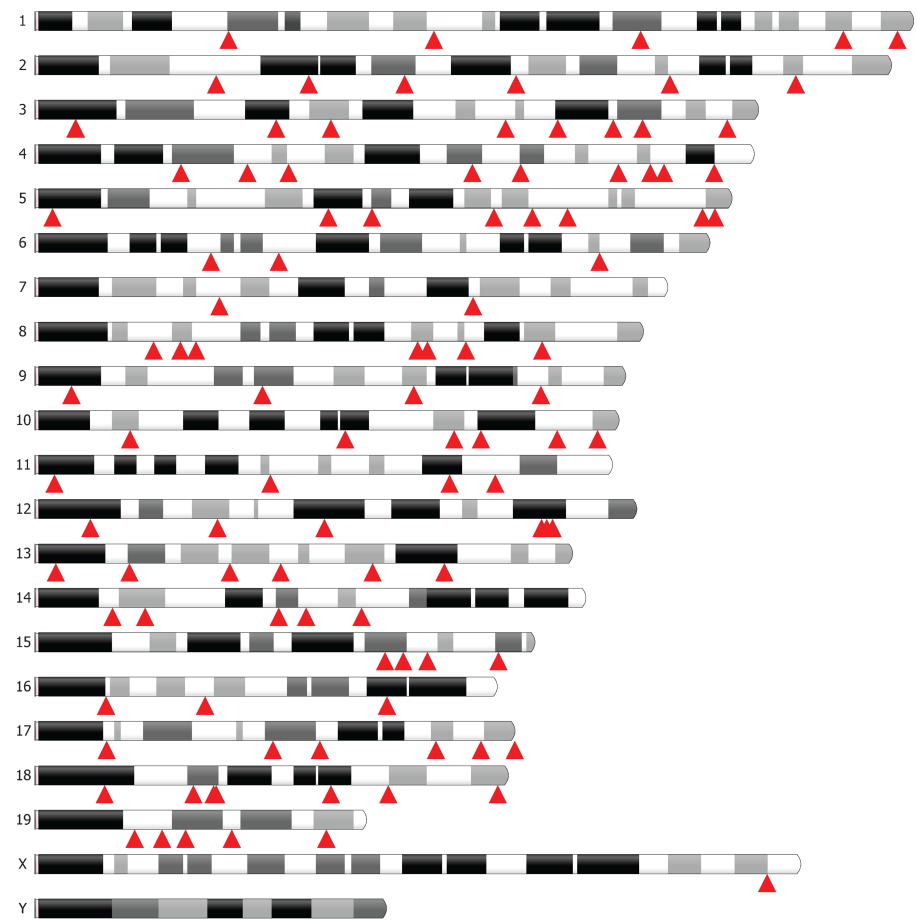


Figure S1: **Homologous hits of an uORF (Rpl29) on mouse chromosome ideogram.** Each red triangle denotes a region which have over 90% sequence identity to an uORF of Rpl29 in ENSMUST00000059802.

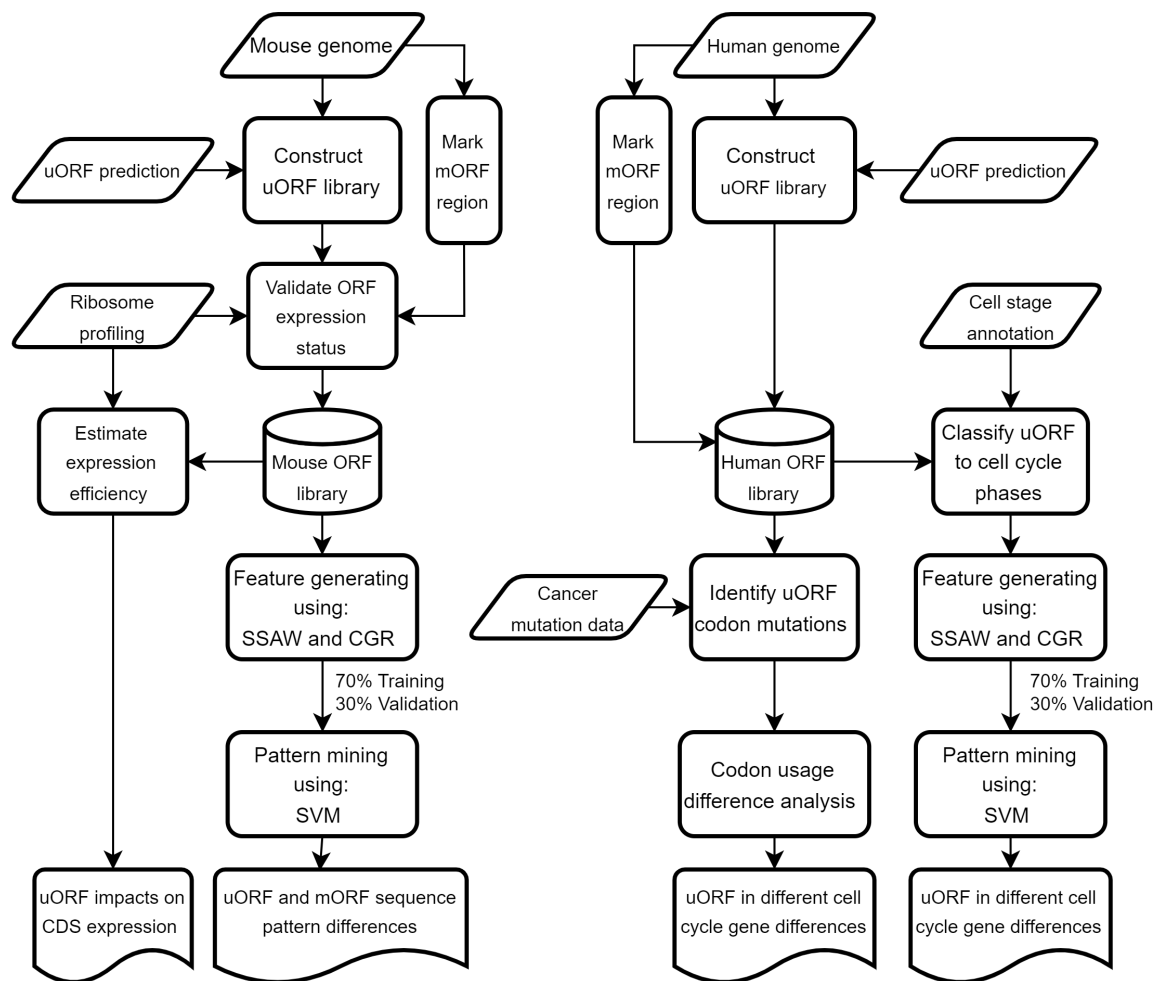


Figure S2: **uORF identification workflow of Year 1.** Analyses were performed based on *Mus musculus* and *Homo sapiens* genome assemblies from Ensembl (Cunningham et al. 2022). The uORF prediction data were obtained from Zhang et al. (2021). The Cell stage annotations were obtained from Cyclebase (Santos, Wernersson, and Jensen 2015). The ribosome profiling data was provided by Joseph Walrdon (Bushell's lab; unpublished data). The cancer mutation data was gathered from TCGA (Grossman et al. 2016). All analyses were performed in R and BioConductor.

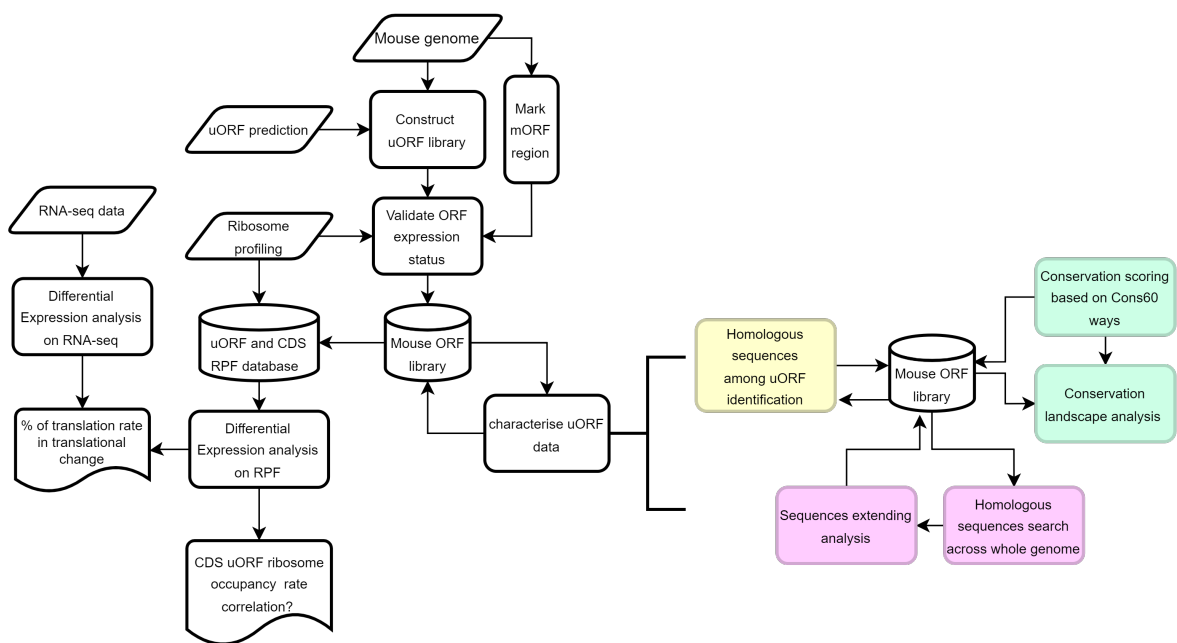


Figure S3: **ORF analysis workflow of Year 2.** This workflow was developed based on the previous year's ORF discovery workflow. New feature analysis modules were introduced. The paralogous sequence search and conservation comparison were performed.

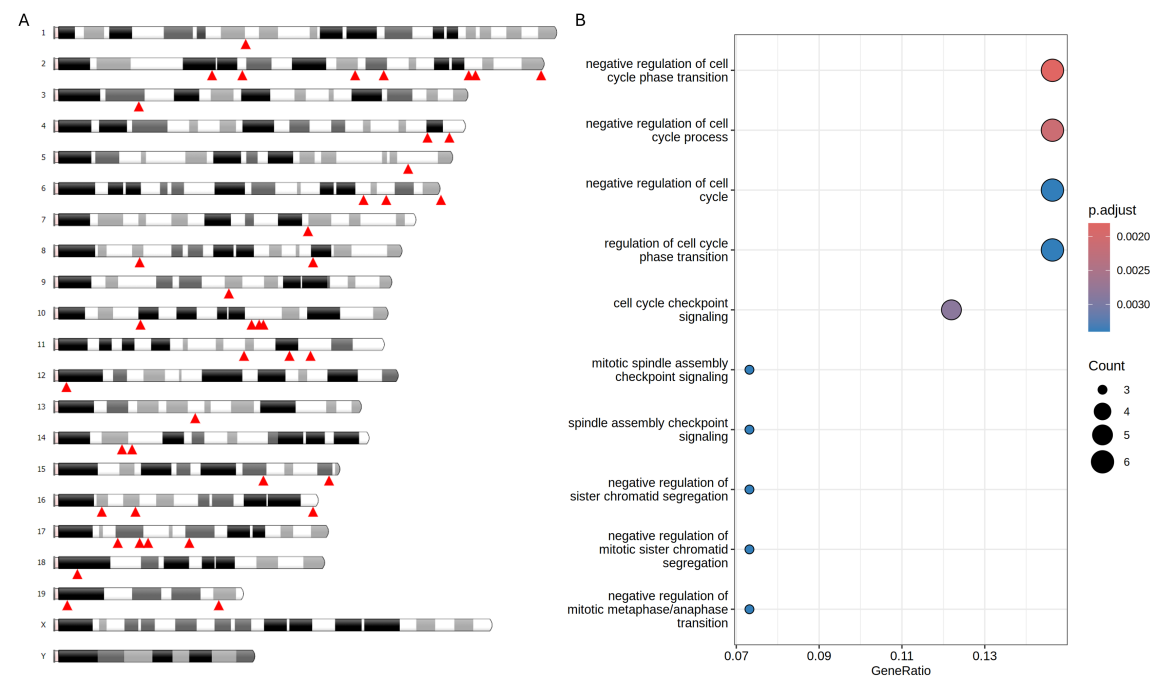


Figure S4: **Homologous hits of an uORF (Ptrh2) on mouse chromosome ideogram with corresponding GSEA.** . Figure A depicts the ideogram showing the homologous sequences of a Ptrh2 uORF, where each red triangle represents a region with over 90% sequence identity. Figure B presents the gene set enrichment analysis based on the Gene Ontology of biological processes.

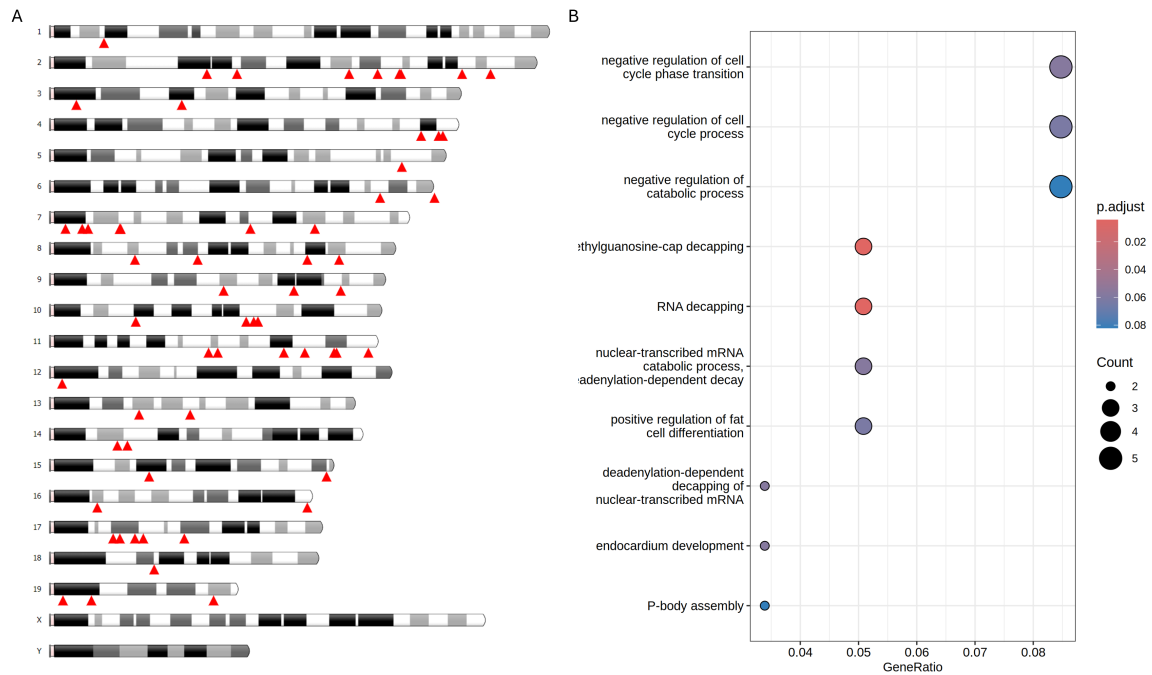


Figure S5: **Homologous hits of an uORF (Nek4) on mouse chromosome ideogram with corresponding GSEA.** . Figure A depicts the ideogram showing the homologous sequences of a Nek4 uORF, where each red triangle represents a region with over 90% sequence identity. Figure B presents the gene set enrichment analysis based on the Gene Ontology of biological processes.

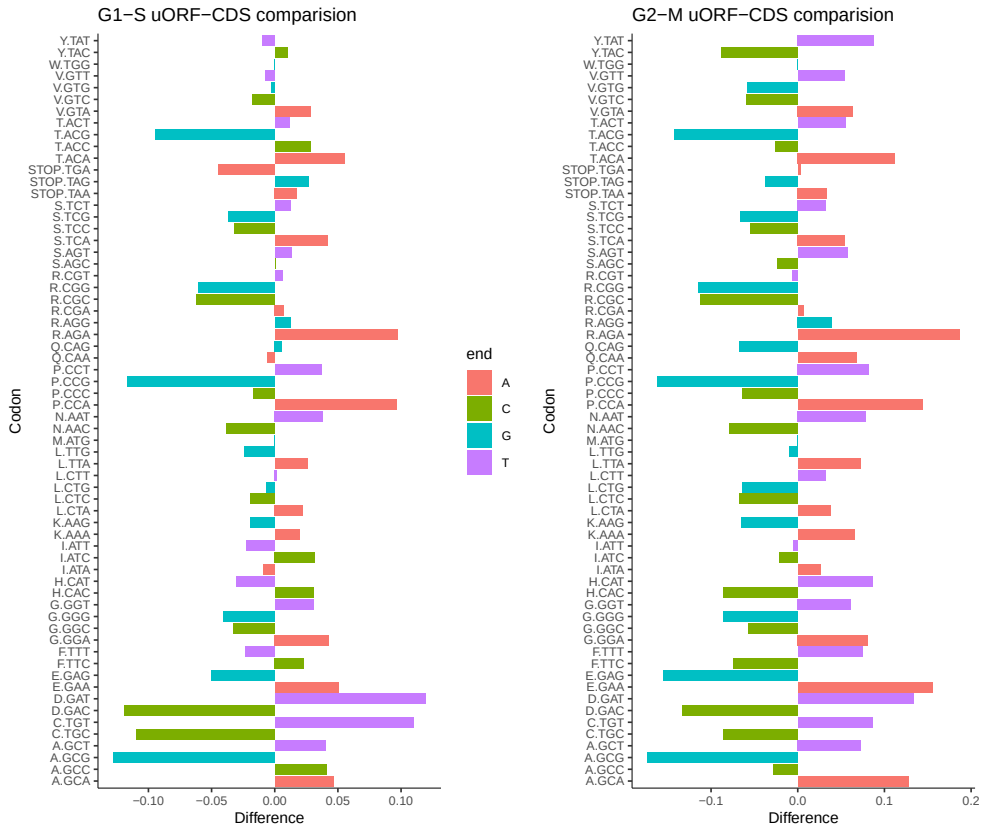


Figure S6: **Codon usage differences between uORFs and CDSs.** The codon frequency differences were determined by calculating the differences in codon occupancy for each encoded amino acid across all transcripts. A positive value indicates that the codon is favoured in the main ORF, while a negative value suggests it is preferred in the uORF. The colour coding represents the codon end, as illustrated in the figure legend. The left panel compares the G1-S gene while the right panel compares the G2-M