Comparative genomics of translational regulation

by

Saket Choudhary

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTATIONAL BIOLOGY AND BIOINFORMATICS)

August 2020

Dedication

> One important idea is that science is a means whereby learning is achieved, not by mere theoretical speculation on the one hand, nor by the undirected accumulation of practical facts on the other, but rather by a motivated iteration between theory and practice.
>
> — George Box

To my family and teachers

Acknowledgements

My journey through graduate school would not have been possible without the support of some wonderful people around me. I feel I have been privileged in multiple aspects, but of everything I have been specially privileged to have some really smart and caring people around me.

I would first like to thank my advisor Dr. Andrew Smith for agreeing to serve as my advisor and giving me a lot of freedom to pursue my ideas. His questions have helped me shape up as a better scientist and I have picked up some really great ideas from him in the process. I would also like to thank my committee members, Dr. Remo Rohs, Dr. Mark Chaisson, Dr. Peter Calabrese, Dr. Zhiping Lu and Dr. Adam MacLean. Remo and Mark, in particular, were great mentors who helped me with their advice everytime I was stuck. I felt privileged reaching out to them whenever I felt like and I will forever be grateful to them. I would also like to specially thank Dr. Peter Calabrese for the wonderful time I had assisting him in teaching BISC-305 - I learned a lot about how to teach students engagingly and had lot of fun doing so.

I was very fortunate to work with two great collaborators - Dr. David Kadosh and Dr. Luiz Penalva at UT Health San Antonio. A small project with Dr. Luiz Penalva played a fundamental role in shaping the rest of my PhD so I will forever be grateful to him. Working with Dr. David Kadosh has been one of the most enlightening parts of my PhD - I came across a number of

interesting biological problems that required development of novel computational methods that eventually formed the core part of my PhD. I will also be grateful to David for the time he spent advising me on life after PhD and specially during my postdoc search.

I was also fortunate to have a great set of lab members and colleagues. I would especially like to thank Dr. Benjamin Decato and Amal Thomas for being both a critical lab member and a great friend. Our discussions about Science and life otherwise have been really insightful. My colleague, lab mate, and friend Dr. Wenzheng Li played a great role in shaping up my research. Our way of challenging each other over Science and our discussions lead to a very productive research life. My special thanks to Dr. Meng Zhou for being a great office mate. I am also grateful to my lab member and friend Dr. Rishvanth Prabhakar for his critical reviews and for fun discussions over the numerous Friday night dinners. I am grateful to Guilherme and Liz for their critical reviews on my writings.

I have been fortunate to have an extended family here in LA. Some relationships go beyond friendship and the only way to do justice to them is to leave them unnamed. Satish, Swati, and Hiteshi have been an integral part of my journey from the beginning. I cannot imagine myself living so far away from family had it not been the support of these three particularly and I will run out of words if I start thanking them in words - my heart fills with gratitude whenever I think about them. I also had the fortune to make some really great friends in these six years. Satya and Puja's company was always cheerful. Sheetal and Anirban were great friends, neighbors and my LA-food-exploring buddies. Nilkanth, my room-mate for being a constant source of fun, a great friend and for all the help he has continued to extend me over all these years In the later years of my journey, my friendship with Angana gave me a new light of confidence in myself and others. She taught me many a things about life, about people and became my perpetual coffee, then tea and finally hot-chocolate buddy. I keep wishing we meet again. Deepak and Soniya, Archith and Aarthy, Nilkanth and Payal, Debanjan, Bibek, Mukul, Gaurav, Karthik, Samprita, Karishma and Rimita for a beautiful friendship and listening to my rants and for having me over for numerous lovely dinners. Asif for being a wonderful-wonderful buddy who always listened to my constant blabbering, for inspiring me to do good in Science and in life. Vasantika for watching my back for me and extending a helping hand whenever I needed it. Nachikethas for being my MATH-buddy, a great friend and for teaching me how to be focused! Shweta Ramdas for constantly checking on me and for reviewing my writings even before we met each other physically for the first time. Madhura Mukhopadhyay for her words of wisdom during my postdoc search. Shailja for being a wonderful friend and for celebrating my little successes with me.

Friends from my undergrad - Akshay Korde, Prateek Maheshwari, Pushkar Godbole and Ashlesh Sonje for constantly checking on me and for constatntly planning awesome trips some of which did see the light of day. Shekhar Mishra and Anirudh Subramanyam for inspiring me to be sane when the times were tough and for their book recommendations some of which changed part of my life (*Siddharatha*). My close friends Seshadri and Shuchi for a wonderful friendship and for making me witness my first wedding in the US! Yamini Krishnan and Parijat Bandyopadhyay for the wonderful conversations on Science and beyond and for going through my writings multiple times. My alma mater, IIT Bombay for introducing me to a wonderful pool of both opportunities and friends!

My cousins Sanju and Chintu bhaiyya, Puja and Neha bhabhi and my wonderful nephews and nieces for their unconditional love and support! Suddu mama and Jaya mami for always keeping me motivated. My aunt, Mitthu didi for all the love she has showered over all of us over the years and for always looking after me, right from my birth. My extended family: Jay and Thelma auntie for showering me with lots of love and keeping me motivated during tough and not so tough times. Jay's insights have made me re-realize some simple but important things in life - 'A no response is also a response' and how to believe in the power of intuition, something that was particularly useful during my postdoc search. Conversations with auntie have made me realize the importance of *realizing* the value of things in life, something I hope will stick with me forever. I am here today because of some really wonderful people, my pillars - my parents. No words that I write here will ever do justice they mean to me and for what they have done for me. My Papa instilled in me the spirit of hard work, patience and perseverance and Maa has taught me the art of keeping calm. Their teaching helped me through graduate school and I believe they will keep doing so. My sister Khushi, who literally lives by her name and has been a source of happiness for me and everyone in her family. She was my only source of inspiration while growing up and continues to inspire me to *think*, about things, about people, how to really care for someone and in whose shadows I will always walk.

Table of Contents

List Of Tables

List Of Figures

Abstract

Abstract

Any biological process is a manifestation of the protein levels. Protein are synthesized inside the body through the process of translation. Translation is one of the most energy-intensive processes and hence is highly regulated to ensure the proteins get synthesized in the right quantity in the cell. The advent of next-generation sequencing has enabled transcriptome-wide monitoring of translation regulation. In particular, Ribo-seq has provided us with a high-throughput sequencing method to profile mRNA undergoing active translation at sub-codon resolution. However, the resulting dataset has inherent noise arising from the sequencing of non-ribosomal and non-active translation fragments. Through this work, I develop a method to identify regions under active translation using Ribo-seq data. The method enables more accurate identification of both short and long open reading frames. I develop a database of uniformly processed public Ribo-seq datasets which have been appropriately de-noised to recover only the actively translating fragments. I also demonstrate the utility of our method as a quality control metric to asses the quality of Ribo-seq data generated in a variety of treatment and biological contexts.

I utilized our method to understand the changes involved in translational landscape in two important biological contexts. First, I present a study on translational landscape changes involved in the irradiation of glioblastoma cell lines thereby providing evidence of radio-resistance in some target genes that can be likely candidates for increasing irradiation efficacy. Second, I provide evidence of translational landscape changes involved in the morphological transition of *Candida albicans*, a fungal pathogen involved in inducing diseases as Candidasis in immunocompromised patients. These target genes can serve as potential targets for anti-fungal drugs.

Finally, by applying our method to identify actively-translating regions on public Ribo-seq datasets, I discovered thousands of upstream open reading frames under active translation across various biological contexts in multiple species. These upstream open reading frames (uORFs) often act as a repressive element for the downstream protein coding regions. Using Ribo-seq datasets from multiple species across different conditions, I characterize the sequence properties of a subset of these uORFs which are almost always under active translation. Further, I analyze it in an evolutionary context providing evidence for the conservation of the upstream sequence context.

Chapter 1

Introduction

The basic structural, functional and biological unit of all organisms is the cell. Cell, the smallest unit of life, is enclosed within a membrane and contains two key biomolecules : DNA and proteins. Since their discovery in 1869 and 1838 respectively, these have become widely-studied and amongst the best-understood biological molecules.

DNA has been known to biologists since 1869 when the young Swiss doctor Friedrich Miescher, working in the laboratory of Felix Hoppe-Seyler at the University of T˝obingen, while performing experiments to determine the chemical composition of leukocytes isolated a precipitate that contained large amounts of phosphorus, lacked sulfur, and was resistant to protease digestion thus essentially ruling out proteins and lipids [5].

By itself, DNA is little more than a blueprint for life. Any biological process is a manifestation of the proteins and their abundance. Proteins were described in scientific literature as early as 1838 by the Dutch chemist Gerardus Johannes Mulder though the name itself was a creation of Swedish chemist J˝ons Jacob Berzeliuswith [6]. The name protein originates from the Greek word $\pi\rho\acute{\omega}\tau\epsilon\iota\varsigma$ (proteios) - primary or leading in the front [7]. As such, proteins act as building blocks for a large variety of large molecules and account for 44% of the human body's dry weight [8]. They also also act as enzymes that catalyze the biochemical reactions in all living beings.

Francis Crick gave a lecture on 19^{th} September 1957 as part of the Society of Experimental Biology Symposium on the Biological Replication of Macromolecules at University College London. This lecture entitled "Protein Synthesis" that was later written up as "On Protein Synthesis" [9] was significant from multiple fronts. First, it laid down the foundation of what is now famously known as the central dogma of molecular biology. The central dogma states that once 'information' has passed into the protein, it cannot be transferred back to either nucleic acid or protein. In other words, information can flow from nucleic acid to nucleic acid or from nucleic acid to protein but transfer of information from protein to protein or protein to nucleic acid is impossible. Second, Crick made a couple of predictions that were eventually proved to be true. First, he predicted the existence of adaptor molecules (later discovered in the same year as tRNA [10]) that work as mediators carrying amino acids to the site of protein synthesis And hence, Crick "permanently altered the logic of biology [11]". Second, he predicted that it would be possible in the future to compare sequencing data to explore the vast amount of evolutionary information between them.

The single largest investment of energy by cells is in the process of translation of mRNAs into protein [12, 13]. Being a substantial energy investment, translation is highly regulated to ensure that the resulting proteins have the desired stoichiometry and are available at the right places within the cell. For example, in a rapidly growing yeast cell, nearly $200,000$ ribosomes occupy as much as $30-40\%$ of the cytoplasm and 2000 ribosomes are synthesized every minute absorbing around 60% of its transcriptional activity [14]. Translation regulation is especially important in maintaining homeostasis and controlling cell proliferation and growth [15]. Dysregulation of translation has been implicated in a wide range of diseases [16].

The notion of translation regulation emerged only a few years after Crick formulated the central dogma of molecular biology. In 1961, Jacob and Monod proposed the idea of 'messenger' intermediate (mRNA) could be subject to different utilization depending on the context [17]. They perceived "that the synthesis of individual proteins may be provoked or suppressed within a cell, under the influence of specific external agents, and more generally that the relative rates at which different proteins are synthesized may be profoundly altered, depending on external conditions" [17] and that such a regulation is "is absolutely essential to the survival of the cell" [17]. This idea of mRNA being utilized differently to regulate the final protein drew very little attention at that time but has been one of the seminal ideas in molecular biology.

The advent of next-generation DNA sequencing technologies and the development of assays such as RNA-seq [18] has facilitated relative ease of measuring mRNA abundance in a high-throughput manner, thus, advancing our knowledge of how cells modulate their gene expression across different physiological and pathological processes. However, biological processes are driven by proteins, while mRNA concentration acts as a mere proxy to protein abundance. At a steady state, protein abundance is a function of the rates of four phenomena: transcription, translation, mRNA decay, and protein decay [19, 20]. Measuring mRNA abundance has been found to be a good predictor of protein levels in multiple contexts [21].

Measuring mRNA abundance though highly informative provides a glimpse into the early steps of a long chain of regulatory events. The correlation between mRNA and protein levels in some contexts has been found to be as low as 0.36 [22]. Such low correlations have been attributed to translational regulation and post-translational buffering. Deciphering the relative contribution of phenomena other than transcription can help us deepen our understanding of biological regulation.

The need to decipher translational regulation has motivated the development of experimental approaches to profile the translational landscape. With the advent of next-generation sequencing technologies, it is now possible to determine the sequence of large DNA and RNA molecules in a high-throughput manner [23]. Ribo-seq [24] is a deep-sequencing based technique that captures snapshots of ribosome protected fragments revealing the positions of the entire pool of ribosomes engaged in translation, hence, providing a global view of the translational process *in-vivo*. Ribo-seq has been used to answer some key biological questions involving the prevalent and dynamic nature of translational regulation of the mammalian cell cycle [25], the discovery of alternative translation initiation sites [26, 27], and translational pauses under induced stress [28].

Ribo-seq provides a rich resource for understanding translation and its regulation. However, ever since the inception of Ribo-seq, it has been known that not all fragments arising from a Ribo-seq experiment are reflective of active translation [29, 30, 31]. Furthermore, Ribo-seq data is highly heterogeneous and noisy making the task of separating "signal" from noise particularly challenging. Multiple approaches have been taken to tackle this shortcoming [32, 33, 34, 35, 36]. These methods can be broadly classified into two paradigms. The first paradigm of methods hypothesizes that the distribution of Ribo-seq fragments is different from a hypothetical null and use this to separate the ones actively-translation from the non-active ones. The second paradigm of methods uses the periodic pattern expected from Ribo-seq signal, since the ribosome moves three nucleotides at a time, to determine actively-translating fragments.

The number of Ribo-seq studies has grown over the years. These studies span several tissues and species. They have been performed under different biologically or technically diverse conditions. However, every new Ribo-seq studywill overlap with previous studies with respect to some biological dimensions of interest. Hence, interpreting new data can leverage on existing public data to accelerate the process of deciphering all biological characteristics for the new experiment. Systematically leveraging on public Ribo-seq data can help us answer some key biological questions involving translation regulation. A key mechanism of regulating translation during the initiating step is via upstream open reading frames (uORFs) located in the 5' untranslated regions (5' UTRs). uORFs are sequences located upstream of the main protein coding region characterized by an initiation codon in-frame with the stop codon. uORFs regulate the translation of downstream protein coding region by multiple mechanisms: by inhibiting initiation at the downstream start codon (AUG) because of their secondary structure; by stimulating cap-independent translation through the internal ribosome entry sites (IRES); by inhibiting or promoting translation because of presence of potential RNA binding protein sites and by the upstream sequence and start codon context [37].

The literature on uORFs thus far has primarily relied on simply using the presence of a 'favorable' sequence context in the 5' UTRs. This favorable sequence context is most often characterized by the presence of a start codon (AUG) and an in-frame stop codon (UAG/UGA/UGA). Ribo-seq, on the other hand, can provide direct evidence of active translation in uORFs as compared to a mere potential arising because of a favorable sequence context. By providing a snapshot of active translation, Ribo-seq has enabled investigating the true coding potential of the uORFs. Given that uORFs mediate translation regulation and it has functional implication, we expect this mechanism to be similar across species as "Nothing in Biology Makes Sense Except in the Light of Evolution [38]".

1.1 Contributions

1. **A method for detecting active translation in both short and long ORFs:** Though tools exist for identifying active translation in Ribo-seq data, they are not sensitive enough to detect active translation in short ORFs (< 100 amino acids). I developed a method, ribotricer, that exploits the inherent periodicity in Ribo-seq data to identify active translation. It gives the highest accuracy and sensitivity compared to other existing approaches as evaluated on multiple datasets across species .

2. **Characterization of translational landscape changes in *C. albicans***: *C. albicans* is a fungal pathogen that inhabits the mucosal surfaces of most healthy individuals as human commensals. Though mostly asymptomatic, they are opportunistic pathogens and are known to causes disease in individuals with a debilitated immune system or a disruption in the host's microbiome. Using ribotricer, I provide insights into changes that take place in its translational landscape and are responsible for its increased virulence of this pathogen.

3. **A uniformly processed database of Ribo-seq studies:** I develop a database, ribopod, of uniformly processed public Ribo-seq datasets across multiple species. Leveraging the accuracy of ribotricer, I provide access to *de-noised* Ribo-seq datasets that will serve as a rich resource for the translation regulation research community.

4. **Conservation of uORF mediated regulation**: uORFs are known to play a role in suppression of translation. Using ribotricer and ribopod, I provide evidence of conservation of uORF mediated regulation across species. I describe universal uORFs (uuORFs) that are almost always translating across different physiological and pathological contexts within a species. Furthermore, I show that the sequence context of the uuORFs is conserved across species.

1.2 Outline

In Chapter 2, I provide a background for translational regulation and Ribo-seq. I provide a history of translation regulation, the known mechanisms that regulate translation. Finally, I describe Ribo-seq and highlight some of the challenges associated with the analysis of Ribo-seq data.

Chapter 3 introduces a new computational method, ribotricer, to identify actively-translating regions in Ribo-seq data. I first provide a review of the existing computational approaches and their limitations. Next, I introduce our method, ribotricer, that overcomes the shortcomings of previous approaches. I provide detailed validation of our approach on multiple datasets across species.

In Chapter 4, I describe changes in the transcriptional and translational landscape of *C. albicans*, a fungal pathogen, that rapidly transitions between yeast-like and filamentous growth patterns where the latter pattern leads to increased virulence. I utilize ribotricer to discover hundreds of genes with altered translation during the morphological translation. Using RNA-seq and Ribo-seq data jointly I re-annotated the partially incomplete transcriptome of *C. albicans* and discovered *novel* exons.

In Chapter 5, with the aim of developing a better targeted therapy for glioblastoma, I characterize the effect of radiation on glioblastoma. Using RNA-seq and Ribo-seq data, we conducted an integrated analysis in the glioblastoma cell lines to profile alterations in the gene expression at multiple levels including transcription, splicing, and translation. Our approach provides a comprehensive view of early response to radiation in and suggests new target options to increase radiation sensitivity and prevent relapse of glioblastoma.

In Chapter 6, I provide a background of existing Ribo-seq datasets and highlight their shortcomings. Next, I try to overcome these shortcomings by developing a database, ribopod, of uniformly processed public Ribo-seq projects from multiple species. Ribopod provides ready access to de-noised Ribo-seq projects that can be used to discover new biological mechanisms or as a resource to test a hypothesis in different biological and physiological contexts.

In Chapter 7, I use ribopod database to investigate uORF mediated regulation across species. I describe uuORFs, a set of uORFs that are almost always expressed across physiological and pathological conditions withing a species. I discover that the sequence context associated with these uORFs is different from the context associated with protein coding regions yet is conservation across species.

Chapter 2

Background: Translational regulation and Ribo-seq

Protein synthesis is a quintessential part of the gene expression pathway and is itself involved in its control. The rate of synthesis of protein depends on both the concentration of the mRNA and its translational efficiency [39]. Translational efficiency is determined by the density of active ribosomes translating the mRNA. Half-life, i.e., the time required for degradation of 50% mRNA molecules for a majority of mammalian mRNAs is $>$ 2 hours [40], and hence tight protein regulation requires is done via translational efficiency and protein decay rates [41].

2.1 Translation

The process of translation is one of the most energy-intensive processes in any cell. The energy requirements are high owing to the multiple GTP hydrolysis steps involved [42]. The process is precise with the frequency of translations errors reported to be as low as 0.1% [43, 44] in incorporating an erroneous amino acid. Aberrations in the translation pathway are known to cause various diseases [45]. Complete translation of a protein involves three steps of initiation, elongation, and termination that we briefly discuss in the sections that follow.

2.1.1 Translation Initiation

Translation initiation begins with recognizing the 5' -m^{7}G cap by the protein initiation complex. This protein initiation complex consists of the 40S (smaller) subunit, translation initiation factors, and tRNA conjugated to methionine and attached to the peptidyl transfer (P-) site of the 40S subunit. The eukaryotic initiation factor 4F (eIF4F), which consists of factors eIF4E, eIF4A, and eIF4G bind to the m^{7}GpppN cap structure located at the 5' end of pol-II transcribed mRNAs. The 40S subunit carries the tRNA-eIF2-GTP complex that gets attached at or near the 5' end of the mRNA. The 40-S associated factor eIF3 and eIF4G aid this linking [46].

The pre-initiation complex scans for the start codon 'AUG' in the 5' to 3' direction along the mRNA. Once an AUG is encountered, the methionine tRNA canonically base pairs with the AUG start codon thus arresting scanning and releasing initiation-factors. It makes way for the more prominent subunit 60S to bind to the 40S at AUG resulting in the 80S ribosome assembly. The presence of m^{7}G cap is quintessential for translation [47], though alternate mechanisms such as the presence of internal ribosomal entry sites (IRES) in the 5' untranslated region (UTR) can also recruit the translation initiation machinery independent of m^{7}G presence [48]. Initiation is the rate-limiting step of the three steps in protein synthesis [49].

2.1.2 Translation Elongation

The 80S assembly traverses the mRNA codon by codon. There are three sites inside the 80S monosome: A, P and E. The A (acceptor) site acts as a binding site for the aminoacyl tRNA, the P (peptidyl) site binds to the tRNA that holds the growing nascent polypeptide and the E (exit) site serves as a threshold for releasing the tRNA devoid of its amino acid. At each codon,

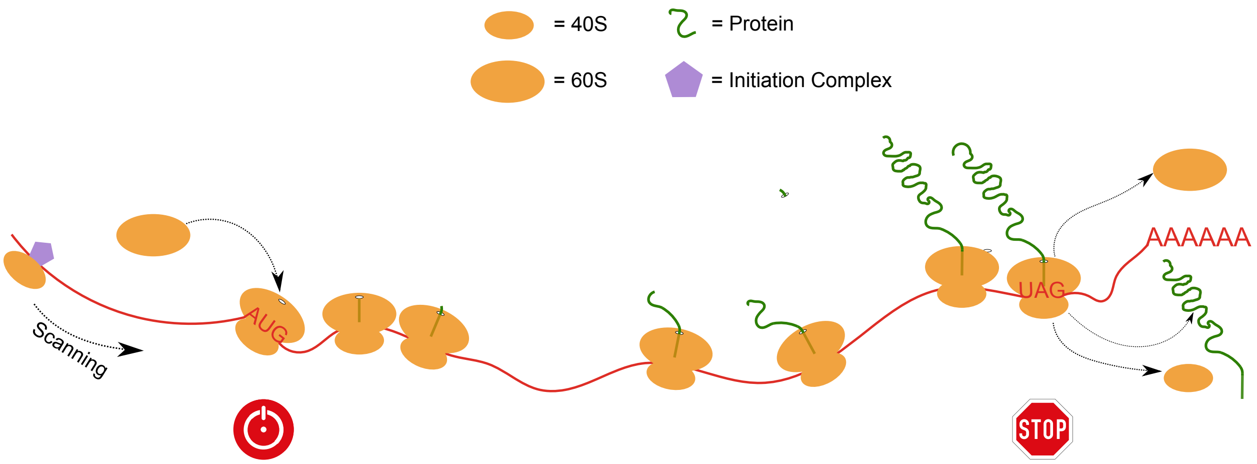


Figure 2.1
Canonical model of translation: initiation, elongation and termination. 40S subunit along with the pre-initiation complex scans for 'AUG'. On encountering AUG, the 60S subunit binds to 40S forming the 80S complex while the pre-initiation complex is let free. The 80S complex traverses the mRNA one codon at a time growing the peptide chain (green) till it encounters a stop codon. The two subunits are released and recycled.

the corresponding aminoacyl-tRNA is base pairs with the codon at the A-site. It results in the formation of a peptide bond and the growing peptide chain (Figure 2.1) is translocated to the P-site. The tRNA-peptide is still base-paired to the mRNA at the P-site while the initiator tRNA now occupies the E-site. It leaves the A-site empty for the next codon to recognize its cognate aminoacyl-tRNA. The elongation process proceeds one codon at a time until it encounters an in-frame stop codon.

2.1.3 Translation Termination

At the stop codon, the linkage between elongated peptide chain and the P-site tRNA site undergoes hydrolysis thus releasing the nascent protein, and the two subunits of the ribosome disassociate. The subunits are recycled and re-used for the translation of other mRNA fragments.

2.1.4 Translation in eukaryotes versus prokaryotes

In prokaryotes, ribosomes engage in co-transcriptional regulation. As soon as the mRNA emerges from the RNA polymerase, ribosomes can recognize the start codon on these mRNA and start translating. On the other hand, the eukaryotic translation starts only after some quality control checks. The mRNA in eukaryotes gets processed in the nucleus before being exported to the cytoplasm where the ribosomes can translate it. The mRNA goes through capping and polyadenylation, among other processing steps in the nucleus. These processing steps ensure that the ribosomes engage with only mRNAs that are complete [50].

2.2 Factors affecting translation

There are five main contexts that can modulate translation in eukaryotes and more specifically in vertebrates [51]: 1. m^{7}G cap 2. the sequence context surrounding the AUG start codon 3. the relative position of AUG start codon in the transcript ("first" versus others) 4. upstream and downstream secondary structure 5. leader length We briefly discuss the effects of each of these contexts in the sections that follow.

2.2.1 m^{7}G cap

The m^{7}G cap stabilizes the mRNA by preventing it from degrading by the attack of phosphatases and other nucleases. The cap generally gets added to mRNA precursors synthesized by RNA polymerase II and to viral transcripts that are replicated in the nucleus. It increases translational efficiency [52] and is quintessential for translation except in viral mRNAs.

2.2.2 Sequence context

Marilyn Kozak characterized **GCC(A/G)CCAUGG** as the optimal sequence context for translation initiation [47] at the AUG start codon in vertebrates. This context affects both fidelity and the efficiency of initiation. AUGs around sub-optimal contexts can cause some 40S subunits to bypass it and initiate translation at a downstream AUG. It is possible that the first AUG if lying in a sub-optimal context, is recognized inefficiently which is consistent with the leaky scanning model. The leaky scanning model proposes that the 40S subunit continues scanning the mRNA for AUGs even after encountering an AUG. This leaky scanning model enables some viral mRNAs to produce two proteins by initiating at two AUGs lying close to each other [53]. Based on the two most crucial nucleotides -3 (A/G) and +4G, the surrounding sequence context has been classified as 'optimal', GCC(A/G)CCAUGG; 'strong' NNN(A/G)CCAUGG when the two important nucleotides are present; 'adequate', when either of the two important nucleotides are present: NNNRNNAUG(A/C/U) or NNN(C/U)NNAUGG, and 'weak' NNN(C/U)NNAUG(A/C/U) that lacks both of the two important nucleotides. This context was assumed to be universal amongst eukaryotes. Later, however, it was eastablished that sequence context might be species dependent. Amongst the characterized species, the sequence context is ACAACCAAAAUGGC for Drosophila, AAAAAAAAAAUGTC for Saccharomyces cerevisiae, and UAAAT(A/C)AACAUG(A/G)C for other invertebrates [54].

2.2.3 The relative position of AUG

The scanning model of the ribosome hypothesizes that the ribosomes will initiate translation at the first encountered AUG as long as it lies in a 'optimal' context [55, 56]. An experiment

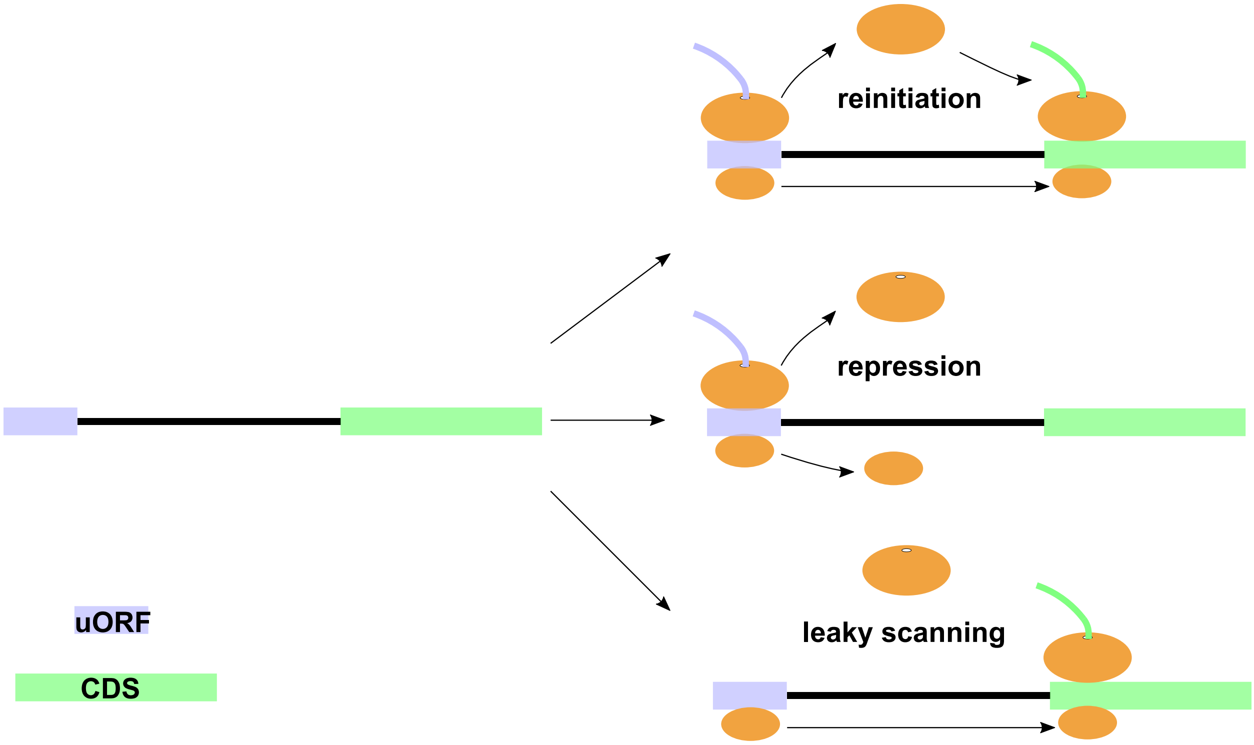


Figure 2.2
uORFs and their potential effects on the downstream CDS. uORF translation can result in short peptides while the free sub-units can re-initiate translation at the downstream CDS. Presence of uORFs can repress the production from downstream CDS. Leaky scanning by 40S subunit can skip uORFs and initiate translation at the downstream CDS.

involving the insertion of out-of-frame AUG codons showed dramatic inhibition of translation [57]. Thus, efficient translation requires that no spurious AUG codons be present upstream. However, the presence of upstream AUGs in the 5' UTR regions may not lead to complete inhibition of translation if it is followed by a stop codon. The 80S ribosome may be released after the translation of this small region, but the 40S will remain bound and continue to scan the mRNA. It may then re-initiate translation at a downstream start codon (the original CDS), however, the re-initiation will happen efficiently only if the upstream ORF is short (Figure 2.2).

2.2.4 Secondary Structure

Secondary structure can have both positive and negative effects on translation initiation. A small amount of secondary structure near the AUG can up-regulate its recognition by 40S. The secondary structure downstream tends to slow down the scanning by the 40S thus giving it more time to recognize the AUG. On the other hand, the presence of stem-and-loop secondary structure between the AUG codon and the 5' cap can inhibit translation. This stem-and-loop prevents the 40S subunit from binding if it inhibits the entry sites for ribosomes. If the secondary structure occurs far away from the cap such that binding of the 40S subunit is not impaired, the inhibition of translation depends on the stability of the stem-and-loop as a strong stable structure will inhibit 40S' movement.

2.2.5 Leader length

If the AUG is located very close to the 5'cap, the recognition might be impaired. In an experiment with synthetic transcripts where the AUG was located in a favorable Kozak context too close to the cap, the initiation efficiency was significantly lower as compared to when the leader sequences were elongated independent of the sequence context. This can be attributed to the fact that longer 5' leaders might provide more avenues for loading 40S as compared to shorter sequences.

2.3 Quantitative analysis of *in-vivo* translation by Ribo-seq

Ingolia *et al.* [24] developed a high throughput sequencing-based approach for ribosome profiling called Ribo-seq. It has enabled a transcriptome-wide quantitative analysis of *in-vivo* translation. The assay involves deep-sequencing of mRNA fragments protected by ribosomes which are believed to be undergoing active translation. It gives a codon level resolution for the dwell time of ribosomes [58]. The abundance of Ribo-seq fragments has been shown to be more correlated with the absolute protein abundances [24].

Ribosomes protect around $28 - 30$ nucleotides of the mRNA from digestion [24, 59] while tightly packed ribosome pairs can protect around $58 - 62$ nucleotides [60]. On the other hand, 48S pre-initiation complex can protect $50 - 70$ nucleotides [61] because of the presence of eIFs.

2.3.0.1 Comparison to polysome profiling

Polysome profiling measures the constituent mRNA bound by different ribosome number fractions using microarrays [62]. Ribo-seq has some clear advantages over the microarray based approach. Ribo-seq can distinguish between ribosomes engaged at the protein-coding regions from those engaged in the upstream regulatory open reading frames (uORFs) [31]. mRNAs that undergo rapid degradation might not be profiled by polysome profiling [62]. Ribo-seq, on the other hand, relies only on nuclease footprint from single ribosomes and as such has lesser sensitivity to the integrity of mRNA. At the same point, there are certain key properties of polysome profiling that Ribo-seq is not able to capture. For instance, polysome profiling has a greater ability to measure differences in the translation of alternate isoforms, particularly so when they differ in their ribosomal occupancy in the 5' or 3' untranslated regions. Also, polysome profiling monitors the translation status of entire full length transcripts while Ribo-seq focuses on determining activity of individual ribosomes. As such, using polysome profiling it is possible to capture the difference arising from a uniform decrease in ribosomal occupancy on all copies of a transcript from the

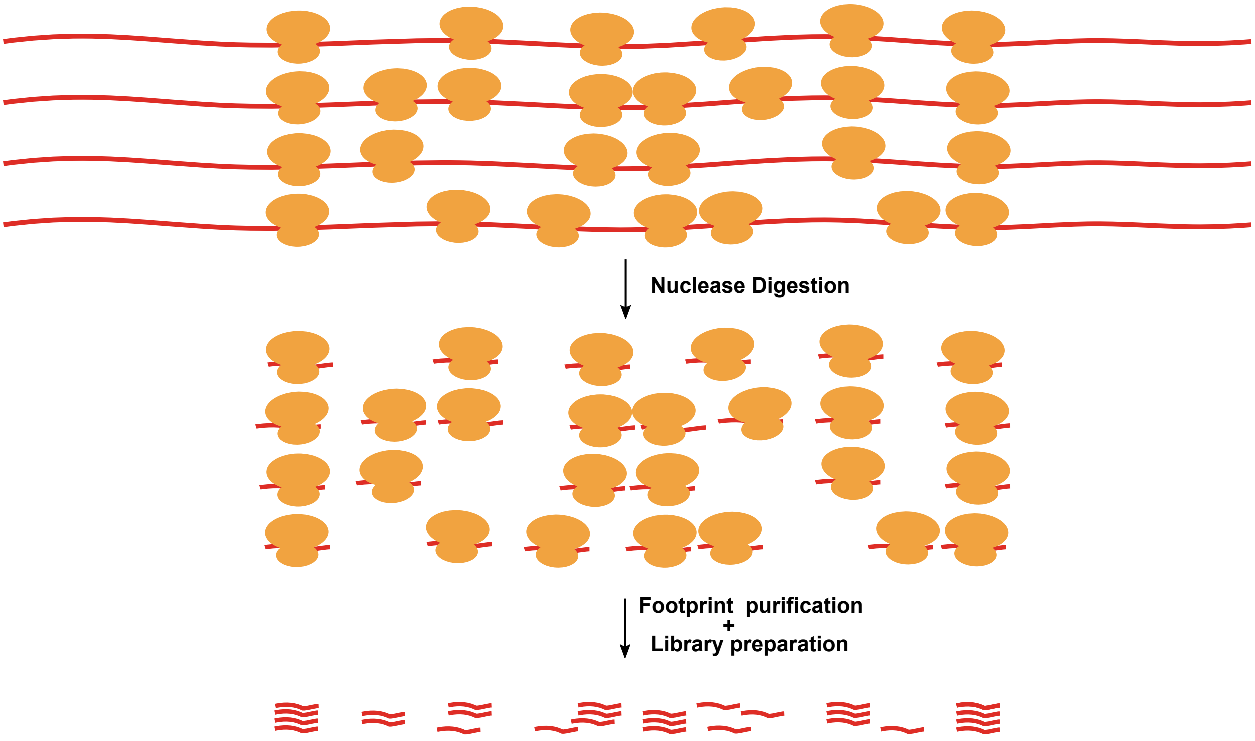


Figure 2.3
Protocol of Ribo-seq. A translating ribosome encloses 30 nucleotides of the mRNA and protects it from digestion. mRNA protected fragments are obtained by nuclease digestion which digests the unprotected mRNA. The protected fragments are then sequenced post ribosome separation.

scenario where a sub-population of mRNAs is under complete repression, a scenario that has been observed in mouse embryonic stem cells [63].

2.3.1 Ribo-seq protocol

Ribo-seq protocol consists of five major steps: 1. Cell lysis and ribosome arrest 2. Nuclease footprinting (RNAse digestion) 3. Isolation of ribosome footprints 4. Library preparation and high-throughput sequencing

Cell lysis and ribosome arrest: In order to profile the translatome, the polysomes need to be first stabilized. This has traditionally been carried out by treating the cells with translation elongation inhibitors before lysing them. One of such inhibitors, cycloheximide has been shown to be biased towards certain codons, altering the distribution of ribosomes on the mRNA, especially near the start codons [64, 65, 66, 67]. Based on these observations, flash-freezing of cells appears to be the most robust approach [68].

Nuclease footprinting (RNAse digestion): This step involves digesting the portion of mRNA that is not protected by ribosomes. RNAse I in eukaryotes and micrococcal nuclease (MNAse) has been used in bacterial cells to perform digestion step [69].

Isolation of Ribosome footprints: After RNA digestion, the ribosome protected fragments that are intact need to separated from the cell lysates. This was originally performed using a sucrose density gradient centrifugation but is now performed by passing the lysate through a sucrose cushion during ultracentrifugation to ensure purification ribosome-bound mRNA.

Library preparation: The library preparation step encompasses rRNA deletion and linker ligation. Each ribosome-footprint complex has several kilo-bases of rRNA while the mRNA protected fragments are only 28 bases

2.3.2 Using Ribo-seq to decipher translation regulation

Ribo-seq data can be leveraged to answer the "how much-where-how" questions involving translation regulation. The simplest of these questions,"how much", is answered by the profiled mRNA fragments serving as a proxy to the protein levels. On the other hand, proximity-specific ribosome profiling has enabled deciphering the "where" questions. For example, the majority of mitochondrial inner membrane proteins are co-translationally translocated except the proteins that are targeted to other mitochondrial sites [70]. Not all aspects of translational control cannot be emulated *in-vitro*. By enabling *in-vivo* measurements, Ribo-seq facilitates identification of

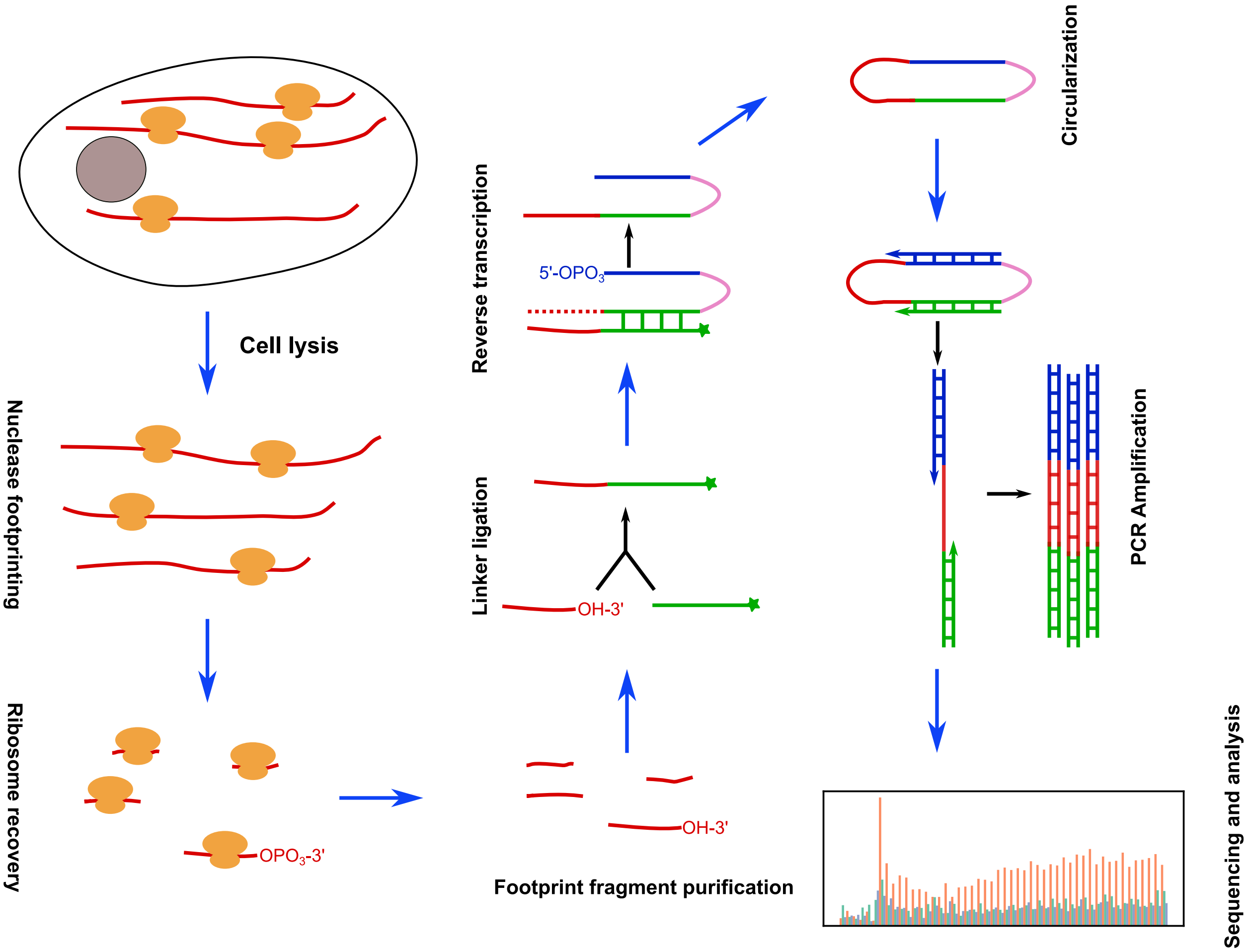


Figure 2.4

Ribo-seq protocol The protocol encompasses four major steps. In the first step, the cell is lysed and harvested under conditions to ensure *in-vivo* positions of the ribosomes are unaffected. The cell lysate can then be digested using a nuclease which will digest all portions of the mRNA not being protected by the ribosome. The portion being protected by the ribosome called the ribosome footprints can then be purified and ligated to a single-stranded linker molecule that serves as a priming site for reverse transcription. The products of first-strand reverse transcription can be circularized to provide a second priming site flanking the original captured footprint sequence which is then used for PCR amplification of a deep-sequencing library.

translation mechanism that vary across cell states and organisms thus answering the “how” questions. For example, Ribo-seq has been used to understand the role of Dom34 in rescuing stalled ribosomes [71].

If multiple isoforms are co-expressed, translation initiated at the upstream initiation sites can obscure the strength of initiation signal downstream. However, treatment with certain drugs such as harringtonine [72], or lactimidomycin [26] which is capable of immobilizing the initiating ribosomes will result in an overabundance poof ribosome protected fragments (RPFs) at the initiation sites. Translation initiation inhibitors have been used to discover alternate translation initiation sites [73].

The number of footprints originating from a transcript directly correspond to the number of ribosomes engaged in translation. This is also equivalent to a) the amount of protein being produced and b) the time required to synthesize it. The time to translate an an ORF is proportional to its length. Ribo-seq has been used to provide empirical evidence that the speed dynamics of translation are consistent across different group of genes [74].

The number of ribosomes over a gene indicates how many ribosomes are translating it. Similarly, the number of footprints over a particular codon indicates how often do ribosomes hit that particular spot. If ribosomes pause at a particular location while translating a gene, then ribosomes will spend a longer duration of time at these codons which would lead to an over-abundance of footprints at this location. Ribo-seq data has been used to discover ribosomal pausing sites in bacteria [69], yeast [75] and mammals [76].

In the upcoming chapters we rely on Ribo-seq data to answer some key questions on mechanisms of translational regulation. In particular, we first present a new computational method to identify short regions which are actively engaged by the ribosomes and apply this method to discover prevalent translation in the upstream leaders across species.

Chapter 3

Accurate detection of short and long ORFs from Ribo-seq data

3.1 Introduction

The process of translating messenger RNA into protein is among the greatest investments of energy by cells [12]. Consequently, translation is highly regulated to ensure that each cell synthesizes the right amount of each protein. Our understanding of the mechanisms regulating the translational process remains limited, which has motivated the development of experimental approaches to profile the translation landscape globally. Ribo-seq [24] is a technology that uses deep-sequencing to identify ribosome-protected fragments, revealing the positions of the entire pool of ribosomes engaged in translation.

Ribo-seq has led to the surprising discovery of prevalent translation through non-canonical ORFs [77]. The non-canonical ORFs include the upstream ORFs (uORFs) located in the 5' untranslated region (UTR), the downstream ORFs (dORFs) located in the 3' UTR, and the ORFs within presumed non-coding genes [78].

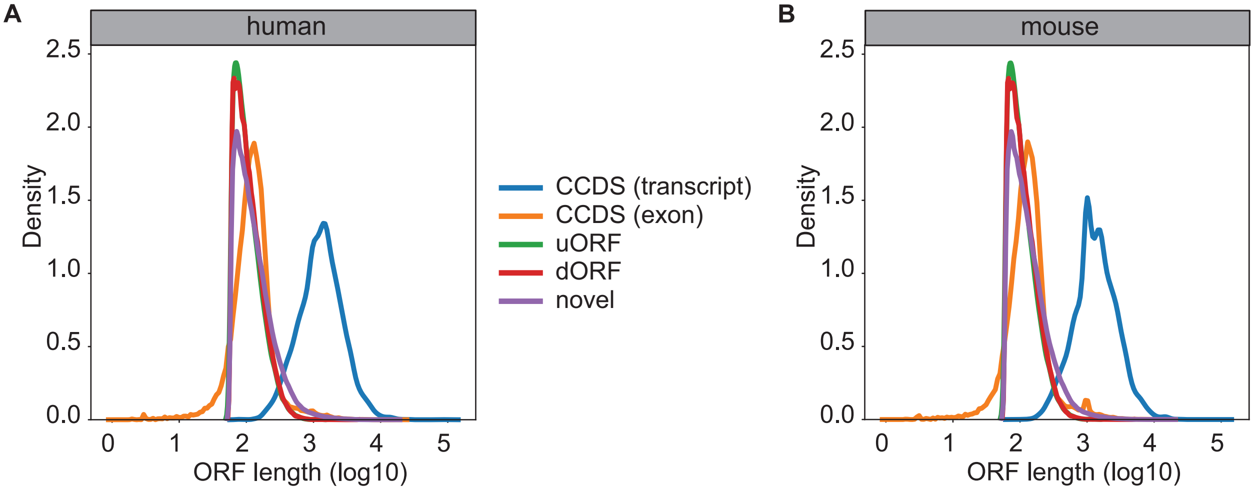


Figure 3.1

Length distribution of candidate ORFs for human and mouse. The length distribution of uORFs, dORFs, and novel ORFs predicted from presumably non-coding genes compared with the CCDS exon and CCDS transcript lengths (CCDS = Canonical Coding Sequence; uORF = upstream ORF in 5' UTR; dORF = downstream ORF in 3' UTR; novel = candidate ORFs in annotated non-coding genes.)

Transcriptome-wide searches for pairs of in-frame start and stop codons defining potential ORFs in human, and mouse genomes reveal that the sizes of such non-canonical ORFs are generally 10-20 fold shorter [79] than the conventional coding sequences (CDS) (Figure 3.1). Their short size presents challenges in detecting the resulting peptides through proteomic approaches [80]. However, there is emerging evidence that these short ORFs, or the products of their translation, serve some function [81, 58]. In particular, the role of uORFs in regulating the translation of downstream CDS has been well documented [82] for individual genes [83], and they are correlated with substantial (30%-80%) repression of protein production [79]. The same mechanism is also used to encode condition-specific activation: in integrated stress response, where the repressed state is the default, uORF-associated repression is released following the stress stimulus [84].

The presence of amplification bias, non-ribosomal RNA-protein complexes or other non-ribosomal contamination can often result in apparent RPFs that do not represent actively-translating ribosomes. Some RNAs such as telomerase RNA, RNAse P, snRNAs, and snoRNAs that are known to be "classical" non-coding RNAs and are predominantly localized in the nucleus have also been reported as origin for RPFs [30]. This is an indication that not all RPFs represent actively-translating ribosomes. Such fragments could represent non-ribosomal protected regions such as those protected by RNA binding proteins. When drawing any conclusion about translational regulation from Ribo-seq data it is imperative to focus only on those fragments that represent actively-translating ribosomes. However, the presence of noise in the data makes the task of identifying actively translated regions challenging. A shorter translation unit means less total data on average for inference, so detection of short ORFs in Ribo-seq has remained especially difficult.

Several methods exist for analyzing Ribo-seq data to determine the coding potential of the transcribed RNA. FLOSS [31], one of the earliest methods, identifies actively translating ORF by focusing on the read length distribution. The key assumption is that the distribution of sequenced fragments contains both RPFs and technical noise, and the true RPFs should exhibit a particular length distribution. FLOSS first learns a reference distribution of RPF lengths on a set of protein-coding genes likely to represent active translation, and then compares fragment lengths through the other regions in the transcriptome to this reference distribution. The idea of treating different fragment lengths separately has been adopted in several subsequent methods. Most other methods can be understood broadly through two paradigms. The first hypothesizes that the distribution of number of mapped fragments differs over actively translated regions, and compares this distribution with some selected null model. The other general approach exploits the periodic pattern in the mapped fragment profiles to distinguish actively-translating regions.

In the first paradigm of methods, ORFscore [85] compares the distribution of reads falling in the three frames to a uniform distribution. ORF-RATER [33] uses a combination of regression and random-forest based classification to predict actively-translating ORFs. It uses a non-negative least squares fit for regressing Ribo-seq read profile of the transcript against the profile obtained from known protein-coding genes. A random-forest classifier then uses these scores to predict the translational status of the ORF. RiboHMM [86], on the other hand, uses a hidden Markov model (HMM) to detect translating ORFs. It models the contribution of each fragment length separately and then combines them to increase sensitivity. The HMM learns the distributions of Ribo-seq coverage over the start/stop codons and the translated CDS; the distributions are then used to predict translation status for candidate ORFs. Rp-Bp [36] uses probabilistic modeling to estimate if read counts at each position belong to an enriched model or a null uniform model. RiboCode [87] uses a modified Wilcoxon signed-rank test [88] to assess periodicity by testing for differential enrichment in one of the frames against the other two.

The second paradigm typically leverages spectral approaches to examine the periodic pattern in Ribo-seq data. Mapping RPFs from Ribo-seq onto the mRNA is expected to reveal a "high-low-low" pattern, owing to ribosome's movement over codons, resulting in a three-nucleotide periodicity. RiboTaper [35] uses multi-tapered windows for calculating a Fourier transform to assess periodicity in the Ribo-seq signal. Based on related principles in signal processing, SPECtre [89] makes use of spectral coherence to correlate Ribo-seq signal with the expected "high-low-low" pattern. RiboWave [90] uses a wavelet transform based method to denoise the RPF profile by extracting the three-nucleotide periodicity. This denoised RPF profile leads to a better performance when identifying active translation.

Methods within both paradigms have enabled discovery of actively-translating ORFs. Each method makes assumptions about the data that are not always satisfied in practice, for different data sets or different data analysis goals. The detection of short ORFs is an example of the latter. However, these methods provide a conceptual foundation that we borrow from to design a simplified method that is more robust to varying statistical features across datasets, and that is capable of detecting both short and long ORFs. Our method, called ribotricer, directly assesses the three-nucleotide periodicity in Ribo-seq data. Ribotricer can account for read length specific P-site offsets and sparsity in Ribo-seq data. Its underlying model emphasizes consistency in the qualitative profile through each codon while down-weighting the influence of the magnitude of the individual values contributing to that profile. This approach helps ribotricer overcome the challenge of detecting short ORFs in regions of low signal to noise ratio.

3.2 Methods

To detect actively-translating ORFs, ribotricer focuses on the characteristic three-nucleotide periodicity in Ribo-seq data. The workflow of ribotricer consists of five major steps. Ribotricer first prepares a candidate set of all potentially translatable ORFs by searching for pairs of start and stop codons genome-wide but inside annotated transcription units. This requires providing gene annotations and the reference genome but is only done once for each genome and gene annotation. Next, ribotricer partitions the mapped reads based on their length. The rationale for processing reads by their length is that each length may be associated with a different P-site offset relative to the 5' end of the mapped fragment. For each read length, ribotricer generates a metagene profile using 5' ends of the mapped reads (accounting for strand as appropriate). The metagene profiles are used to infer P-site offsets for different read lengths by choosing the offsets that maximize the cross-correlation of these profiles with the profile for the most abundant read length. The read profiles corresponding to different read-lengths can then be merged using the corresponding inferred P-site offsets, an approach taken previously by Calviello *et al.* for RiboTaper [35] and Xiao *et al.* for RiboCode [87]. The previous step produces a single RPF profile for each candidate ORF. In its final step, ribotricer assesses the periodicity of the merged RPF profile using a novel approach to predict its translation status.

Our key contribution is a novel method for assessing the three-nucleotide periodicity of RPF profile based on 3D to 2D projection (Figure 3.2). Within each codon, we may observe reads with 5' ends at each of the three nucleotides, providing three unconstrained count values. These count values can be imagined as vectors in a three-dimensional space with each nucleotide position representing one dimension. Based on early observations, we hypothesized that in practice, using information related to total read abundance might obscure the signal. This would happen, for example, if codons that are assigned a higher total read count are also more likely to receive reads associated with non-active translation. Applying any spectral method would require that the profiles satisfy conditions to ensure stationarity. Instead, we rely on using the qualitative information at each codon in the form of “high-low-low” or related pattern. This approach discards much of the quantitative information associated with individual read counts but also simplifies the problem while eliminating the need to explicitly model random variation or systematic trend in total read counts along the RPF profile.

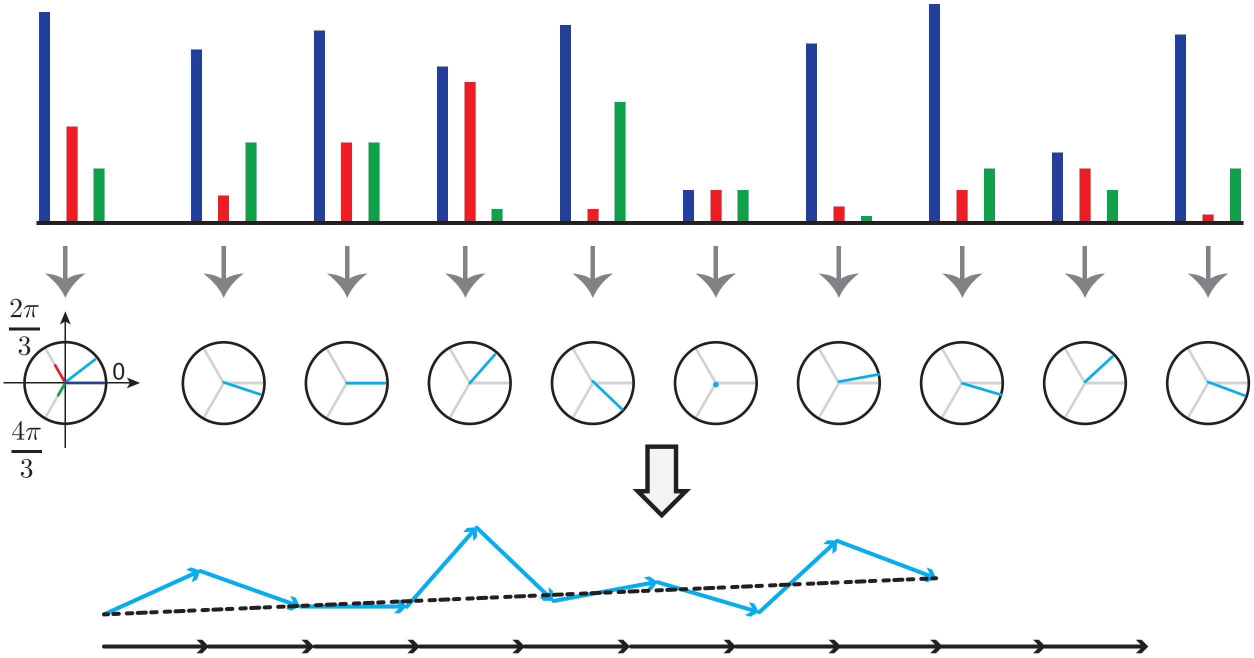


Figure 3.2
Methodology design of ribotricer.

For a given ORF consisting of N codons, let x_{ij} denote the number of P-sites inferred from the reads of Ribo-seq experiment aligning to the i-th codon and j-th frame of the ORF, where $i = 1, 2, \dots, N$ and $j = 1, 2, 3$. The RPF profile of the ORF can then be denoted as $P = (x_{11}, x_{12}, x_{13}, \dots, x_{N1}, x_{N2}, x_{N3})$. For each codon profile $x_i = (x_{i1}, x_{i2}, x_{i3})$, a 3D vector, we perform the following transformation to convert it into a 2D unit vector $\phi_i = (a_i, b_i)^{\mathrm{T}}$:

$$\phi_i = \frac{w x_i^{\mathrm{T}}}{\|w x_i^{\mathrm{T}}\|}, \tag{3.1}$$

where

$$w = \begin{pmatrix} 1 & \cos(-2\pi/3) & \cos(-4\pi/3) \\ 0 & \sin(-2\pi/3) & \sin(-4\pi/3) \end{pmatrix}.$$

With this transformation, the three basis vectors $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ are mapped as

$$
\begin{aligned}
&(1, 0, 0) \rightarrow (1, 0),\\
&(0, 1, 0) \rightarrow (\cos(-2\pi/3), \sin(-2\pi/3)),\\
&(0, 0, 1) \rightarrow (\cos(-4\pi/3), \sin(-4\pi/3)).
\end{aligned}
$$

The three mapped unit vectors lie $2\pi/3$ away from each other to ensure the direction of the transformed vector ϕ_i is equally determined by reads of each frame. These can be replaced by any three unit vectors that are equally spaced on the unit circle, and the results would not change.

For the transformation performed, the direction of the resulting vector is determined by the relative values of x_{i1}, x_{i2}, and x_{i3}. For an actively-translating ORF, we expect to see a "high-low-low" pattern for each codon. This is equivalent to observing x_{i1} as the largest value consistently over all codons. If this holds, we expect the directions of the resulting unit vectors ϕ_i to be consistent across codons. As indicated above, the motivation behind unit normalization of each vector is to help ensure that each codon contributes equally to our assessment of translation status, avoiding bias from the fraction of codons with an over-abundance of reads. This transformation disregards the total read counts at each of the three positions. For example, the two codon profiles $(100, 20, 10)$ and $(10, 2, 1)$ will result in the same unit vectors when applying equation Eq., (3.1). While this discards quantitative information, it still captures the qualitative "high-low-low" pattern of the profile. This approach helps ribotricer handle the heterogeneous nature of Ribo-seq data where despite of pervasive active-translation, different codons could have completely different coverages either because of the actual difference in ribosome's dwell time or because of usage of drugs like cycloheximide which can alter codon-specific elongation rates [91].

The l^2-norm of the mean vector of the transformed vectors can be used to assess the periodicity of RPF profile. More consistent directions of the vectors would result in a larger l^2-norm. The mean vector of the transformed vectors is

$$\bar{\phi} = \frac{1}{N} \sum_{i=1}^{N} \phi_i,$$

and its l^2-norm $\|\bar{\phi}\|$ is

$$\|\bar{\phi}\| = \sqrt{\left(\tfrac{1}{N} \sum_{i=1}^{N} a_i\right)^2 + \left(\tfrac{1}{N} \sum_{i=1}^{N} b_i\right)^2},$$

which falls in [0, 1], with a value of 1 if and only if

$$a_1 = a_2 = \cdots = a_N,$$

$$b_1 = b_2 = \cdots = b_N,$$

in which case the directions for all vectors are the same.

Besides heterogeneity arising from uneven distribution of read counts across codons [92], another key challenge in Ribo-seq data is sparsity leading to profiles with many empty codons, i.e., codons to which no reads map. Since both actively-translating and non-translating ORFs can have empty codons, they do not contribute any information about the translation status and hence need to be handled specially. For a particular data set with N codons, define the set V of non-empty codons as

$$V = \{i = 1, 2, \dots, N \mid x_i \neq (0, 0, 0)\},$$

and let $N_v = |V|$. If we define $\bar{\phi}^*$ as the mean vector including only non-empty codons, the ratio between $\|\bar{\phi}\|$ and $\|\bar{\phi}^*\|$ is

$$\frac{\|\bar{\phi}\|}{\|\bar{\phi}^*\|} = \frac{N_v}{N}.$$

With the reasoning outlined above, we use $\|\bar{\phi}^*\|$ as our measure for assessing the periodicity of the RPF profile of an ORF. This score describes how "aligned" all the vectors are, and is equivalent to measuring how similar the phases are, i.e., the angles created by the resulting vectors with respect to the abscissa. We will refer to this score as the "phase score" hereafter. Note that in theory, a high phase score may result from strong consistency of some pattern other than the anticipated "high-low-low". In designing our approach, we hypothesized that the only source of consistency in the signal would be an active translation. A consistent "low-high-low" or "low-low-high" pattern would most likely result from an inaccurate estimate of the P-site offsets, in which case our assumptions add a layer of robustness.

The angles made by the resultant vectors when all the codons follow a "high-low-low" pattern should be concentrated around 0. The distribution we observe for the Ribo-seq data is centered around 0 (Figures 3.7 and 3.8), which confirms that most codons follow the "high-low-low" pattern. For the RNA-seq data, the resulting angles follow a multimodal distribution with the highest peaks at $\{-2\pi/3, 0, 2\pi/3\}$ (Figures 3.7 and 3.8) which corresponds to the three unit vectors. To interpret the multimodal distribution observed in RNA-seq data, we simulated read counts using a Poisson distribution. To account for variation in total data between genes, we simulated means of the Poisson distribution using the per nucleotide coverage from the RNA-seq.

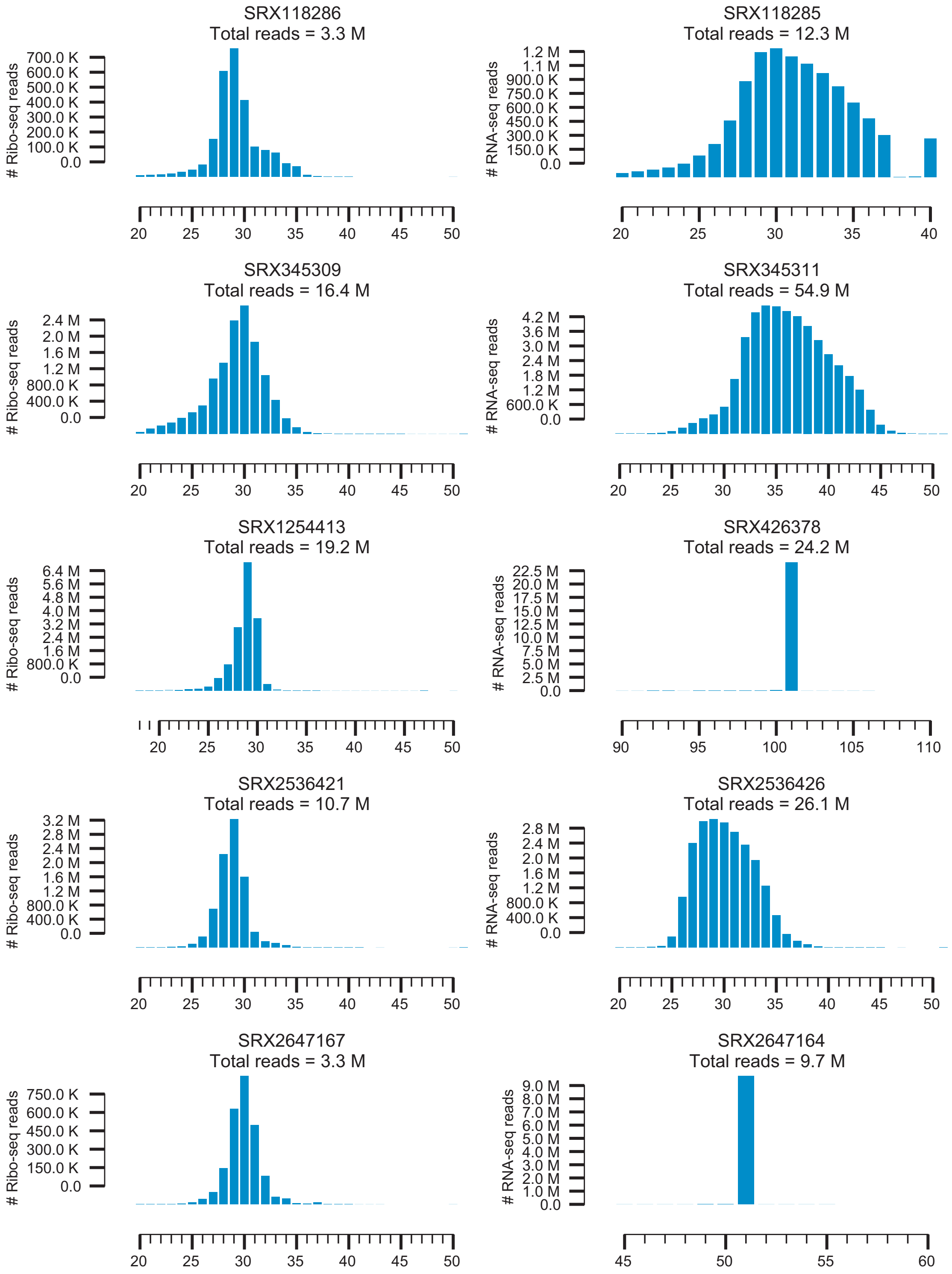


Figure 3.3
Read length distribution of Ribo-seq and RNA-seq samples from human datasets. SRA sample accession and total uniquely mapping reads are shown in individual subplots.

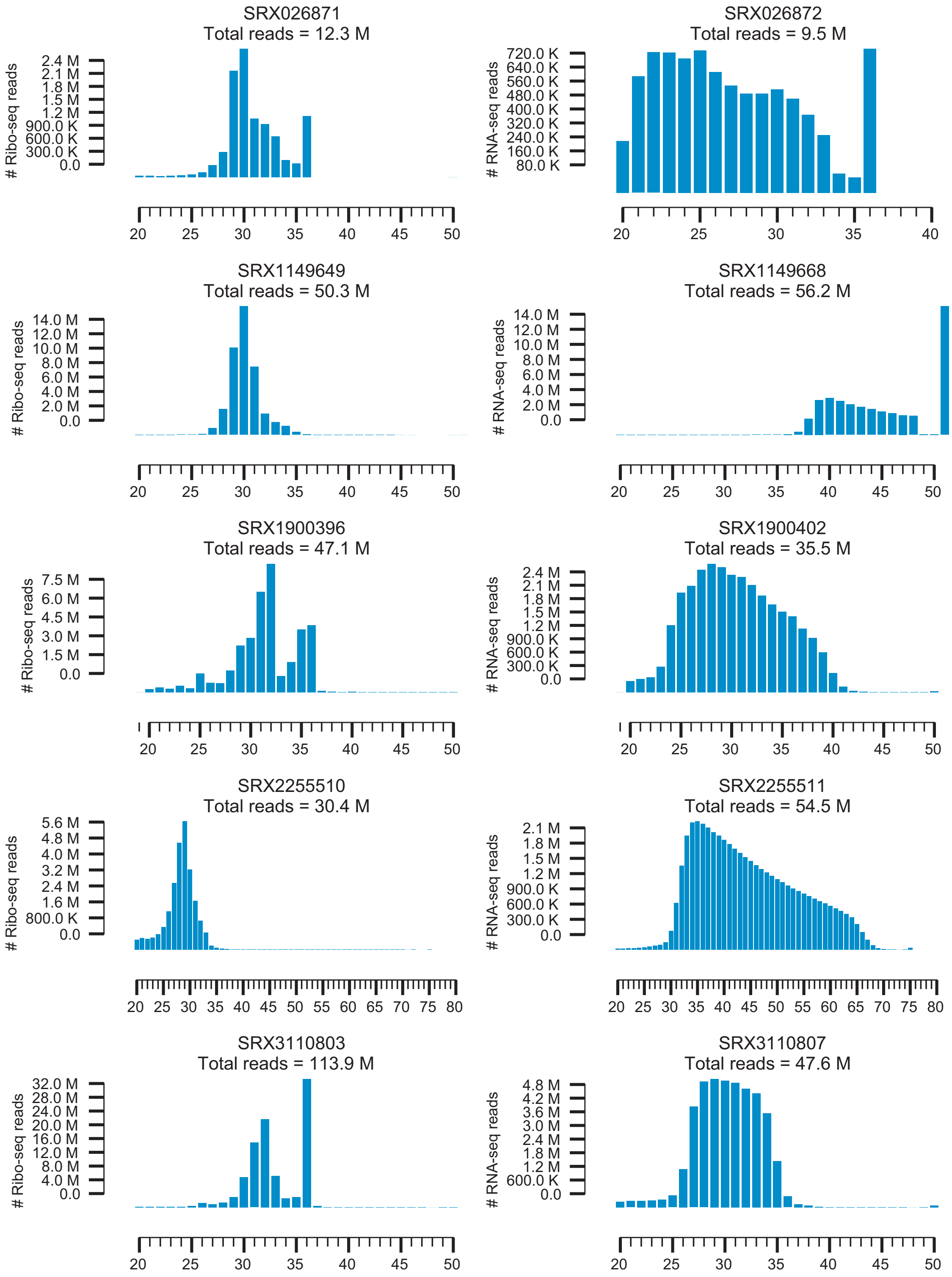
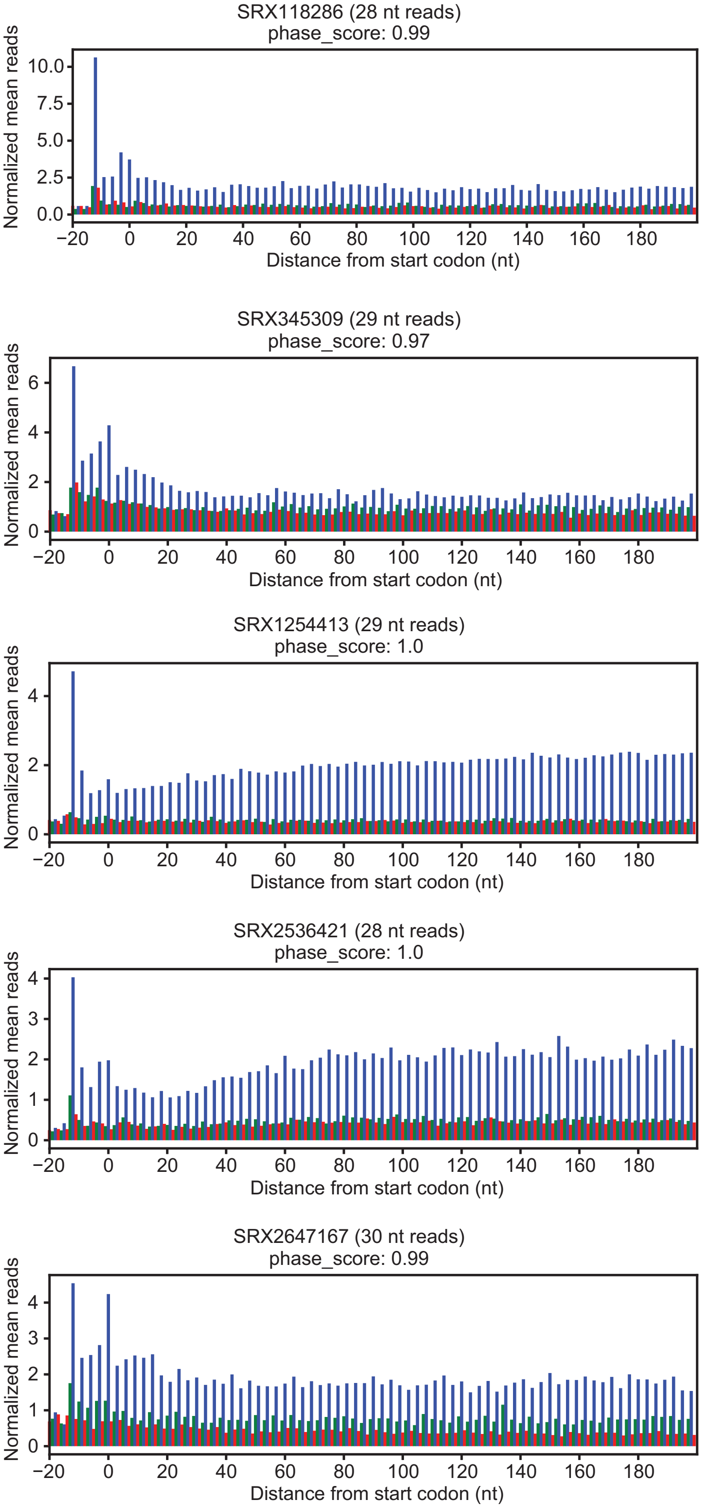


Figure 3.4
Read length distribution of Ribo-seq and RNA-seq samples from mouse datasets. SRA sample accession and total uniquely mapping reads are shown in individual subplots.


SRX118286 (28 nt reads)
phase_score: 0.99

SRX345309 (29 nt reads)
phase_score: 0.97

SRX1254413 (29 nt reads)
phase_score: 1.0

SRX2536421 (28 nt reads)
phase_score: 1.0

SRX2647167 (30 nt reads)
phase_score: 0.99

Figure 3.5
Metagene plots for representative read lengths for human Ribo-seq samples. SRA sample accession, read length and phase score are shown in individual subplots.

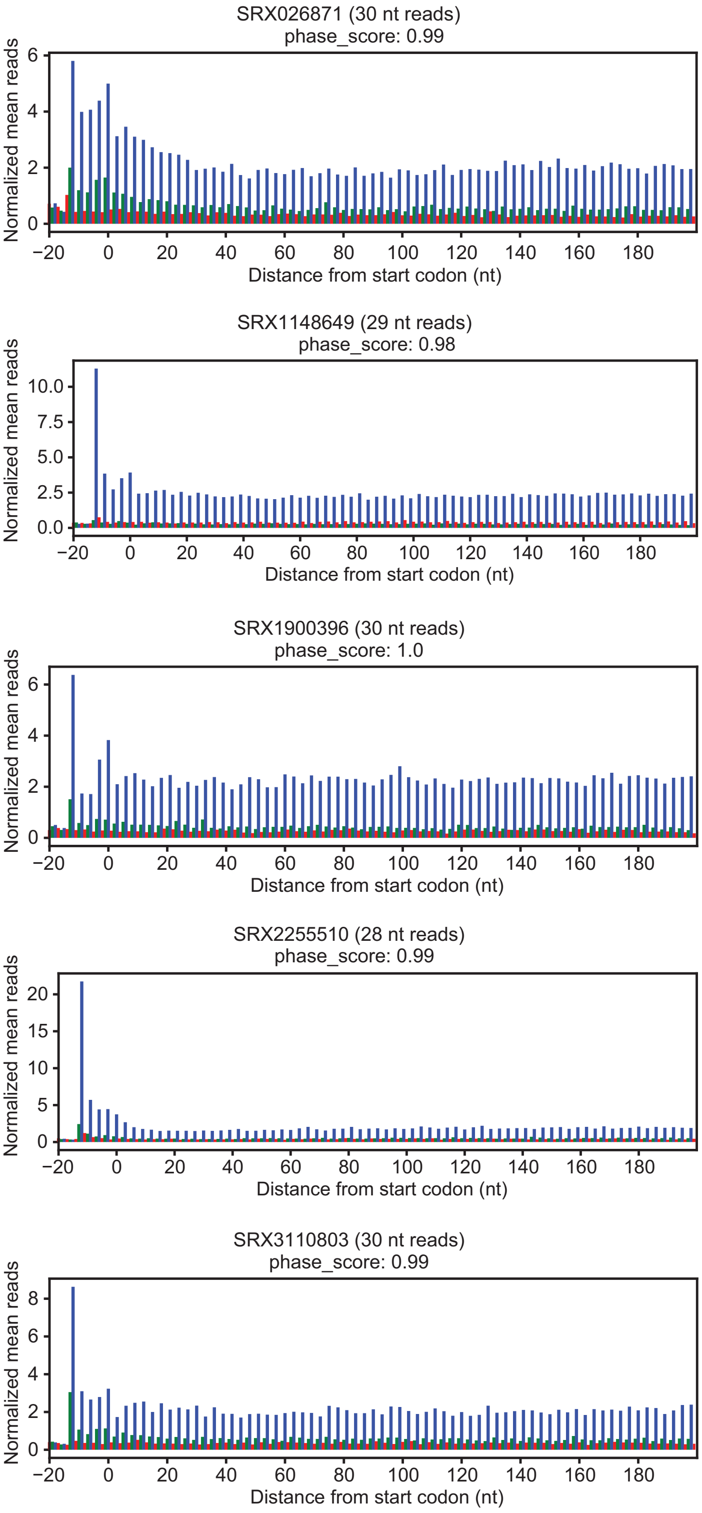


Figure 3.6

Metagene plots for representative read lengths for mouse Ribo-seq samples. SRA sample accession, read length and phase score are shown in individual subplots.

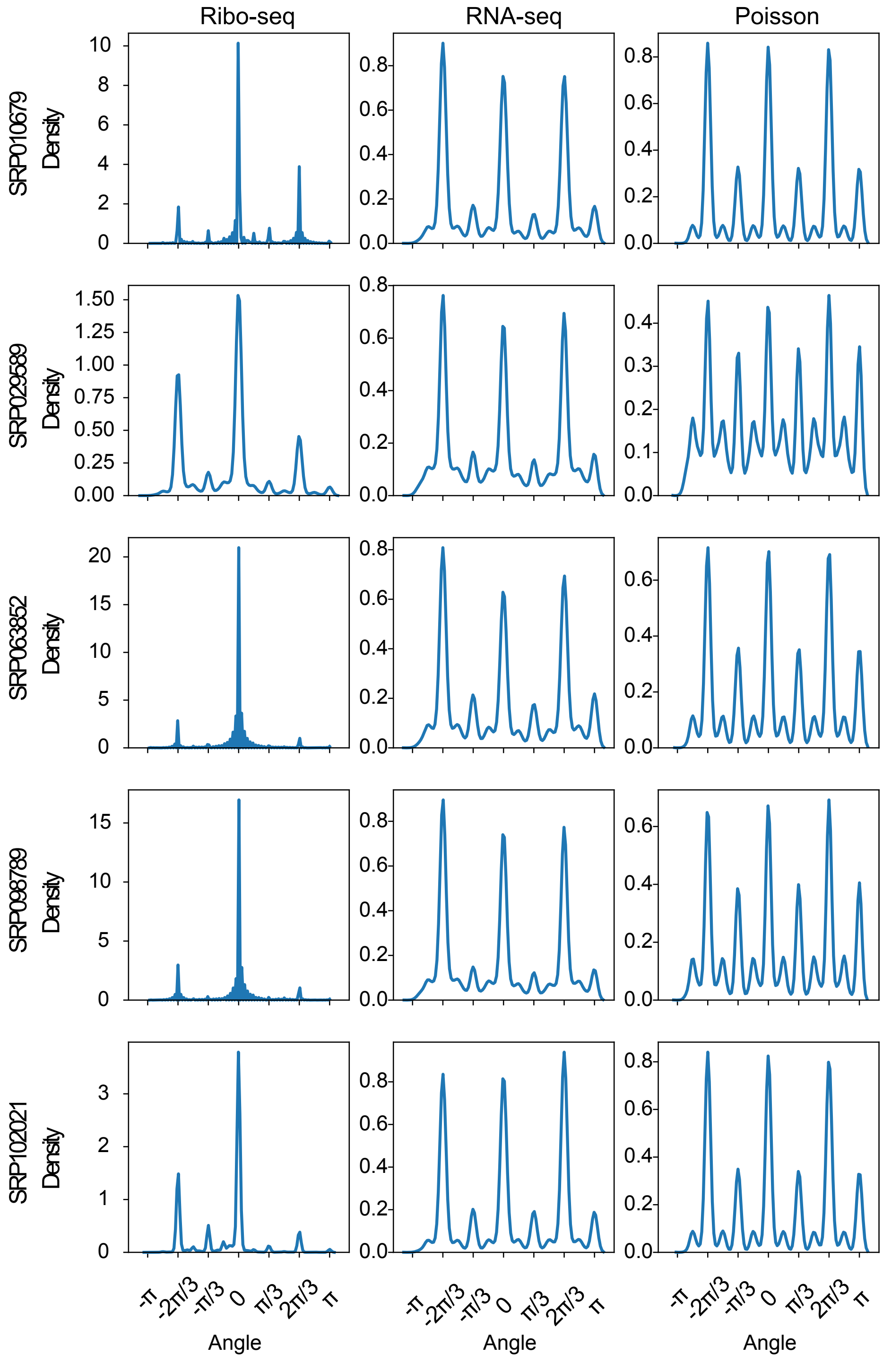


Figure 3.7
Distribution of the resulting vector angles for datasets in human. Angles are formed by projecting the CCDS 3D codon profiles to 2D unit vectors. The left sub-panel indicates the distribution for Ribo-seq sample; the center sub-panel shows the distribution for its corresponding RNA-seq sample; the right sub-panel shows the distribution of angles resulting from a RNA-seq profile simulated from a Poisson distribution with the mean parameter estimated from the RNA-seq data.

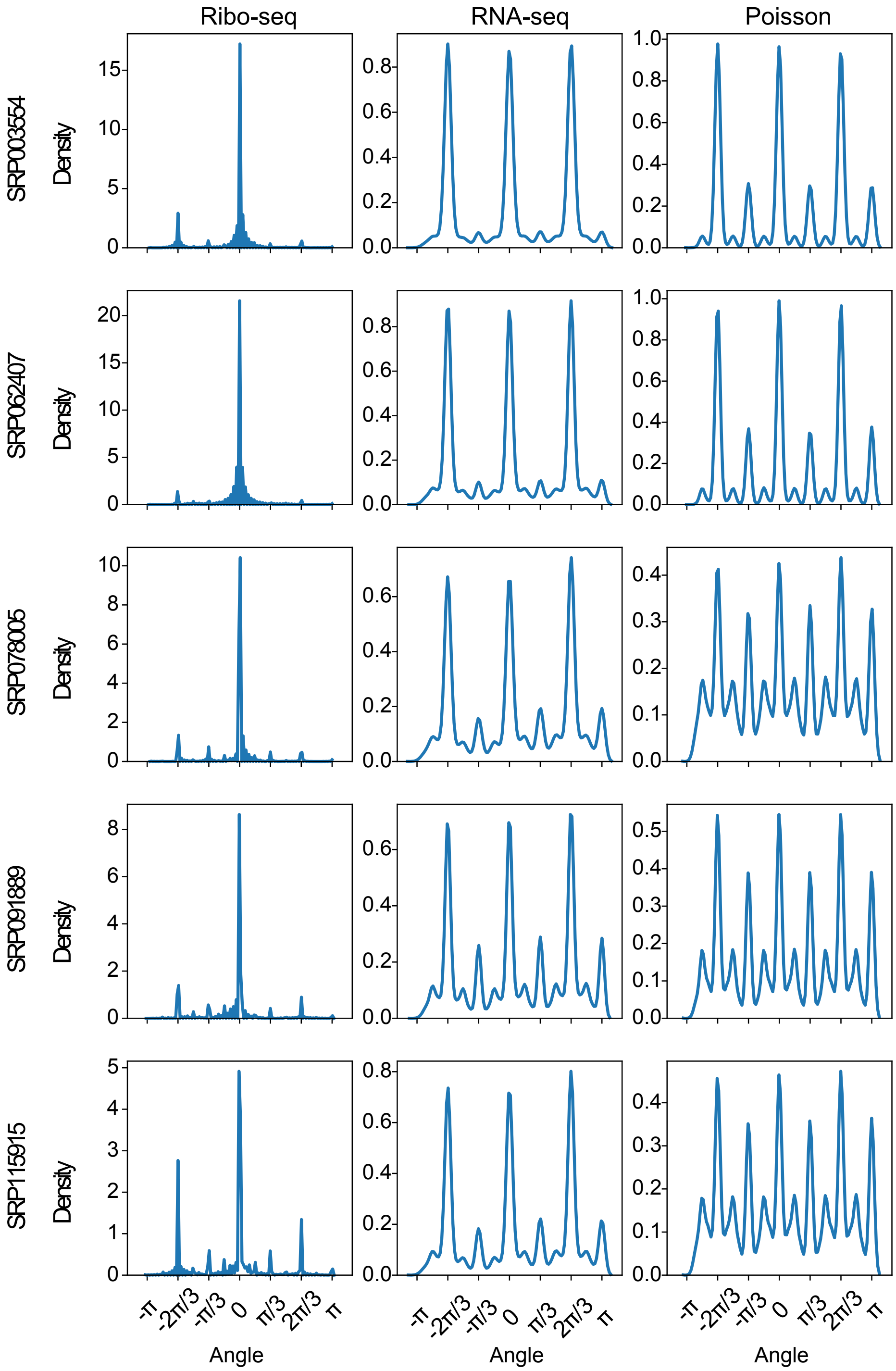


Figure 3.8

Distribution of the resulting vector angles for datasets in mouse. Angles are formed by projecting the CCDS 3D codon profiles to 2D unit vectors. The left sub-panel indicates the distribution for Ribo-seq sample; the center sub-panel shows the distribution for its corresponding RNA-seq sample; the right sub-panel shows the distribution of angles resulting from a RNA-seq profile simulated from a Poisson distribution with the mean parameter estimated from the RNA-seq data.

The resulting angle distribution of the simulated codon profiles is similar to that obtained from profiles of the RNA-seq data (Figures 3.7 and 3.8) which explains the observed multimodality.

3.2.1 Obtaining and pre-processing data

We downloaded the raw data (Table 3.4) from NCBI's Sequence Read Archive (SRA) using `pysradb` [93]. We used `cutadapt` [94] to perform adapter trimming. The specific adapters for each dataset are either obtained from the corresponding papers or were automatically inferred by checking for over-represented $k-$mers at the 3'end. Sequences of the adapters for each dataset is documented in Table 3.1. All the Ribo-seq and RNA-seq data were mapped using `STAR` [95] by allowing at most two mismatches (`--outFilterMismatchNmax 2`) and forcing end-to-end (`--alignEndsType EndToEnd`) read alignment. Only uniquely mapping reads were retained (`-outFilterMultimapNmax 1`). For human and mouse, we relied on the GENCODE [96] GTF for annotation. For all other species except *C. albicans*, we used ENSEMBL [97]. For *C. albicans*, both the FASTA and the GTF were obtained from the Candida Genomes database [98]. The assembly and GTF information is summarized in Table 3.2. FASTA is handled using the pyfaidx package [99].

The strand-specific protocol, either forward stranded, reverse stranded or unstranded, is inferred by checking the first $20,000$ reads from the mapping results. Since most tools we compared with can only deal with forward stranded protocol, our ten datasets are all forward stranded for both RNA-seq and Ribo-seq samples. BAM files are processed using pysam, a python interface to samtools [100].

Table 3.1
Adapters trimmed from Ribo-seq and RNA-seq samples for each dataset.

SRA Accession	Ribo-seq adapter	RNA-seq adapter
SRP010679	CTGTAGGCAC	CTGTAGGCAC
SRP029589	CTGTAGGCACCATCAAT	CTGTAGGCACCATCAAT
SRP063852	None	None
SRP098789	CTGTAGGCACCATCAAT	CTGTAGGCACCATCAAT
SRP102021	TCGTATGCCGTCTTCTGCTTG	None
SRP003554	TCGTATG	TCGTATG
SRP062407	TGGAATTCTCGGGTGCCAAGG	TGGAATTCTCGGGTGCCAAGG
SRP078005	TGGAATTCTCGGGTGCCAAGG	TGGAATTCTCGGGTGCCAAGG
SRP091889	AGATCGGAAGAGCACACGTCT	AGATCGGAAGAGCACACGTCT
SRP115915	TGGAATTCTCGGGTGCCAAGG	TGGAATTCTCGGGTGCCAAGG
SRP108862	TGGAATTCTCGG	AGATCGGAAGAGC
SRP087624	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP029587	TCGTATGCCGTCTTCTGCTTG	TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC
SRP059391	TGGAATTCTCGG	TGGAATTCTCGG
SRP018118	TGGAATTCTCGG	TGGAATTCTCGG
SRP075766	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP033499	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP028614	AAAAAAAAAAA_AGATCGGAAGAGC	AAAAAAAAAAA_AGATCGGAAGAGC
SRP028552	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP000637	AAAAAAAA_AGATCGGAAGAGC	AAAAAAAA_AGATCGGAAGAGC
SRP056647	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP026198	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP014427	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP010374	AAAAAAA_AGATCGGAAGAGC	AGATCGGAAGAGC
SRP108999	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP028243	CTGTAGGCACCATCAAT	AGATCGGAAGAGC
SRP076919	AGATCGGAAGAGC	TGGAATTCTCGG
SRP045475	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP056012	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP045777	AGATCGGAAGAGC	AGATCGGAAGAGC
ERP007231	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP034750	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA	AGATCGGAAGAGCACACGTCTGAACTCCAGTCA
SRP010040	ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAA	ATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAA
SRP023492	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP032814	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP107240	AGATCGGAAGAGC	AGATCGGAAGAGC
SRP062129	TGGAATTCTCGG	AGATCGGAAGAGC

Table 3.2
Reference assemblies and GTF for each species

Species	Reference assembly	GTF
Human	GRCh38	Gencode (v94)
Mouse	GRCm38	Gencode (v94)
Arabidopsis	TAIR10	ENSEMBL (v96)
C.elegans	WBcel235	ENSEMBL (v96)
Drosophila	BDGP6	ENSEMBL (v96)
Rat	Rnor6.0	ENSEMBL (v96)
Zebrafish	GRCz11	ENSEMBL (v96)
C. albicans	SC5314	Candida Genomes Database (r27)
S. pombe	ASM294v2	ENSEMBL (v96)
Chimpanzee	Pantro3	ENSEMBL (v96)
Macaque	Mmul8	ENSEMBL (v96)

To create fragment length specific metagene profile, we counted the number of 5' end of reads at each nucleotide per fragment length. Figures S2 and S3 show the distribution of fragment lengths for Ribo-seq and RNA-seq samples across different datasets in human and mouse, respectively. Metagene plots for individual fragment lengths which were retained for downstream analysis for different datasets are shown in Figures 3.5 and 3.6

The specific Ribo-seq and RNA-seq samples used from each dataset for the benchmarking along with the read lengths and the corresponding P-site offsets used for the Ribo-seq samples can be found in Table 3.3.

To evaluate the performance of ribotricer and other existing methods, acknowledging the heterogeneity and appreciable noise levels in Ribo-seq data, we selected five human and five mouse datasets for performance comparison (Table 3.4). This includes the human HEK293 cells dataset (SRA accession: SRP063852) [35], which was originally used as a benchmark dataset when RiboTaper was introduced [35] and subsequently used in other studies. We followed the strategy previously established by Calviello *et al.* in assessing RiboTaper [35] and Xiao *et al.* in assessing RiboCode [87]. For all the ten datasets, we obtained the RPF profiles for all the CCDS

Table 3.3
Ribo- and RNA-seq samples used for the benchmarking along with the read lengths and P-site offsets used for Ribo-seq samples.

SRA Accession	Ribo-seq sample	Read lengths (nt)	P-site offsets (nt)	RNA-seq sample	Species
SRP010679	SRX118286	28,29,30	12,13,13	SRX118285	Human
SRP029589	SRX345309	29,30,32	12,12,13	SRX345311	Human
SRP063852	SRX1254413	28,29,30	12,12,12	SRX426378	Human
SRP098789	SRX2536421	28,30	12,13	SRX2536426	Human
SRP102021	SRX2647167	28,29,30,31	12,12,12,12	SRX2647164	Human
SRP003554	SRX026871	28,29,30	12,12,12	SRX026872	Mouse
SRP062407	SRX1149649	28,29,30,31	12,12,12,12	SRX1149668	Mouse
SRP078005	SRX1900396	26,27,28,29,30	12,12,12,12,12	SRX1900402	Mouse
SRP091889	SRX2255510	26,27,28,29,30	12,12,12,12,12	SRX2255511	Mouse
SRP115915	SRX3110803	29,30,31,32,33,34	12,12,12,13,13,13	SRX3110807	Mouse
SRP108862	SRX2896566	23	12	SRX2896570	Arabidopsis
SRP087624	SRX2148419	28,29,30,31,32	12,12,12,12,12	SRX2148418	Arabidopsis
SRP029587	SRX345240	26,27	12,12	SRX345251	Arabidopsis
SRP059391	SRX1056790	27,30	12,12	SRX1056791	Arabidopsis
SRP018118	SRX219170	28,29,30,31	11,12,13,13	SRX347226	Arabidopsis
SRP075766	SRX1801603	26,27,28	11,12,13	SRX1801650	Baker's Yeast
SRP033499	SRX386988	29,30,31	12,12,12	SRX386983	Baker's Yeast
SRP028614	SRX333052	28,29,30	12,13,13	SRX334053	Baker's Yeast
SRP028552	SRX332185	28,29,30	11,12,12	SRX332188	Baker's Yeast
SRP000637	SRX003187	28,29,30,31	12,12,12,12	SRX003191	Baker's Yeast
SRP056647	SRX971770	28,29,30,31,32	12,12,12,12,12	SRX971774	*C. elegans*
SRP026198	SRX311784	29,30,31,32	12,12,12,12	SRX311777	*C. elegans*
SRP014427	SRX160518	28,29,30,31,32	12,12,12,12,12	SRX160149	*C. elegans*
SRP010374	SRX118118	28,29,30,31,32	12,12,12,12,12	SRX118116	*C. elegans*
SRP108999	SRX2902857	29,30,31,32	12,13,10,12	SRX2902867	Drosophila
SRP028243	SRX327686	28,29,30,32,33,34	12,12,12,12,12,13	SRX327688	Drosophila
SRP076919	SRX1870218	34	12	SRX1870191	Drosophila
SRP045475	SRX679371	28,29,30,31,32	12,12,12,12,12	SRX679372	Drosophila
SRP056012	SRX915217	29,30,31,32	12,12,13,13	SRX915210	Rat
SRP045777	SRX686499	28,29,30,31	12,12,12,13	SRX686500	Rat
ERP007231	ERX609893	28,29,30,31,32	12,12,12,12,12	ERX609898	Rat
SRP034750	SRX399800	28,29,30,31	12,12,12,12	SRX399817	Zebrafish
SRP010040	SRX113357	27,28,30,31,33,34	12,12,12,12,12,12	SRX113344	Zebrafish
SRP023492	SRX288475	28,29,30	12,12,12	SRX288474	Zebrafish
SRP032814	SRX375317	28,29,30	12,12,12	SRX375318	*C. albicans*
SRP107240	SRX2825796	28,29,30	12,13,13	SRX2825805	*S. pombe*
SRP062129	SRX1135820	28,29,30	12,12,12	SRX333018 (SRP028612)	Chimpanzee
SRP062129	SRX1135825	28,29,30	12,12,12	SRX333023 (SRP028612)	Macaque

from the results generated by RiboTaper and used the expressed CCDS profiles from Ribo-seq data as true positives and the corresponding CCDS profiles from RNA-seq data as true negatives. Since RiboTaper was designed and benchmarked for detecting active translation at the exon level, we split the existing methods for active translation detection into two groups; those that support detection at the exon level and those that only allow detection at the transcript level. We compared the performance of ribotricer at both the exon and transcript levels.

3.2.2 Learning cutoff of phase score

The phase score is indicative of how consistent the profile is through a defined region. We require some cutoff to distinguish phase scores that differentiate active from non-active translation, with the latter representing either some form of noise or inactive translation. Our approach is to learn this cutoff empirically using ten published datasets (Table 3.4) with an assumed ground truth set for regions of active translation and regions lacking active translation. Taking this strategy, we used RPF profiles of expressed Consensus Coding Sequence (CCDS) [101] exons from Ribo-seq data as the true positives, and mapped read profiles from RNA-seq data for a negative control, as previously described [35, 87]. In order to choose the best cutoff, we relied on maximizing the F1 score statistic. F1 score represents the harmonic mean of precision and recall and is considered a more realistic measure of a classifier's performance than precision or recall in isolation. The chosen cutoff should apply universally across all Ribo-seq datasets and hence should remain independent of our choice of the datasets. The ten datasets we chose in human and mouse (Table S1) might not necessarily be representative of all Ribo-seq datasets. Hence, we performed a bootstrap [102] analysis with $50,000$ bootstraps using selected datasets, defining the bootstrap statistic as the phase score that maximizes the F1 score resulting in a 95% confidence interval of $(0.419, 0.469)$ (Figure 3.10). This resulted in a mean cutoff score of 0.444. However, for our benchmarking at the exon and the transcript level, we wanted to take the most conservative approach avoiding any bias that might arise from learning this cutoff on all the ten datasets. Hence, we used only two human and two mouse datasets (SRA accession: SRP010679 [103], SRP098789 [1], SRP003554 [104], and SRP115915 [105]), to learn the cutoff (0.428) and used the remaining six datasets to assess the performance. Note that this chosen cutoff still resides in the 95% confidence interval of our bootstrap analysis and hence can be applied universally for all other datasets.

3.3 Results

3.3.1 Ribotricer accurately detects translating ORFs at the exon level

We evaluated the performances of methods that support exon-level detection of translation, including ORFscore [85], RiboTaper [35], and RiboCode [87], and compared their performance with that of ribotricer.

We first compared the ability of each method to distinguish Ribo-seq profiles from RNA-seq using the area under the receiver operating characteristic (ROC) and precision-recall (PR) curve. For human HEK293 cells dataset (SRA accession: SRP063852) [35], ribotricer achieved an area under the ROC (AUROC) of 0.97. The second best one was achieved by RiboCode with an AUROC of 0.93. RiboTaper and ORFscore achieved an AUROC of 0.88 and 0.87, respectively (Figure 3.15A). For the mouse liver tissue dataset (SRA accession: SRP078005) [108], ribotricer achieved an AUROC of 0.99 while RiboCode, RiboTaper, and ORFscore achieved AUROC of 0.97,

Table 3.4
List of datasets.

SRA Accession	Species	Cell type	Treatment	Citation
SRP010679	Human	PC3	100 g/ml cycloheximide	[103]
SRP029589	Human	HeLa	cycloheximide	[25]
SRP063852	Human	HEK293	100 g/ml cycloheximide	[35]
SRP098789	Human	HeLa	100 g/ml cycloheximide	[1]
SRP102021	Human	H1933	100 g/ml cycloheximide	[106]
SRP003554	Mouse	neutrophils	100 g/ml cycloheximide	[104]
SRP062407	Mouse	hippocampal neurons	100 g/ml cycloheximide	[107]
SRP078005	Mouse	liver	200 g/ml cycloheximide	[108]
SRP091889	Mouse	ESC	cycloheximide	[109]
SRP115915	Mouse	liver	200 g/ml cycloheximide	[105]
SRP108862	Arabidopsis	inflorescences	unavailable	unpublished
SRP087624	Arabidopsis	leaf tissue	50 g/ml cycloheximide	[110]
SRP029587	Arabidopsis	whole seedlings	50 g/ml cycloheximide	[111]
SRP059391	Arabidopsis	leaf tissue	100 g/ml cycloheximide	[112]
SRP018118	Arabidopsis	etiolated seedling	100 g/ml cycloheximide	[113]
SRP075766	Baker's Yeast	strain by4743	100 g/ml cycloheximide	[114]
SRP033499	Baker's Yeast	strain: by4741	0.1 mg/ml cycloheximide	[115]
SRP028614	Baker's Yeast	strain: by4176	cycloheximide	[116]
SRP028552	Baker's Yeast	strain: s288	cycloheximide	[117]
SRP000637	Baker's Yeast	strain: by4741	100 g/ml cycloheximide	[118]
SRP056647	*C. elegans*	strain: n2	100 g/ml cycloheximide	[119]
SRP026198	*C. elegans*	strain: n2	100 g/ml cycloheximide	[120]
SRP014427	*C. elegans*	strain: n2	cycloheximide	[121]
SRP010374	*C. elegans*	strain: n2	cycloheximide	[122]
SRP108999	Drosophila	body wall muscle	100 g/ml cycloheximide	[123]
SRP028243	Drosophila	embryo	20 g/ml emetine	[124]
SRP076919	Drosophila	oocytes	100 g/ml cycloheximide	[125]
SRP045475	Drosophila	S2 cell	100 g/ml cycloheximide	[126]
SRP056012	Rat	PC12 Cells	100 g/ml streptomycin	[127]
SRP045777	Rat	Pheochromocytoma	streptomycin	[128]
ERP007231	Rat	strain: bn/shr	0.1 mg/ml cycloheximide	[129]
SRP034750	Zebrafish	strain: tuab	100 g/ml cycloheximide	[32]
SRP010040	Zebrafish	strain: tuab	100 g/ml cycloheximide	[130]
SRP023492	Zebrafish	strain: tuab	50 g/ml cycloheximide	[131]
SRP032814	*C. albicans*	strain: sc5314	10 g/mL Blasticidin S	[132]
SRP107240	*S. pombe*	strain: WT	0.15 g/ml tunicamycin	[133]
SRP062129	Chimpanzee	Lymphoblastoid cell line	flash freezing	[134]
SRP062129	Macaque	Lymphoblastoid cell line	flash freezing	[134]

Table 3.5
Datasets used to learn ribotricer phase score cutoffs.

SRA Accession	Species	Used to learn cutoff
SRP010679	Human	Yes
SRP029589	Human	No
SRP063852	Human	No
SRP098789	Human	Yes
SRP102021	Human	No
SRP003554	Mouse	Yes
SRP062407	Mouse	No
SRP078005	Mouse	No
SRP091889	Mouse	No
SRP115915	Mouse	Yes
SRP108862	Arabidopsis	No
SRP087624	Arabidopsis	No
SRP029587	Arabidopsis	No
SRP059391	Arabidopsis	Yes
SRP018118	Arabidopsis	Yes
SRP075766	Baker's Yeast	Yes
SRP033499	Baker's Yeast	No
SRP028614	Baker's Yeast	No
SRP028552	Baker's Yeast	Yes
SRP000637	Baker's Yeast	No
SRP056647	*C. elegans*	No
SRP026198	*C. elegans*	Yes
SRP014427	*C. elegans*	No
SRP010374	*C. elegans*	Yes
SRP108999	Drosophila	Yes
SRP028243	Drosophila	Yes
SRP076919	Drosophila	No
SRP045475	Drosophila	No
SRP056012	Rat	Yes
SRP045777	Rat	No
ERP007231	Rat	Yes
SRP034750	Zebrafish	Yes
SRP010040	Zebrafish	Yes
SRP023492	Zebrafish	No

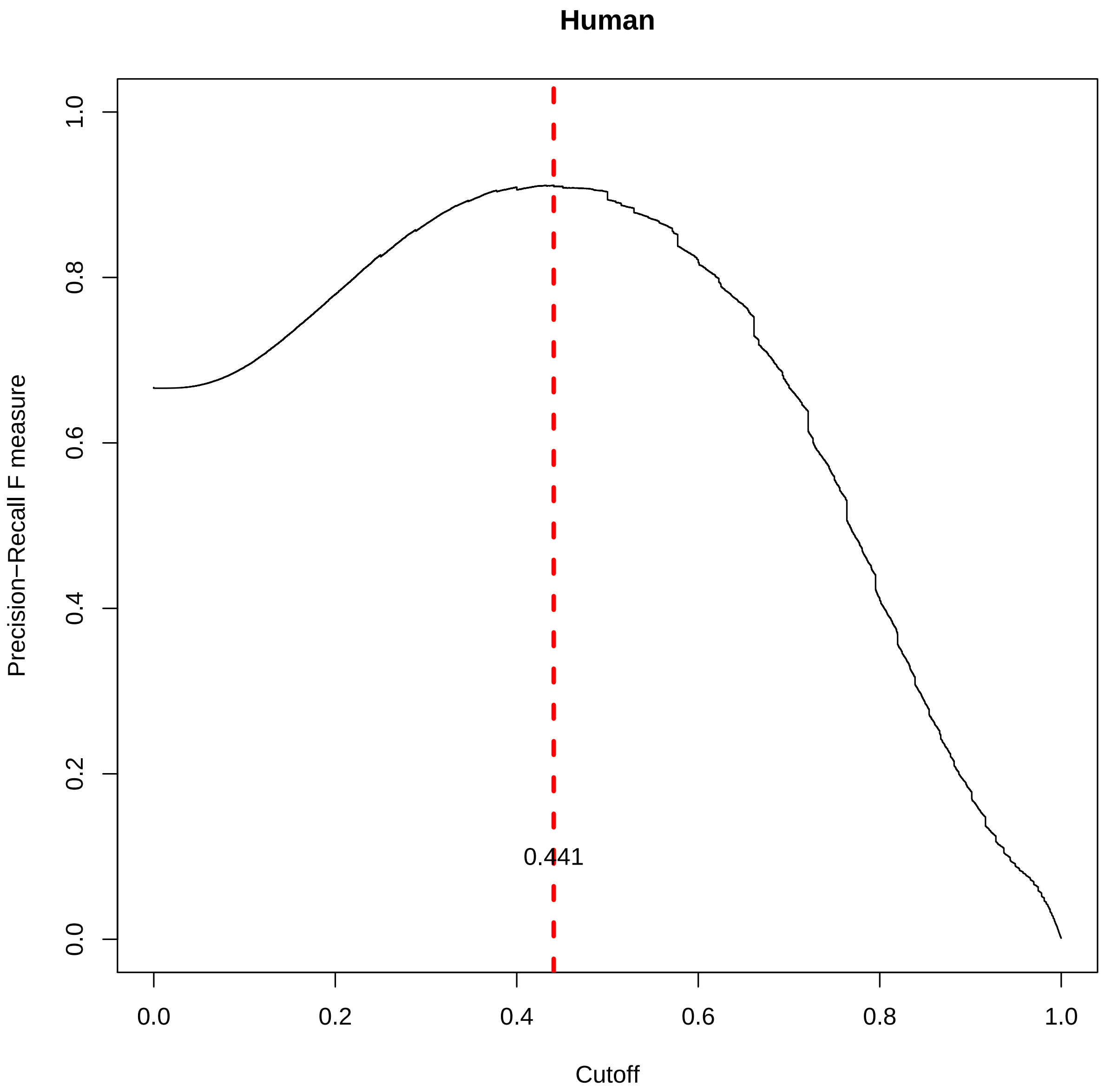


Figure 3.9
Learning the cutoff for phase scores for human datasets. The optimum cutoff for distinguishing actively translating regions from non-active translation was learned by maximizing the F1 score. The profiles from expressed CCDS exons in Ribo-seq data were treated as positives and corresponding profiles from RNA-seq were treated as negatives. Two datasets in human (SRA accession: SRP010679, SRP098789) were used for learning this cutoff.

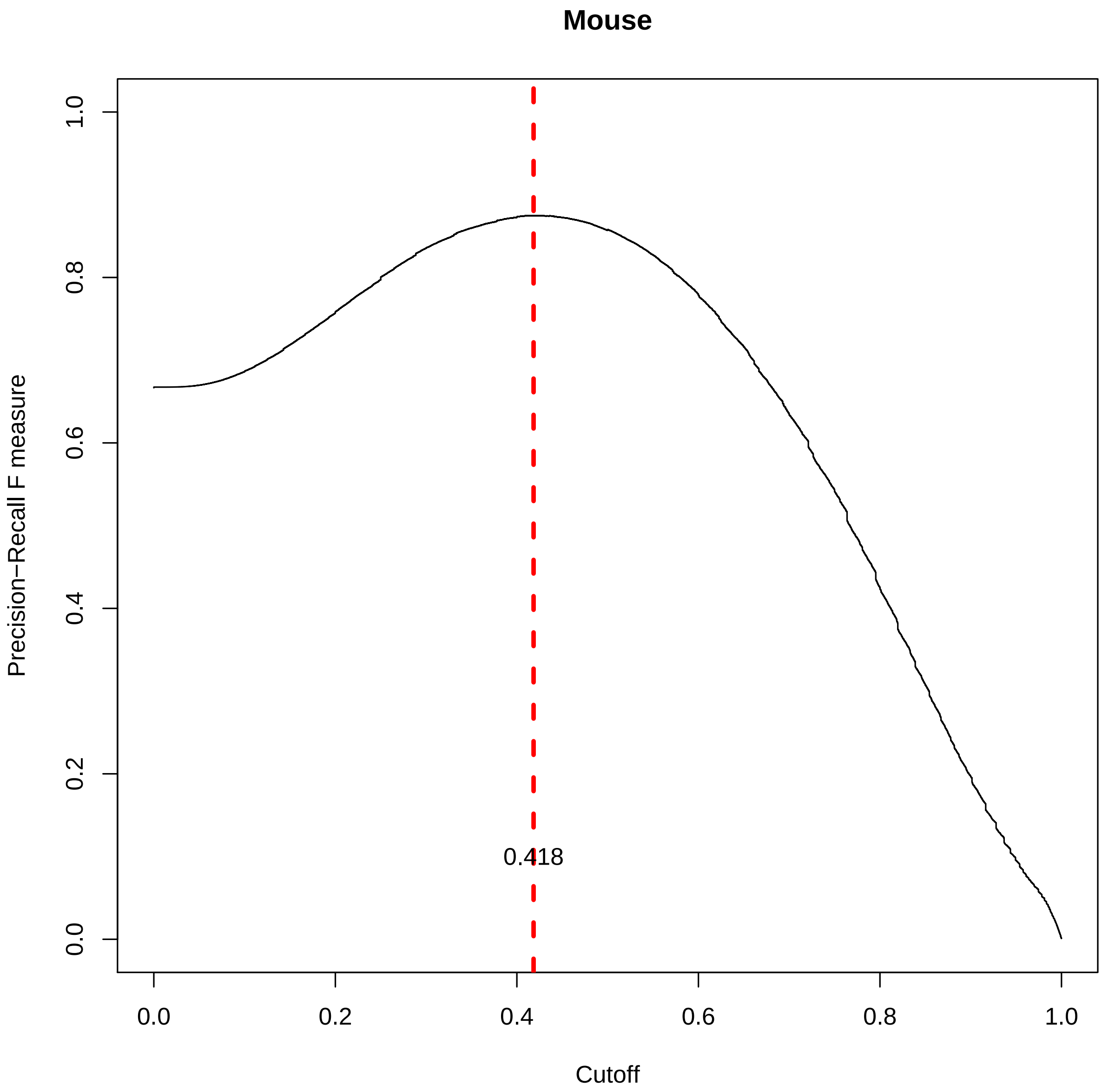


Figure 3.10
Learning the cutoff for phase scores for mouse datasets. The optimum cutoff for distinguishing actively translating regions from non-active translation was learned by maximizing the F1 score. The profiles from expressed CCDS exons in Ribo-seq data were treated as positives and corresponding profiles from RNA-seq were treated as negatives. Two datasets in mouse (SRA accession: SRP003554, and SRP115915)were used for learning this cutoff.

0.92, and 0.92, respectively (Figure 3.15A). Ribotricer also outperformed the other three methods consistently under the PR metric (Figure 3.15A). Ribotricer displayed the best performance on almost all the datasets at both ROC and PR metrics (Figures 3.35 and 3.36).

Next, we compared the performance of ribotricer, ORFscore, RiboTaper, and RiboCode by contrasting the number of true positives detected by each method while controlling the false positive rate at 0.1. We calibrated the cutoffs for each method so that the number of false positives reported by each method is 10% of the number of negatives. For human HEK293 cell dataset (SRA accession: SRP063852), ribotricer recovered $39,517$ truly translating exons, while RiboCode recovered $33,665$. RiboTaper, and ORFscore recovered $28,333$ and $26,486$ translating exons, respectively (Figure 3.15B). For mouse liver tissue dataset (SRA accession: SRP078005), ribotricer recovered $46,380$ truly translating exons, RiboCode recovered $43,332$, while RiboTaper and ORFscore recovered $35,746$ and $36,120$ translating exons, respectively (Figure 3.15B). We observed a similar performance for the other eight datasets where ribotricer consistently recovered more truly translating exons compared to the other three methods (Figures S12).

Longer ORFs have a higher chance of accumulating more ribosomes which leads to richer features when analyzing the associated RPF profiles. In order for ribotricer to be capable of detecting both short and long ORFs, the phase scores generated should be independent of the length of the ORF assessed. We investigated the effect of ORF length on the scores or the p-values generated by each method. The phase score generated by ribotricer is unaffected by the length of ORF while RiboCode, RiboTaper, and ORFscore generate a higher score or more significant P value as the ORF gets longer (Figure 3.14).

Finally, we compared the performance of ribotricer with other methods in terms of F1 score using the default cutoff for each method. Since we learned the cutoff for ribotricer from four

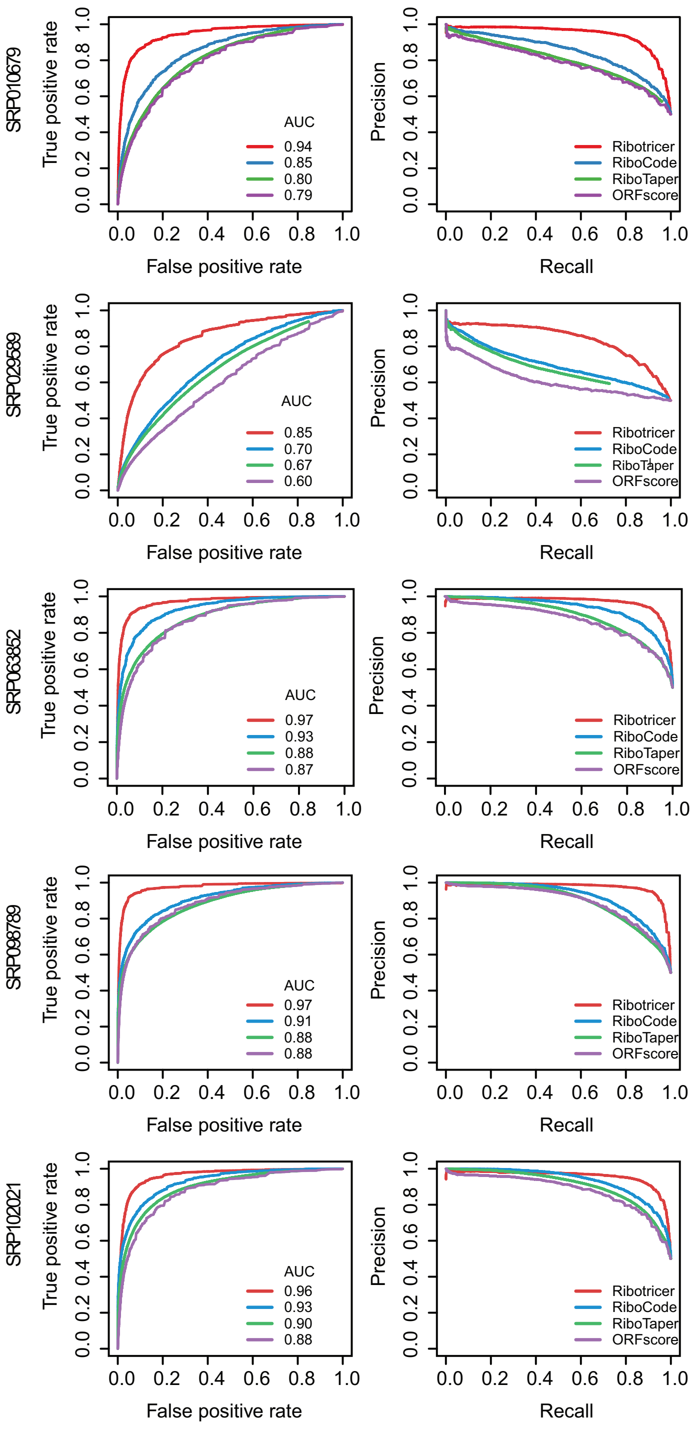


Figure 3.11

ROC plots and Precision-Recall plots for human datasets for exon level classification. Performance of ribotricer for detecting translating ORFs at exon level is compared with RiboCode, RiboTaper and ORFscore. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

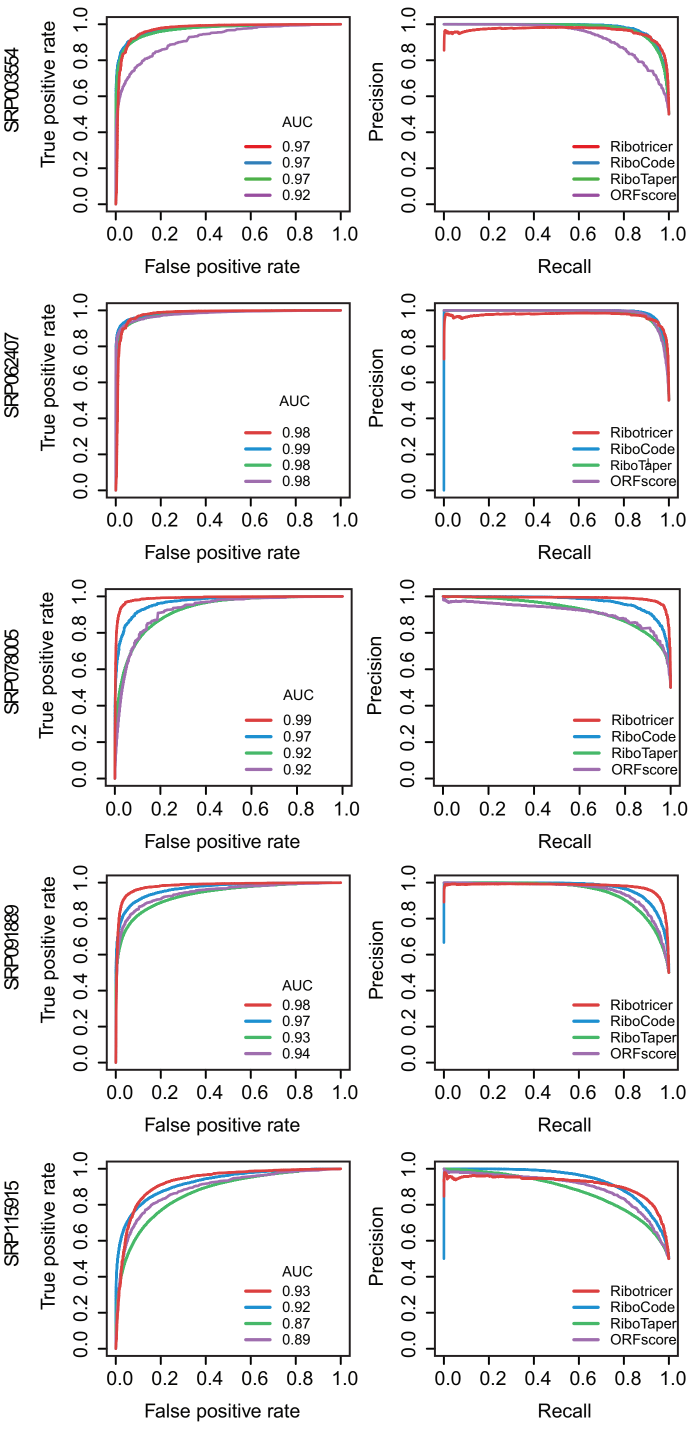


Figure 3.12

ROC plots and Precision-Recall plots for mouse datasets for exon level classification. Performance of ribotricer for detecting translating ORFs at exon level is compared with RiboCode, RiboTaper and ORFscore. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

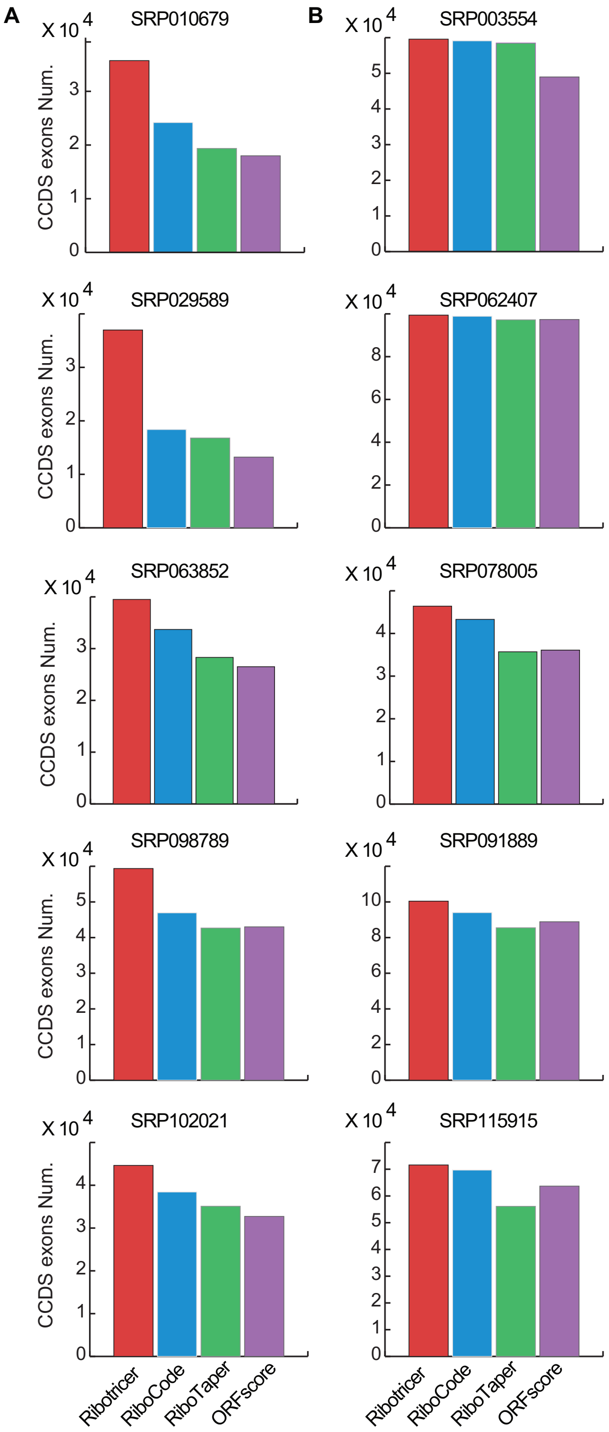


Figure 3.13
Number of translating exons recovered when controlling the false positive rate to be the same. Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore when the false positive rate is controlled to be 0.1. The number of truly translating exons are shown for both human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

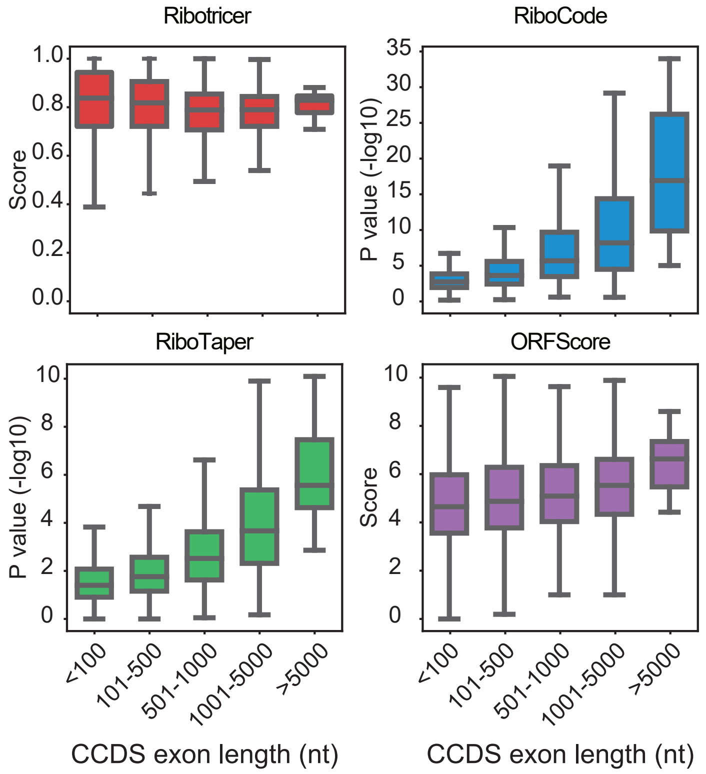


Figure 3.14

Effect of ORF length on output scores. Distribution of scores generated by ribotricer and ORFscore, and the P-values generated by RiboCode and RiboTaper over different CCDS exon lengths.

real datasets, we summarized the performance of ribotricer on the remaining six datasets that were not used to learn the empirical cutoff (Figure S14-S17). Notably, for human HeLa cell dataset (SRA accession: SRP029589) [25], all methods achieved relatively low F1 score with the best one to be 0.67 achieved by ribotricer. We checked the angle distribution of the 3D to 2D projection described earlier for this dataset (Figure 3.7), and found that it displays high noise level compared to other datasets analyzed, which indicates low data quality. Consequently, we excluded this dataset from further analysis. For the other two human datasets, ribotricer achieved an average F1 score of 0.91, and RiboCode achieved an average F1 score of 0.84. RiboTaper and ORFscore achieved an average F1 score of 0.73 and 0.12, respectively. For the three mouse datasets, ribotricer achieved an average F1 score of 0.93, and RiboCode achieved an average F1 score of 0.90. RiboTaper and ORFscore achieved an average F1 score of 0.85 and 0.55, respectively.

3.3.2 Ribotricer accurately detects translating ORFs at the transcript level

ORF-RATER [33], RibORF [34], Rp-Bp [36] and RiboWave [90] only detect translating ORFs at the full transcript level. To evaluate ribotricer against these methods we use a similar to the comparison strategy as used for exon-level benchmarking. For transcript level comparison, we first used the area under ROC/PR curves to assess the ability of different methods to distinguish Ribo-seq profiles from those from RNA-seq data. For human HEK293 cell dataset (SRA accession: SRP063852), ribotricer correctly distinguished Ribo-seq profiles from the simulated RNA-seq profiles with an AUROC of 1.0, while both Rp-Bp and RibORF achieved an AUROC of 0.96.

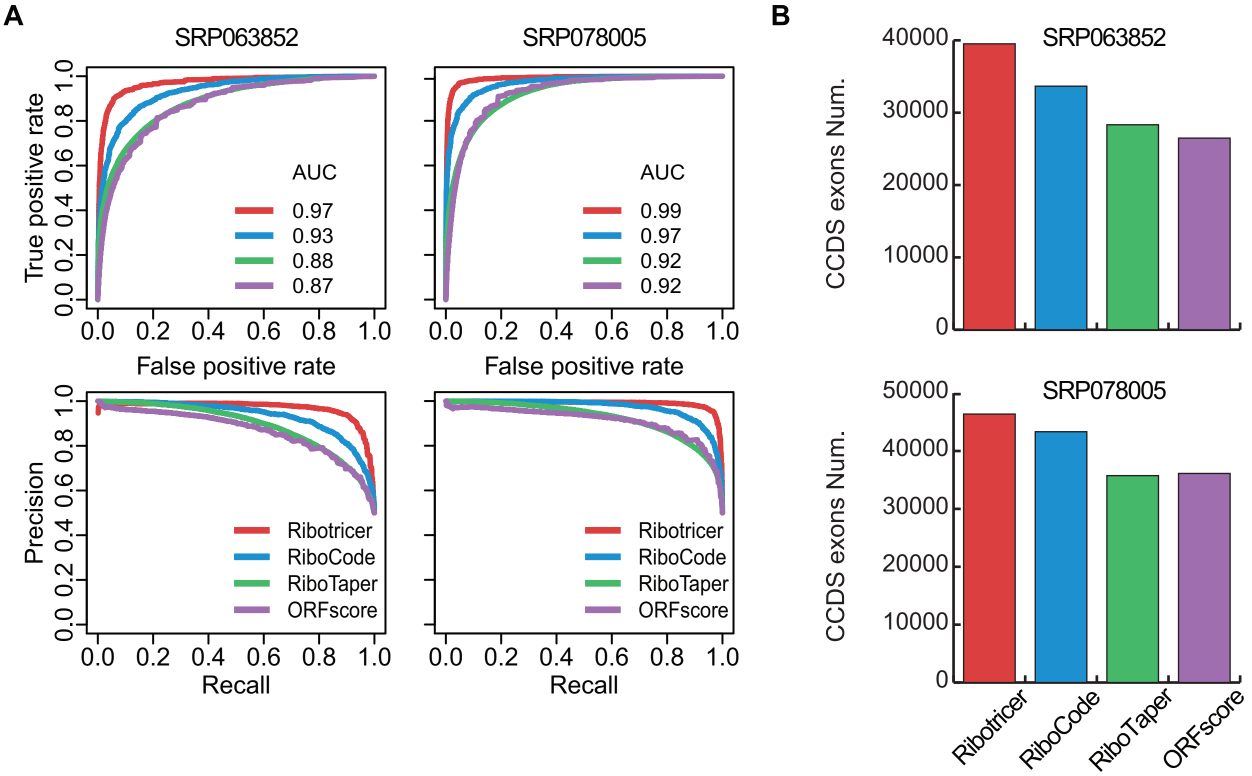


Figure 3.15

Comparison of performance on detecting translating exons. The performance of ribotricer is compared with that of RiboCode, RiboTaper, and ORFscore. (A) The ROC and precision recall curves summarizing performance of ribotricer, RiboCode, RiboTaper and ORFscore on one human and one mouse dataset. (B) The number of translating exons recovered when controlling the false positive rate to be the same.

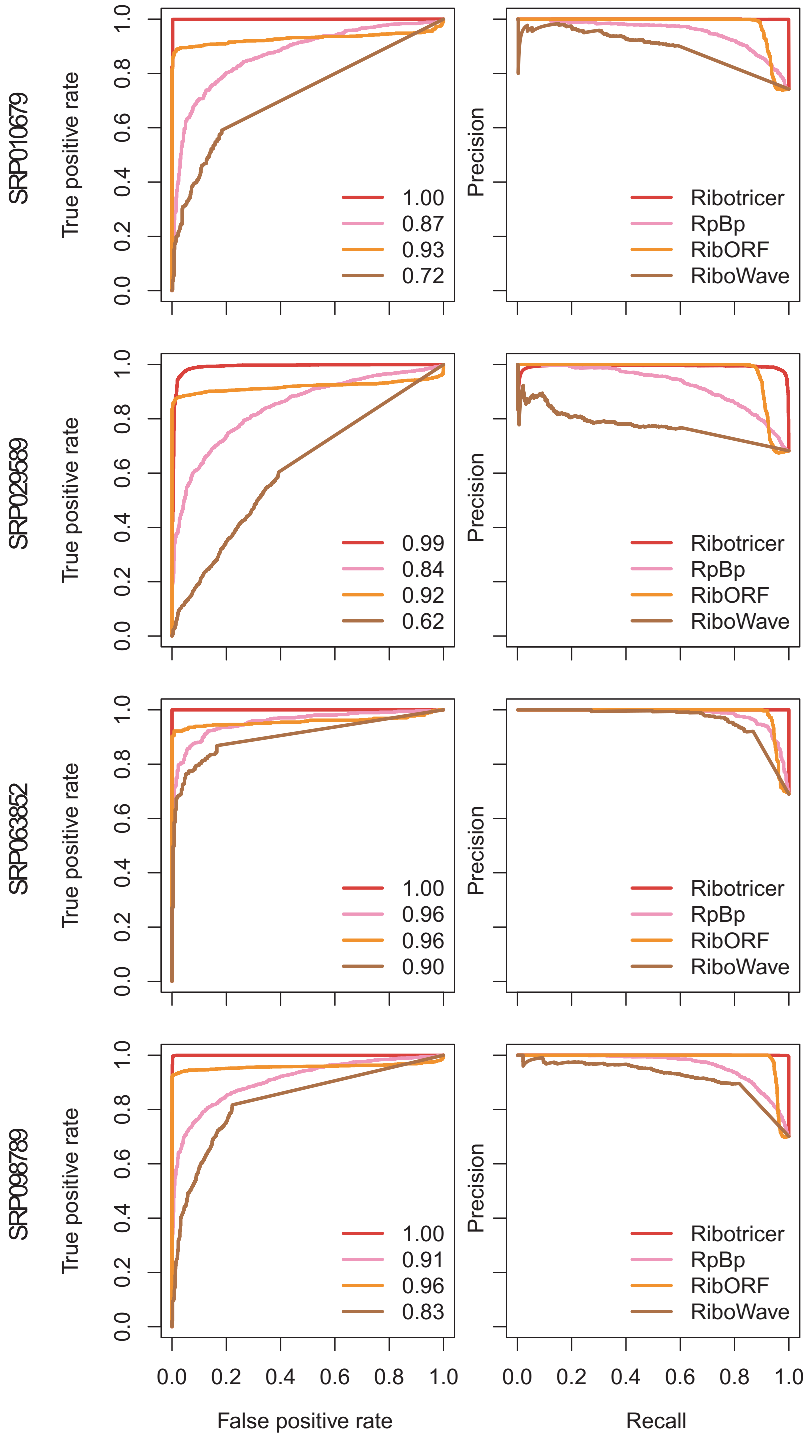


Figure 3.16

ROC plots and Precision-Recall plots on transcript level for human datasets. Performance of ribotricer for detecting translating ORFs at transcript level is compared with RpBp, ribORF and RiboWave. The profiles of expressed CCDS transcripts in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

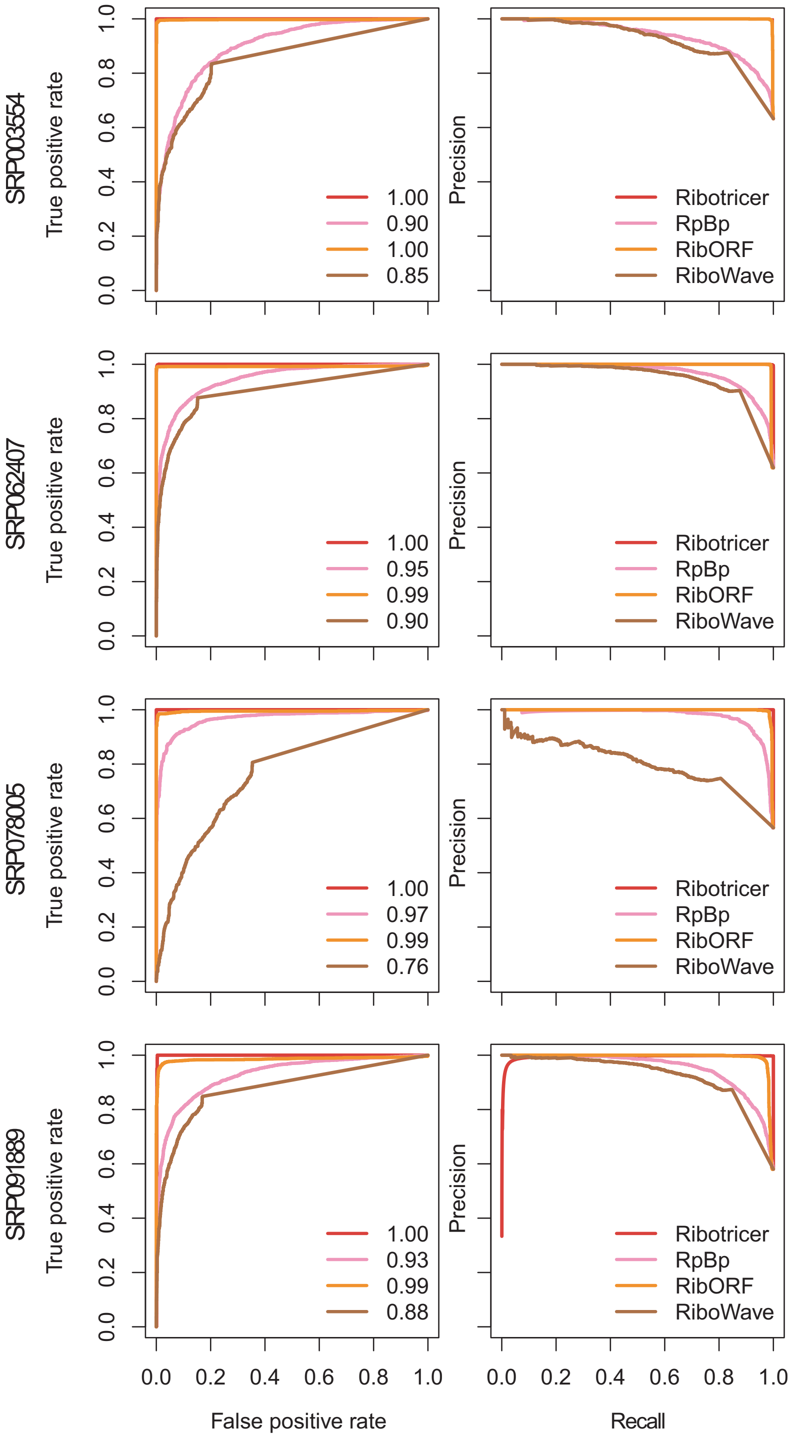


Figure 3.17
ROC plots and Precision-Recall plots on transcript level for mouse datasets. Performance of ribotricer for detecting translating ORFs at transcript level is compared with RpBp, ribORF and RiboWave. The profiles of expressed CCDS transcripts in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

RiboWave achieved an AUROC of 0.90 (Figure 3.26A). For human HeLa cell dataset (SRA accession: SRP098789) [1], ribotricer again perfectly distinguished Ribo-seq profiles from the simulated RNA-seq ones with an AUROC of 1.0, and Rp-Bp achieved an AUROC of 0.91. RibORF and RiboWave achieved an AUROC of 0.96 and 0.83, respectively (Figure 3.26A). Ribotricer also consistently outperformed other methods under the PR metric (Figure 3.26A). The complete results for all human and mouse samples can be found in Figure 3.16 and 3.17. It is worth mentioning that RibORF [34] uses a classification based method which trains its model by selecting one-third of the CDS profiles as true positives which might give it an extra advantage in this comparison. Notably, here we excluded ORF-RATER from the comparison because it always reports around half the number of detected ORFs compared with other methods, as noticed by Xiao *et al.* previously [87].

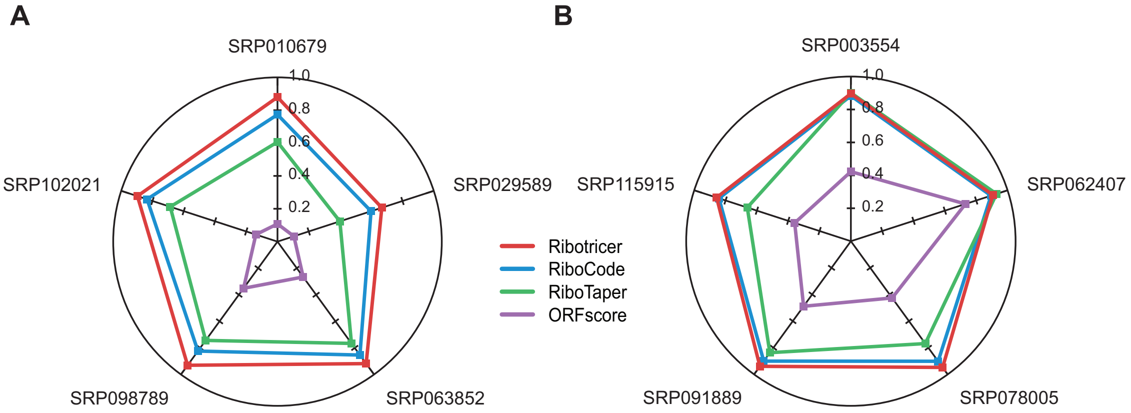


Figure 3.18
Comparison of F1 score (exon level) of ribotricer with RiboCode, RiboTaper, and ORFscore. Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore in terms of F1 score when the default threshold score is used for each tool. Results are shown for human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

Next, we compared the performances of different methods by checking the number of truly translating transcripts recovered when controlling the false positive rate to be the same as 0.1.

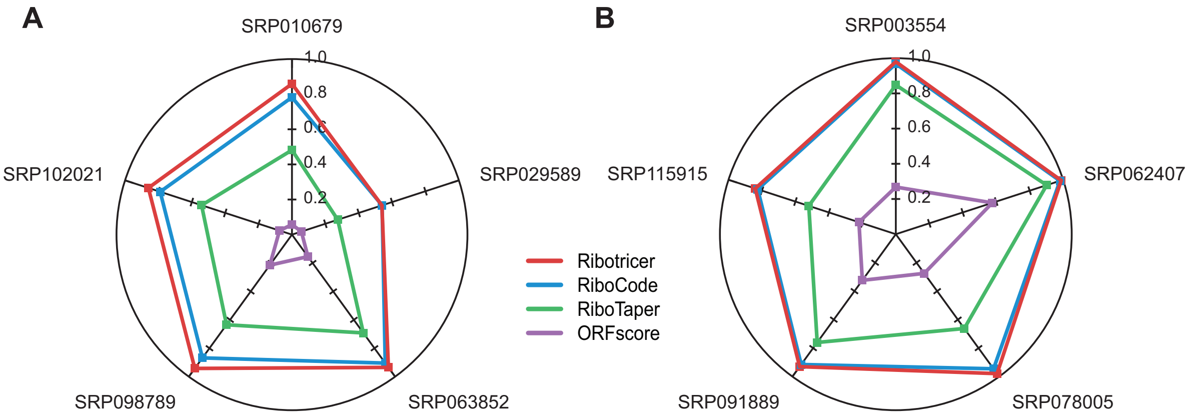


Figure 3.19
Comparison of sensitivity (exon level) of ribotricer with RiboCode, RiboTaper, and ORFscore. Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore in terms of sensitivity when the default threshold score is used for each tool. Results are shown for human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

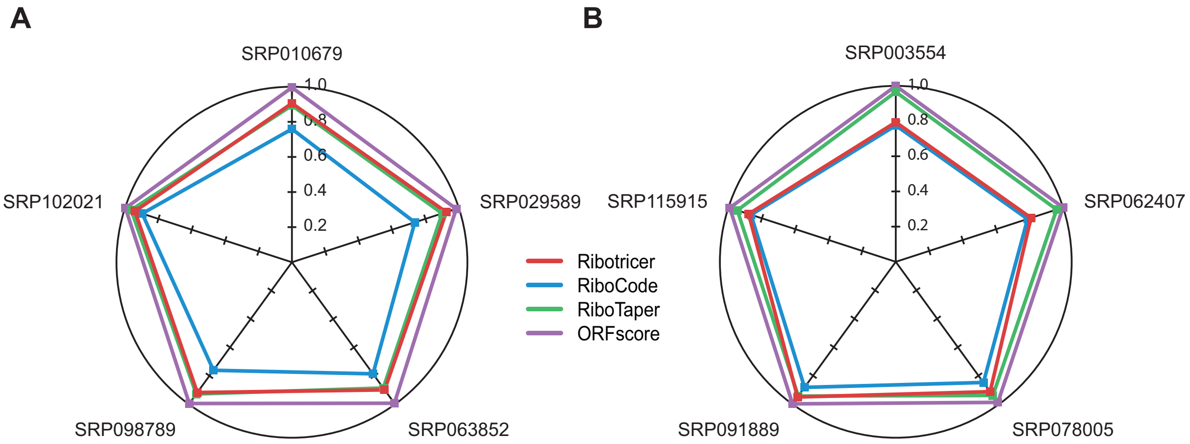


Figure 3.20
Comparison of specificity (exon level) of ribotricer with RiboCode, RiboTaper, and ORFscore. Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore in terms of specificity when the default threshold score is used for each tool. Results are shown for human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

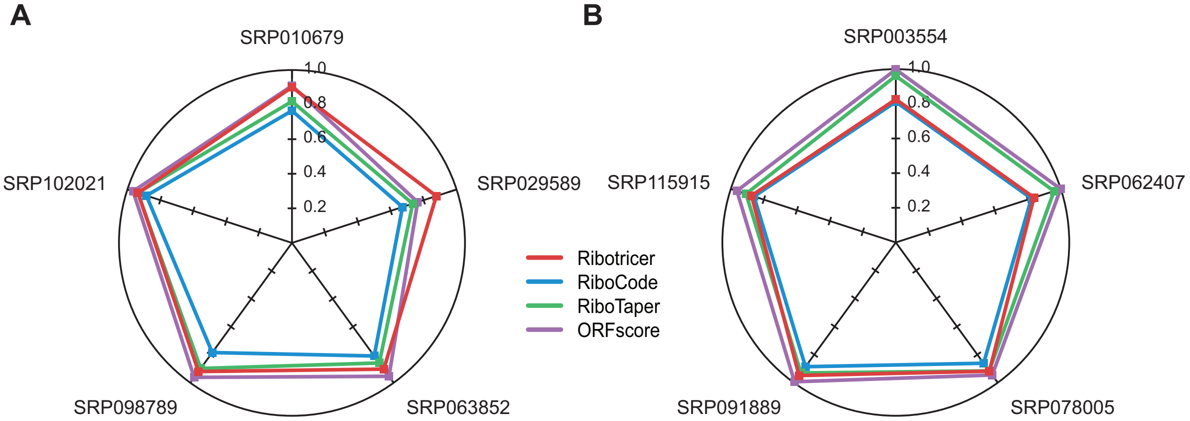


Figure 3.21
Comparison of precision (exon level) of ribotricer with RiboCode, RiboTaper, and ORFscore. Performance of ribotricer is compared with RiboCode, RiboTaper, and ORFscore in terms of precision when the default threshold score is used for each tool. Results are shown for human (A) and mouse (B) datasets. The profiles of expressed CCDS exons in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

For the human HEK293 cell dataset (SRA accession: SRP063852), ribotricer recovered 577 truly translating transcripts, while Rp-Bp, RibORF, and RiboWave recovered 508, 542, and 459 translating transcripts, respectively (Figure 3.26B). For the human HeLa cell dataset (SRA accession: SRP098789), ribotricer recovered $2,251$ truly translating transcripts, and Rp-Bp recovered $1,730$. RibORF and RiboWave recovered $2,130$ and $1,308$ truly translating transcripts, respectively (Figure 3.26B).

Finally, we used the F1 score to assess the performance of ribotricer in detecting actively-translating transcripts in comparison with other tools. For the two human samples, ribotricer achieved an average F1 score of 0.99, and Rp-Bp achieved an average F1 score of 0.89. RibORF and RiboWave achieved an average F1 score of 0.91 and 0.75, respectively. For the three mouse samples, ribotricer achieved an average F1 score of 0.99, and Rp-Bp achieved an average F1 score of 0.87. RibORF and RiboWave achieved an average F1 score of 0.97 and 0.69, respectively (Figure 3.23).

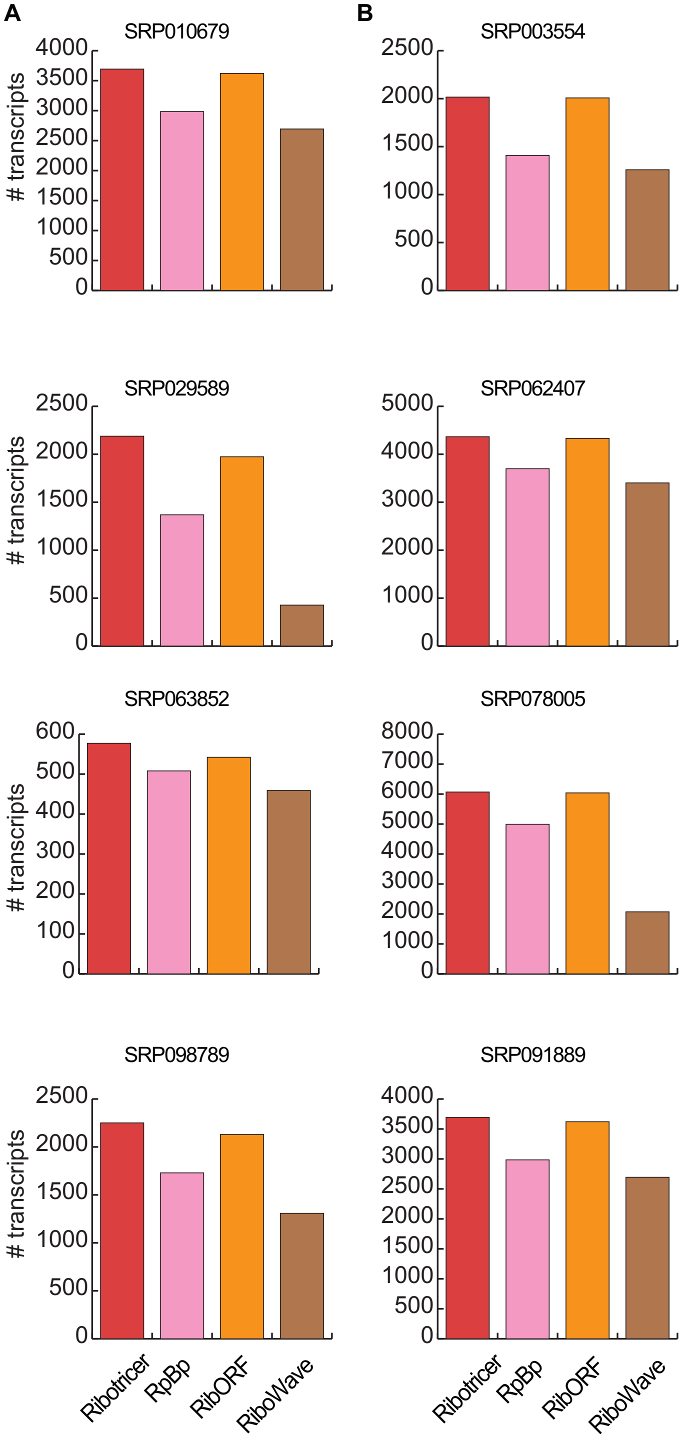


Figure 3.22

Number of translating transcripts recovered when controlling the false positive rate to be the same. Performance of ribotricer is compared with RpBp, ribORF, and RiboWave when the false positive rate is controlled to be 0.1. The number of truly translating transcripts are shown for both human (A) and mouse (B) datasets. The profiles of expressed CCDS transcripts in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

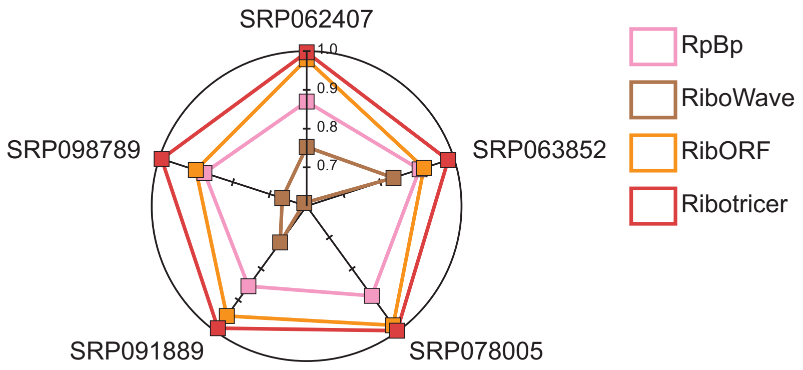


Figure 3.23
Performance of different methods on transcript level measured using F1 score. Performance of ribotricer is compared with RpBp, ribORF, and RiboWave in terms of F1 score when the default threshold score is used for each tool. The profiles of expressed CCDS transcripts in Ribo-seq data were treated as true positive and the corresponding RNA-seq profile as true negative.

AUROC, F1 scores, and p-values for AUROC difference were calculated using the `pROC` [135] package in R. For calculating p-values, we used the `bootstrap` method and set `alternative=`greater'`.

3.4 Ribotricer's phase score remains stable on truncated ORFs

In order to test the ability of ribotricer to correctly predict the translation status of an ORF whose length has been shortened due to truncation we performed a simulation where for all candidate ORFs which have atleast 50% of non-empty codons, *i.e.* codons with non-zero reads, we truncated it from 3' end such that the truncated length was $10-100\%$ of the original length. For each such truncated ORF, we calculated ribotricer's phase score and compared it with the corresponding RiboCode generated p-value. It is worth mentioning, that among the tools of

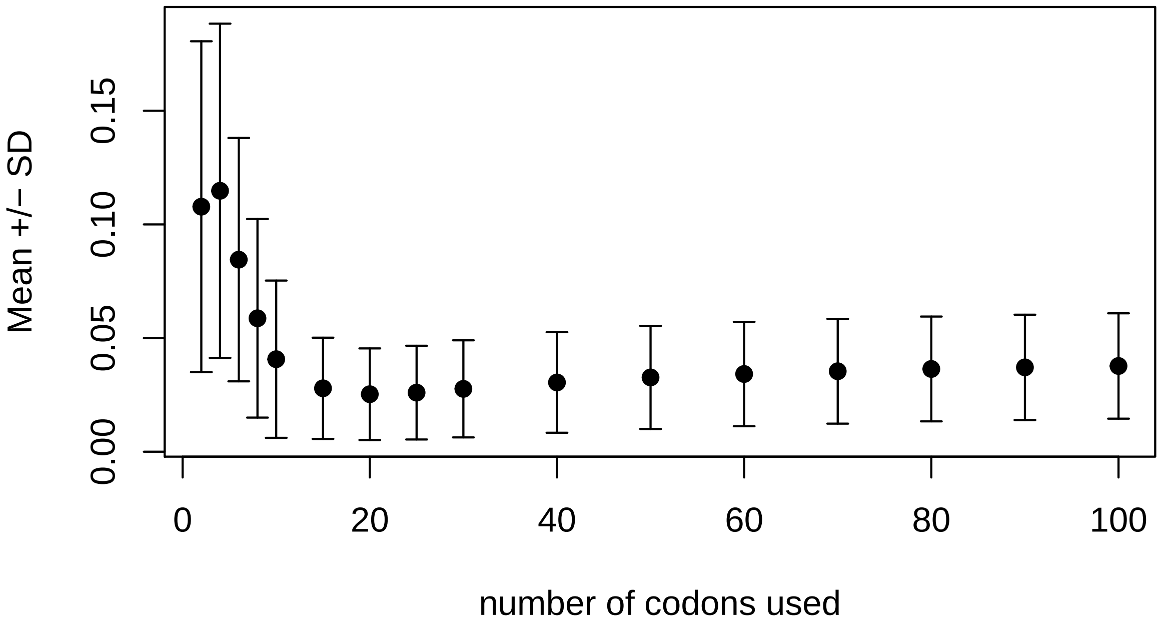


Figure 3.24
Effect of number of codons on ribotricer's phase score in human dataset. Mean absolute difference and standard deviation between original phase score using all codons and the one with down-sampled number of codons. The plot was generated on human dataset (SRA accession: SRP063852) using 5K genes with at least 50% valid codons, the down-sampling is repeated 100 times for each gene. Similar trend is observed for other human datasets.

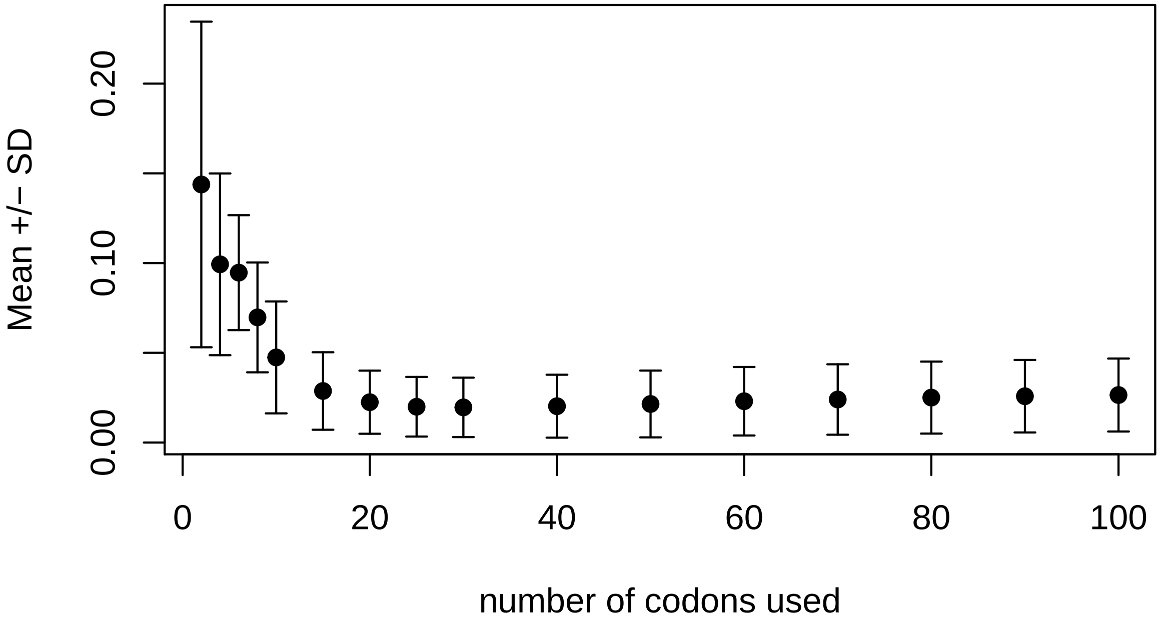


Figure 3.25
Effect of number of codons on ribotricer's phase score in mouse dataset. Mean absolute difference and standard deviation between original phase score using all codons and the one with down-sampled number of codons. The plot was generated on mouse dataset (SRA accession: SRP003554) using 5K genes with at least 50% valid codons, the down-sampling is repeated 100 times for each gene. Similar trend is observed for other mouse datasets.

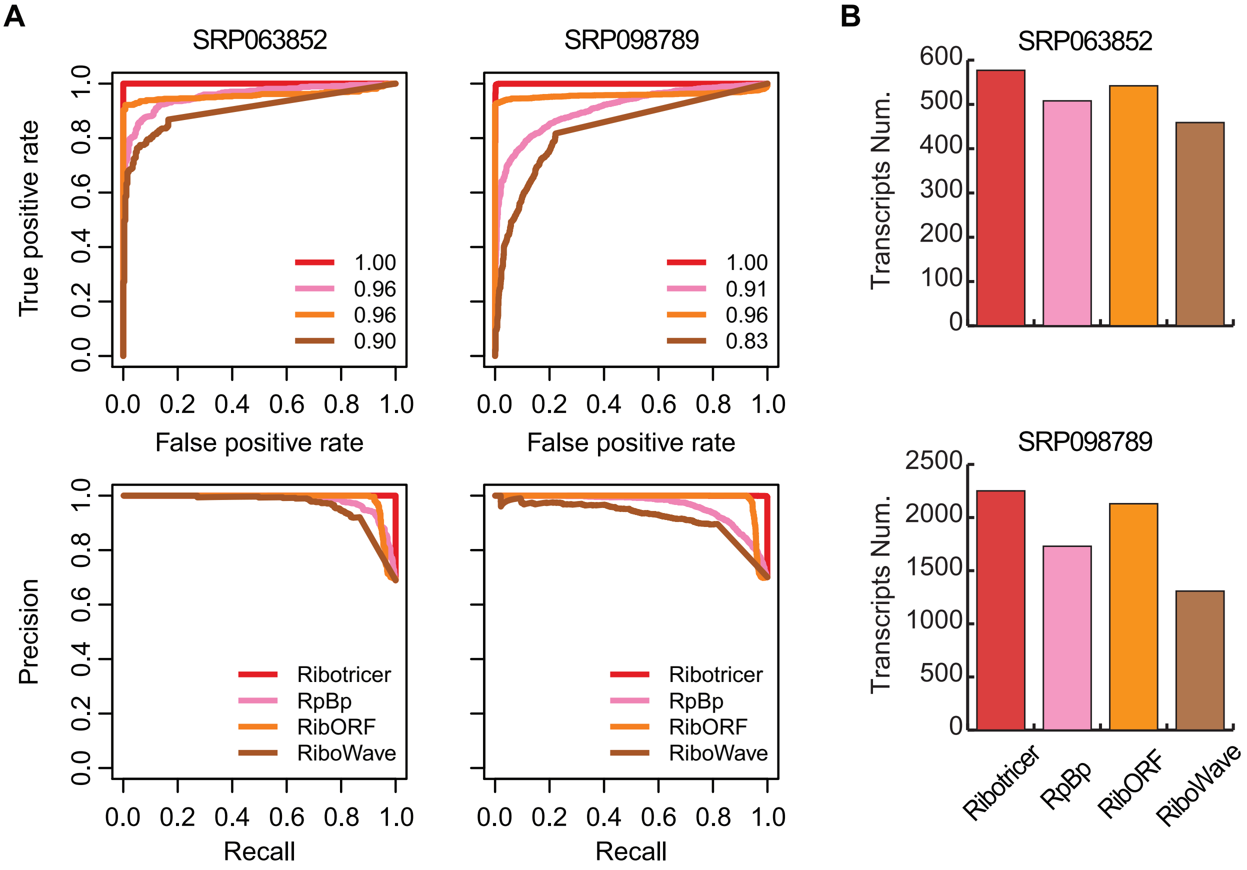


Figure 3.26

Comparison of performance on detecting translating transcripts.The performance of ribotricer is compared with that of RibORF, RiboWave, and Rp-Bp. (A) The ROC and precision recall curves summarizing performance of ribotricer, RibORF, RiboWave and Rp-Bp on one human and one mouse dataset. (B) The number of translating transcripts recovered when controlling the false positive rate to be the same.

capable of performing exon level classification, we were able to benchmark ribotricer against only RiboCode and ORFscore as RiboTaper requires bam files of both RNA-seq and Ribo-seq samples.

Ribotricer's score for the truncated ORF is negligibly different from the original ORF with a maximum difference of ± 0.05 (Figure 3.27 and 3.28) as demonstrated using a human (SRA accession: SRP063852) and a mouse dataset (SRA accession: SRP003554). On the other hand, the RiboCode generated p-values show a clear dependence on the ORF length with the deviation from original score being as high as $\pm$ 100. It is worth mentioning that the differences between truncated and original profile for RiboCode are calculated on a $\log_{10}$ scale as it outputs p-values, while for both ribotricer and ORFscore, the differences are calculated on the same scale as the scores.

3.4.1 Ribotricer can detect ORFs as short as 20 codons

In order to determine the minimum length of ORF that can be detected by ribotricer we performed a simulation using the Ribo-seq profiles of genes with total codons > 100 and with at least 50% non-empty codons. We then randomly sampled $10 - 100$ codons, without maintaining their order explicitly, and generated a "downsampled" profile. The mean absolute difference between the original phase score calculated using the full length profile versus the "downsampled" profile with 20 or more codons is smaller than 0.05 and does not change after increasing the number of codons (Figures 3.24 and 3.25).

3.4.2 Learning species-specific cutoffs

Ribo-seq's protocol was initially developed to profile the translational landscape in yeast [24], but it has been widely used to profile the translational status of ORFs in multiple species [136, 137]. We benchmarked ribotricer first using human and mouse datasets where we have access to CCDS annotated regions as a high confidence ground truth for known protein coding status (Figures 3.13-3.23). In order to further benchmark ribotricer against other methods, we used additional public Ribo-seq datasets from *Arabidopsis*, *C. elegans*, *Drosophila*, rat, yeast, and zebrafish (Table 3.4). Unlike human and mouse, CCDS annotations are not available for these species. Hence, for such species, we considered the Ribo-seq profile of annotated CDS regions as

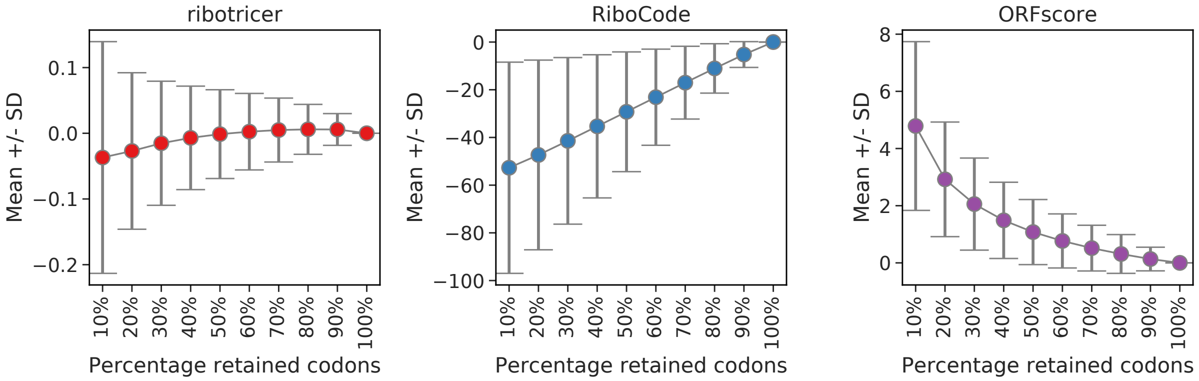


Figure 3.27
Effect of truncating an ORF on ribotricer's phase score, RiboCode's p-values and ORFscore in human dataset. Mean difference and standard deviation between original phase score using full length ORF and the ones after truncating it from the 3' end. The plot was generated on human dataset (SRA accession: SRP063852) using 5K genes with at least 50% valid codons and truncating it to have indicated percentage (X-axis) of codons.The differences between truncated and original profile for RiboCode are calculated on a $\log_{10}$ scale as it outputs p-values, while for both ribotricer and ORFscore, the differences are calculated on the same scale as the scores.

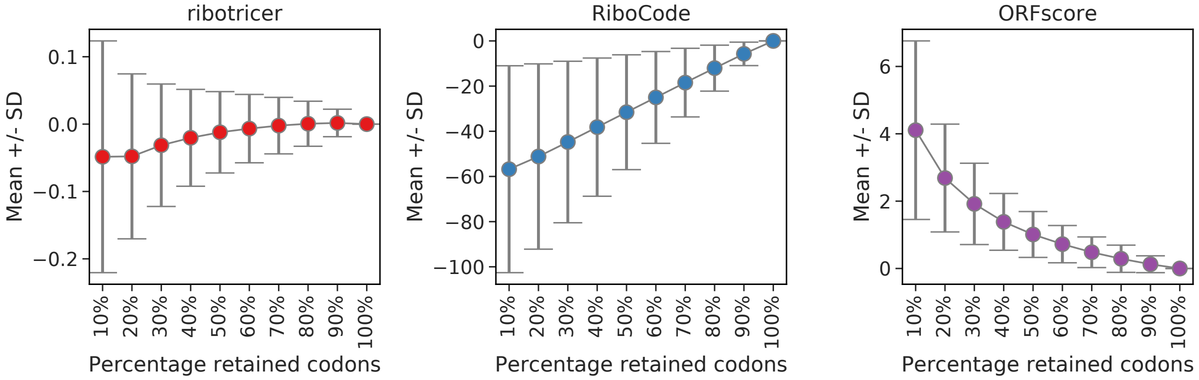


Figure 3.28
Effect of truncating an ORF on ribotricer's phase score, RiboCode's p-values and ORFscore in mouse dataset. Mean difference and standard deviation between original phase score using full length ORF and the ones after truncating it from the 3' end. The plot was generated on mouse dataset (SRA accession: SRP003554) using 5K genes with at least 50% valid codons and truncating it to have indicated percentage (X-axis) of codons. The differences between truncated and original profile for RiboCode are calculated on a $\log_{10}$ scale as it outputs p-values, while for both ribotricer and ORFscore, the differences are calculated on the same scale as the scores.

the true positive and the corresponding RNA-seq profile as the true negative. In order to establish if we needed to re-adjust our phase score cutoff for each species separately, we summarized the phase scores for both Ribo-seq and RNA-seq samples from multiple public datasets (Figure 3.33). We observed that phase scores of both RNA-seq and Ribo-seq samples vary across species (Figures 3.30, 3.31, and 3.32) with higher variation arising from the Ribo-seq samples. The variation in phase scores for RNA-seq samples in the same species is limited, though it also exhibits a species related trend (Figure 3.32). Ribo-seq samples on the other hand exhibit higher intra-species and across-species heterogeneity. Hence, in order to capture this species-specific differences in RNA-seq and Ribo-seq scores, we learned cutoffs for each species separately (Table 3.5 and 3.6; Figure 3.34). It is worth noting that, human and mouse samples that we previously used for our benchmark exhibit similar variation in RNA-seq and Ribo-seq phase scores besides having higher Ribo-seq phase scores as compared to all other species. On the other hand, the difference between Ribo- and RNA-seq phase scores appears to be particularly low in *Drosophila* datasets (Figure 3.32).

3.4.3 Learning dataset-specific cutoffs

In studies where both Ribo-seq and RNA-seq experiment are available, it is possible to fine-tune the phase-score cutoff to be dataset-specific. The Ribo-seq and RNA-seq samples within the same species can show variation in terms of their phase score (Figure 3.32) and hence, it is possible that learning dataset-specific cutoffs leads to an overall better performance (Figures 3.38-3.42). To learn the dataset-specific cutoffs, we calculated the median difference between phase scores of Ribo-seq and RNA-seq profiles for each dataset over only protein-coding regions. Using a

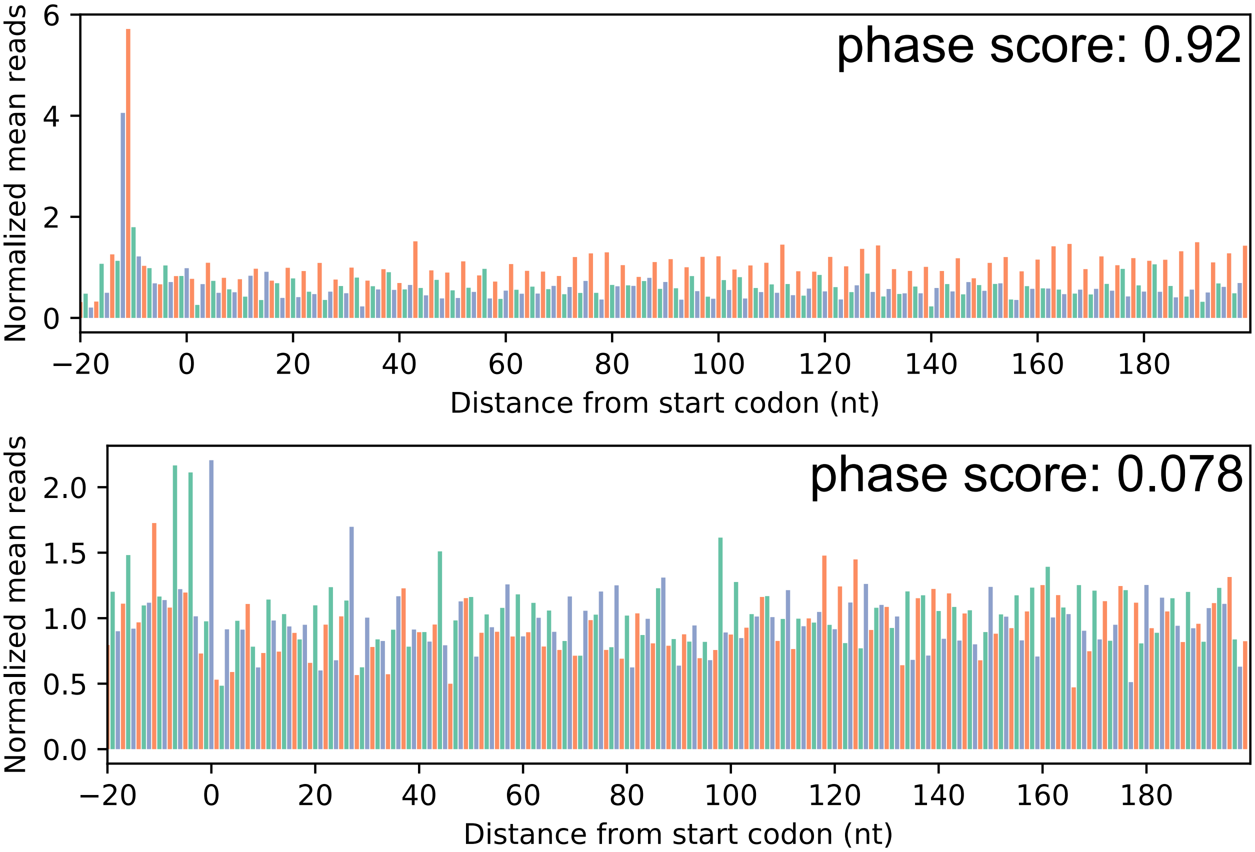


Figure 3.29

Example of phase scores for an active and a non-active ORF. Phase score generated by ribotricer for two different profiles.

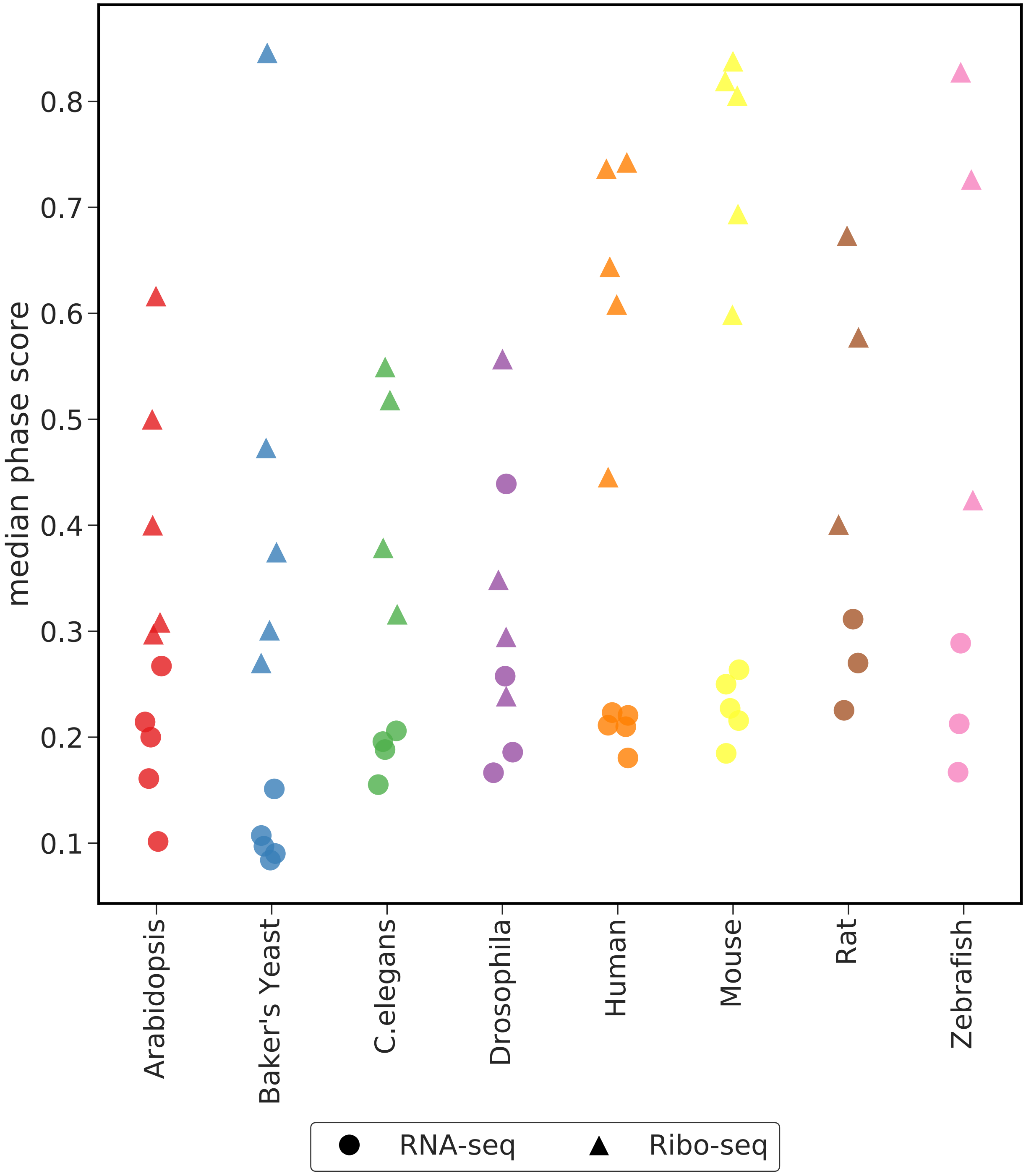


Figure 3.30
Summarized median phase score for RNA-seq and Ribo-seq for all datasets. For each dataset, the median phase score was calculated for all the candidate ORFs for both Ribo-seq and the corresponding RNA-seq sample.

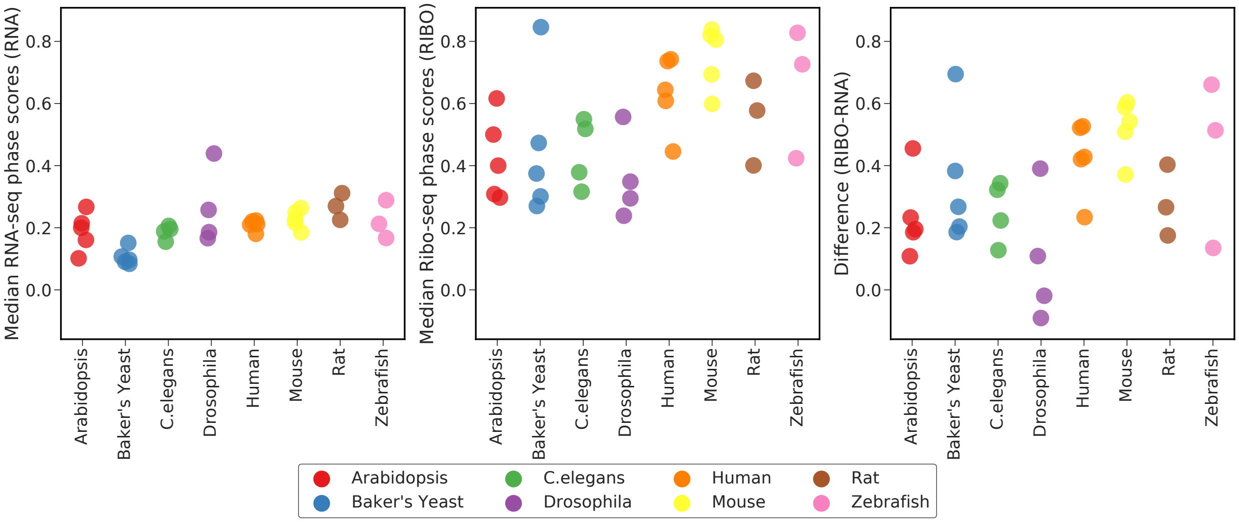


Figure 3.31
Median phase score for RNA-seq and Ribo-seq and their differences across multiple species. For each dataset, the median phase score was calculated for all the candidate ORFs for both Ribo-seq and the corresponding RNA-seq sample. Same as Figure 3.30 except that the RNA- and Ribo-seq samples have been separated into individual panels.

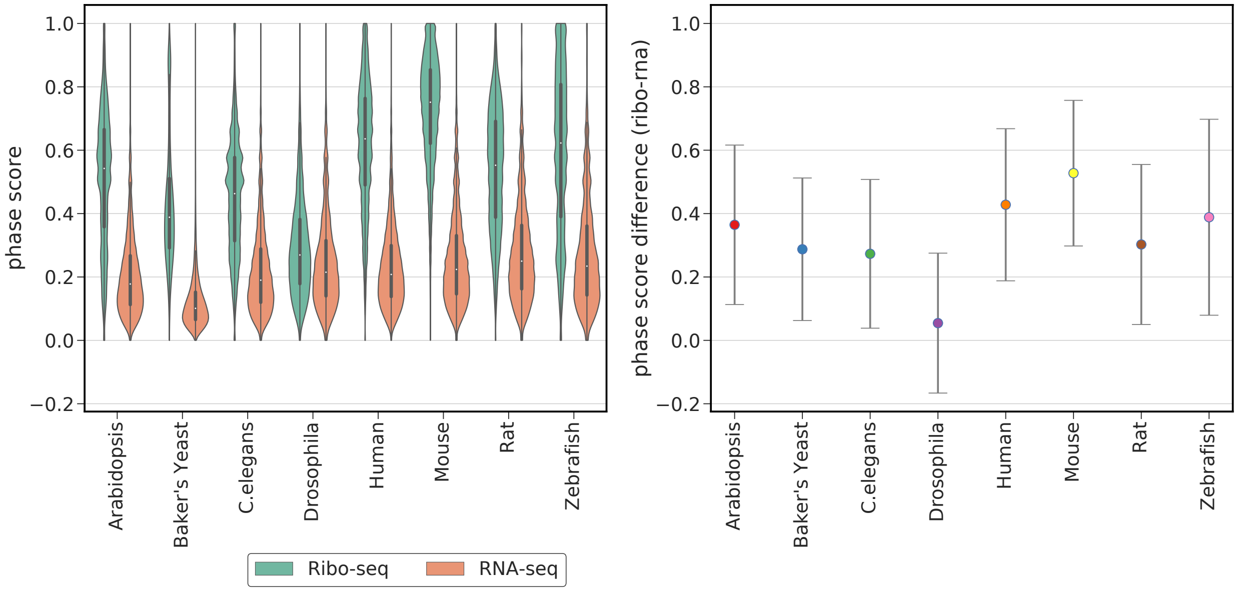


Figure 3.32
Distribution of median phase scores for RNA-seq and Ribo-seq samples and their differences across multiple species. For each species, medians were calculated on the collection of merged datasets for that species.

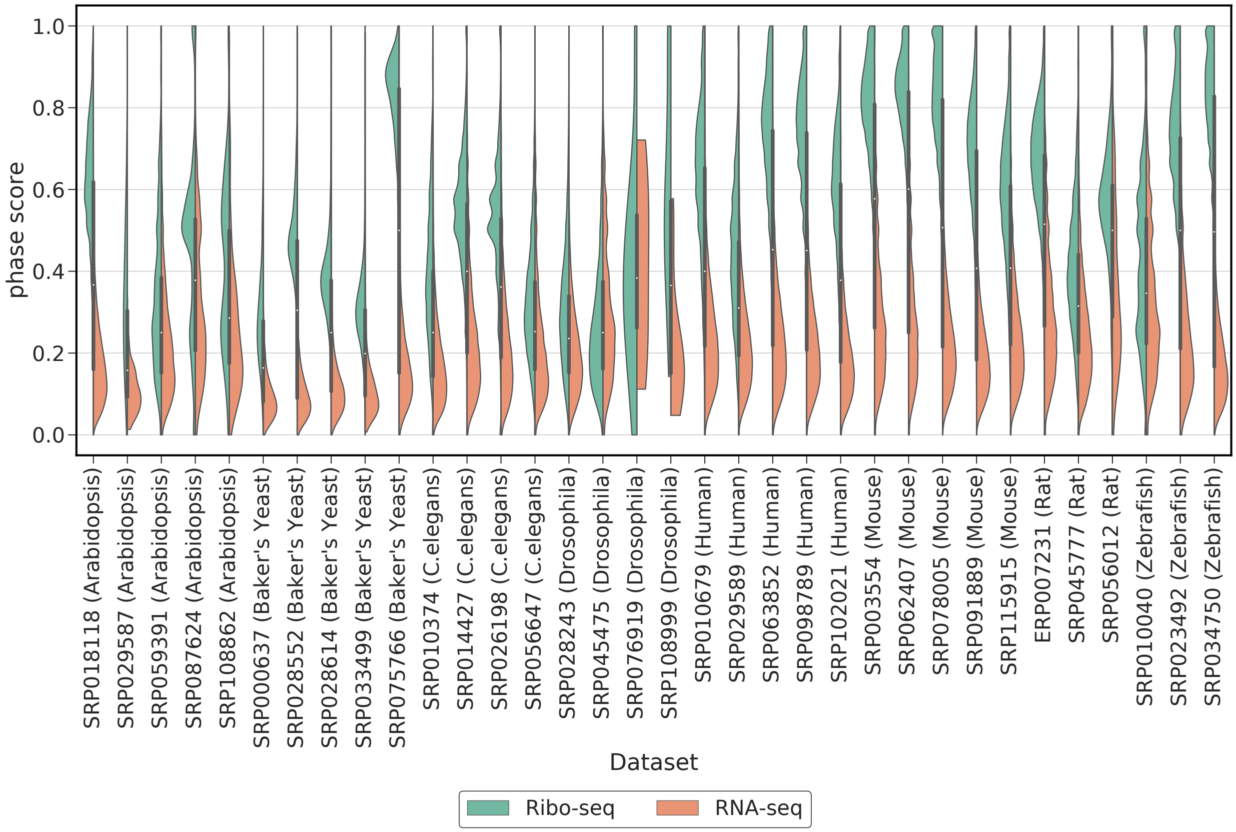


Figure 3.33
Distribution of individual RNA-seq and Ribo-seq samples' phase scores across species. For each dataset phase scores were calculated for all candidate ORFs. For human and mouse, Ribo-seq CCDS profiles were treated as true positive and the corresponding RNA-seq profile was treated as true negative. For all other species Ribo-seq profile of annotated CDS regions were treated as true positive and the corresponding RNA-seq profile treated as true negative.

sampling strategy where a one-third fraction of protein-coding profiles were used to determine the median difference between Ribo-seq and RNA-seq profiles with replacement ($n_{\text{bootstraps}} = 10000$) [102], the dataset-specific cutoff was assigned to be the median of these differences. It is worth mentioning that this approach is only viable for studies where both Ribo-seq and RNA-seq samples are available. The dataset-specific cutoffs result in ribotricer achieving higher F1 scores in some but not all datasets (Tables 3.9-3.11; Figures 3.38-3.42). In all our datasets, a median difference of 0.25 or more between Ribo-seq and RNA-seq protein-coding profiles results in an F1 score greater than 0.73 (Figure 3.41). Given a set of Ribo-seq and RNA-seq mapped files (BAM), the dataset-specific cutoffs can be determined by using `ribotricer learn-cutoff`.

Table 3.6
Species specific recommended phase score cutoffs for ribotricer. A "#" indicates the cutoff for the species is taken to be the median phase score difference between CDS annotated Ribo-seq and RNA-seq profiles since they only had one dataset each.

Species	Cutoff
Arabidopsis	0.330
Baker's Yeast	0.318
C. elegans	0.249
Drosophila	0.181
Human	0.440
Mouse	0.418
Rat	0.453
Zebrafish	0.249
C. albicans#	0.228
S. pombe#	0.409
Chimpanzee#	0.334
Macaque#	0.321

3.4.4 Running ribotricer on a new species

We provide a list of recommended phase score cutoffs (Table 3.6) for most species where there are at least three or more public Ribo-seq datasets (Table 3.4). The cutoffs for each species were

Table 3.7

Species wise mean, median and standard deviation of difference of Ribo-seq and RNA-seq phase scores. SD = Standard Deviation. A "#" indicates that the median phase score difference for these species is also considered as cutoff for ribotricer, since they only had one dataset each.

species	number of samples	mean difference phase score	median difference phase score	SD
Arabidopsis	5	0.308	0.365	0.252
Baker's Yeast	5	0.309	0.287	0.225
C.elegans	4	0.232	0.273	0.235
Drosophila	4	0.048	0.054	0.221
Human	5	0.385	0.428	0.240
Mouse	5	0.468	0.528	0.230
Rat	3	0.260	0.303	0.253
Zebrafish	3	0.325	0.388	0.309
C. albicans#	1	0.228	0.225	0.151
S. pombe#	1	0.380	0.409	0.176
Chimpanzee#	1	0.328	0.334	0.233
Macaque#	1	0.285	0.321	0.218

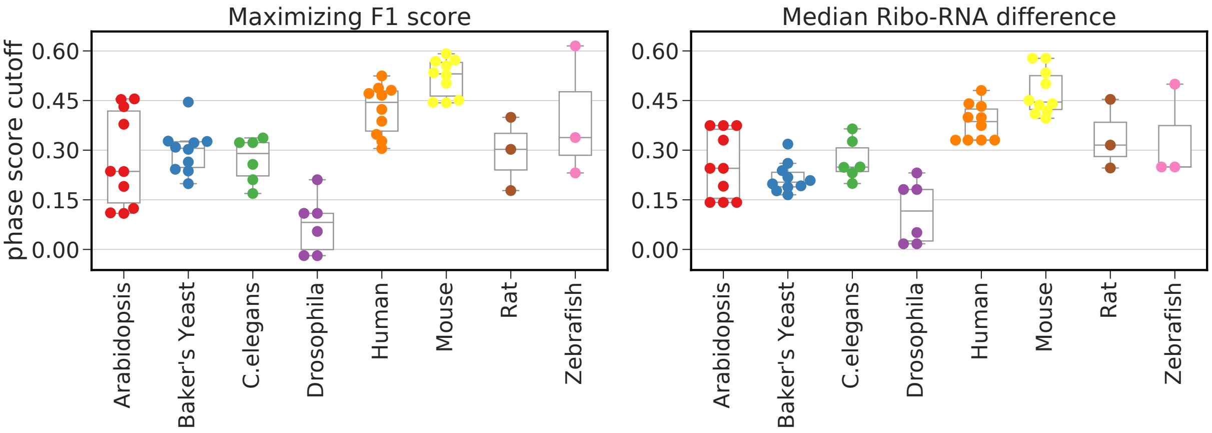


Figure 3.34

Distribution of median difference between Ribo-seq and RNA-seq sample as determined using only two datasets per species. For each species all possible combinations of two datasets were chosen and median difference between phase scores of Ribo-seq and RNA-seq determined.

Table 3.8

Best and second to best performing methods at AUROC metric for each dataset. The p-values were calculated using pROC [135] package using bootstrap method and alternative=`greater'. AUROC (B) and AUROC (SB) denotes area under ROC for the best and the second to best methods respectively. A * indicates the dataset was later used to learn the ribotricer cutoffs by maximizing the F1 score. The AUROC values however do not depend on any cutoff.

SRP	Species	Best (B)	Second Best (SB)	AUROC (B)	AUROC (SB)	p-value
SRP018118*	Arabidopsis	ribotricer	RiboCode	0.982	0.923	$< 2.2 \times 10^{-16}$
SRP029587	Arabidopsis	ribotricer	RiboCode	0.897	0.594	$< 2.2 \times 10^{-16}$
SRP059391*	Arabidopsis	ribotricer	ORFscore	0.690	0.632	$< 2.2 \times 10^{-16}$
SRP087624	Arabidopsis	ribotricer	RiboTaper	0.697	0.523	$< 2.2 \times 10^{-16}$
SRP108862	Arabidopsis	ribotricer	RiboCode	0.732	0.607	$< 2.2 \times 10^{-16}$
SRP000637	Baker's Yeast	ribotricer	RiboCode	0.921	0.837	$< 2.2 \times 10^{-16}$
SRP028552*	Baker's Yeast	ribotricer	RiboCode	0.986	0.951	$< 2.2 \times 10^{-16}$
SRP028614	Baker's Yeast	ribotricer	RiboCode	0.966	0.846	$< 2.2 \times 10^{-16}$
SRP033499	Baker's Yeast	ribotricer	RiboCode	0.947	0.783	$< 2.2 \times 10^{-16}$
SRP075766*	Baker's Yeast	ribotricer	RiboCode	0.996	0.962	$< 2.2 \times 10^{-16}$
SRP010374*	*C. elegans*	ribotricer	RiboCode	0.867	0.776	$< 2.2 \times 10^{-16}$
SRP014427	*C. elegans*	ORFscore	ribotricer	0.927	0.920	3.774×10^{-14}
SRP026198*	*C. elegans*	ORFscore	ribotricer	0.956	0.908	$< 2.2 \times 10^{-16}$
SRP056647	*C. elegans*	RiboCode	RiboTaper	0.745	0.745	0.247
SRP028243*	Drosophila	ribotricer	RiboCode	0.725	0.587	$< 2.2 \times 10^{-16}$
SRP045475	Drosophila	ORFscore	RiboTaper	0.633	0.522	$< 2.2 \times 10^{-16}$
SRP076919	Drosophila	ORFscore	ribotricer	0.638	0.465	0.317
SRP108999*	Drosophila	ribotricer	RiboTaper	0.884	0.727	0.068
SRP010679*	Human	ribotricer	RiboCode	0.944	0.849	$< 2.2 \times 10^{-16}$
SRP029589	Human	ribotricer	RiboCode	0.846	0.701	$< 2.2 \times 10^{-16}$
SRP063852	Human	ribotricer	RiboCode	0.969	0.930	$< 2.2 \times 10^{-16}$
SRP098789*	Human	ribotricer	RiboCode	0.975	0.908	$< 2.2 \times 10^{-16}$
SRP102021	Human	ribotricer	RiboCode	0.961	0.927	$< 2.2 \times 10^{-16}$
SRP003554*	Mouse	RiboCode	ribotricer	0.974	0.972	2.045×10^{-6}
SRP062407	Mouse	RiboCode	ORFscore	0.986	0.981	$< 2.2 \times 10^{-16}$
SRP078005	Mouse	ribotricer	RiboCode	0.989	0.968	$< 2.2 \times 10^{-16}$
SRP091889	Mouse	ribotricer	RiboCode	0.981	0.966	$< 2.2 \times 10^{-16}$
SRP115915*	Mouse	ribotricer	RiboCode	0.926	0.923	1.095×10^{-11}
ERP007231*	Rat	RiboTaper	RiboCode	0.955	0.953	3.321×10^{-9}
SRP045777	Rat	ribotricer	RiboCode	0.793	0.746	$< 2.2 \times 10^{-16}$
SRP056012*	Rat	ORFscore	RiboCode	0.971	0.872	$< 2.2 \times 10^{-16}$
SRP010040*	Zebrafish	ribotricer	ORFscore	0.658	0.562	$< 2.2 \times 10^{-16}$
SRP023492	Zebrafish	ribotricer	ORFscore	0.970	0.958	$< 2.2 \times 10^{-16}$
SRP034750*	Zebrafish	ribotricer	RiboCode	0.995	0.977	$< 2.2 \times 10^{-16}$
SRP032814	*C. albicans*	ribotricer	RiboCode	0.953	0.842	$< 2.2 \times 10^{-16}$
SRP062129	Chimp	ribotricer	ORFscore	0.918	0.883	$< 2.2 \times 10^{-16}$
SRP107240	*S. pombe*	ribotricer	RiboCode	0.972	0.939	$< 2.2 \times 10^{-16}$
SRP062129	Macaque	ribotricer	ORFscore	0.904	0.854	$< 2.2 \times 10^{-16}$

Table 3.9

Best and second to best performing methods at F1 score metric for each dataset using dataset-specific cutoff. F1 (B) and F1 (SB) denotes the F1 scores for the best and the second to best methods respectively. An asterisk (*) indicates that the dataset was used to learn the cutoffs by maximizing the F1 score. A # indicates the ribotricer phase score cutoff for the dataset is taken to be the median phase score difference between CDS annotated Ribo-seq and RNA-seq profiles.

SRP	Species	Best (B)	Second Best (SB)	F1 (B)	F1 (SB)
SRP018118*	Arabidopsis	ribotricer	RiboCode	0.937	0.848
SRP029587	Arabidopsis	ribotricer	RiboCode	0.645	0.176
SRP059391*	Arabidopsis	ribotricer	RiboCode	0.562	0.361
SRP087624	Arabidopsis	ribotricer	ORFscore	0.675	0.338
SRP108862	Arabidopsis	ribotricer	RiboCode	0.628	0.333
SRP000637	Baker's Yeast	RiboCode	ribotricer	0.680	0.503
SRP028552*	Baker's Yeast	ribotricer	RiboCode	0.964	0.859
SRP028614	Baker's Yeast	ribotricer	RiboCode	0.855	0.738
SRP033499	Baker's Yeast	RiboCode	RiboTaper	0.747	0.705
SRP075766*	Baker's Yeast	ribotricer	RiboTaper	0.951	0.877
SRP010374	*C. elegans*	ribotricer	RiboCode	0.799	0.517
SRP014427	*C. elegans*	ribotricer	RiboCode	0.826	0.776
SRP026198*	*C. elegans*	ribotricer	RiboCode	0.828	0.636
SRP056647	*C. elegans*	ribotricer	RiboCode	0.690	0.634
SRP028243*	Drosophila	ribotricer	RiboCode	0.693	0.562
SRP045475	Drosophila	ribotricer	RiboCode	0.561	0.391
SRP076919	Drosophila	ribotricer	RiboCode	0.667	0.125
SRP108999*	Drosophila	ribotricer	RiboCode	0.769	0.400
SRP010679*	Human	ribotricer	RiboCode	0.877	0.773
SRP029589	Human	ribotricer	RiboCode	0.651	0.599
SRP063852	Human	ribotricer	RiboCode	0.919	0.854
SRP098789*	Human	ribotricer	RiboCode	0.932	0.824
SRP102021	Human	ribotricer	RiboCode	0.890	0.835
SRP003554*	Mouse	RiboTaper	ribotricer	0.901	0.899
SRP062407	Mouse	RiboTaper	ribotricer	0.930	0.910
SRP078005	Mouse	ribotricer	RiboCode	0.951	0.901
SRP091889	Mouse	ribotricer	RiboCode	0.938	0.900
SRP115915*	Mouse	ribotricer	RiboCode	0.853	0.842
ERP007231*	Rat	ribotricer	RiboTaper	0.879	0.874
SRP045777	Rat	RiboCode	ribotricer	0.618	0.511
SRP056012*	Rat	ribotricer	RiboCode	0.787	0.786
SRP010040*	Zebrafish	ribotricer	RiboCode	0.670	0.377
SRP023492	Zebrafish	RiboCode	ribotricer	0.838	0.826
SRP034750*	Zebrafish	RiboCode	ribotricer	0.920	0.894
SRP032814#	*C.albicans*	ribotricer	RiboCode	0.883	0.752
SRP062129#	Chimp	ribotricer	RiboCode	0.865	0.436
SRP062129#	Macaque	ribotricer	RiboCode	0.842	0.635
SRP107240#	*S. pombe*	ribotricer	RiboCode	0.913	0.869

Table 3.10

Best and second to best performing methods at F1 score metric for each dataset-specific cutoff. F1 (B) and F1 (SB) denotes the F1 scores for the best and the second to best methods respectively. The cutoff was learned independently for each dataset as the median difference between Ribo-seq and RNA-seq phase scores over protein coding ORFs.

SRP	Species	Best (B)	Second Best (SB)	F1 (B)	F1 (SB)
SRP018118	Arabidopsis	ribotricer	RiboCode	0.920	0.848
SRP029587	Arabidopsis	ribotricer	RiboCode	0.846	0.176
SRP059391	Arabidopsis	ribotricer	RiboCode	0.678	0.361
SRP087624	Arabidopsis	ribotricer	ORFscore	0.671	0.338
SRP108862	Arabidopsis	ribotricer	RiboCode	0.695	0.333
SRP000637	Baker's Yeast	ribotricer	RiboCode	0.850	0.680
SRP028552	Baker's Yeast	ribotricer	RiboCode	0.928	0.859
SRP028614	Baker's Yeast	ribotricer	RiboCode	0.923	0.738
SRP033499	Baker's Yeast	ribotricer	RiboCode	0.904	0.747
SRP075766	Baker's Yeast	ribotricer	RiboTaper	0.935	0.877
SRP010374	*C.elegans*	ribotricer	RiboCode	0.798	0.517
SRP014427	*C.elegans*	ribotricer	RiboCode	0.868	0.776
SRP026198	*C.elegans*	ribotricer	RiboCode	0.846	0.636
SRP056647	*C.elegans*	ribotricer	RiboCode	0.716	0.634
SRP028243	Drosophila	ribotricer	RiboCode	0.679	0.562
SRP045475	Drosophila	ribotricer	RiboCode	0.667	0.391
SRP076919	Drosophila	ribotricer	RiboCode	0.667	0.125
SRP108999	Drosophila	ribotricer	RiboCode	0.818	0.400
SRP010679	Human	ribotricer	RiboCode	0.878	0.773
SRP029589	Human	ribotricer	RiboCode	0.765	0.599
SRP063852	Human	ribotricer	RiboCode	0.919	0.854
SRP098789	Human	ribotricer	RiboCode	0.922	0.824
SRP102021	Human	ribotricer	RiboCode	0.900	0.835
SRP003554	Mouse	ribotricer	RiboTaper	0.919	0.901
SRP062407	Mouse	ribotricer	RiboTaper	0.936	0.930
SRP078005	Mouse	ribotricer	RiboCode	0.944	0.901
SRP091889	Mouse	ribotricer	RiboCode	0.924	0.900
SRP115915	Mouse	ribotricer	RiboCode	0.863	0.842
ERP007231	Rat	RiboTaper	RiboCode	0.874	0.867
SRP045777	Rat	ribotricer	RiboCode	0.722	0.618
SRP056012	Rat	RiboCode	ribotricer	0.786	0.738
SRP010040	Zebrafish	ribotricer	RiboCode	0.668	0.377
SRP023492	Zebrafish	ribotricer	RiboCode	0.918	0.838
SRP034750	Zebrafish	ribotricer	RiboCode	0.937	0.920

Table 3.11
Ribotricer's performance at F1 score when considering species-specific or dataset-specific cutoff F1 (SS) and F1 (DS) denotes the F1 scores for ribotricer when using species-specif and dataset-specific cutoffs respectively. Ribo-RNA indicates the median difference between phase score of protein coding ORFs in Ribo- and RNA-seq samples. 'sampled' indicates the median was calculated using 30% of protein coding ORFs per dataset with resampling ($n_{\textbf{bootstraps}} = 10000$) while 'all' indicates the median was calculated using the complete list of protein coding ORFs.

SRP	species	F1 (SS)	F1 (DS)	Ribo-RNA (sampled)	Ribo-RNA (all)
SRP018118	Arabidopsis	0.937	0.920	0.455	0.447
SRP029587	Arabidopsis	0.645	0.846	0.206	0.191
SRP059391	Arabidopsis	0.562	0.678	0.109	0.104
SRP087624	Arabidopsis	0.675	0.671	0.233	0.145
SRP108862	Arabidopsis	0.628	0.695	0.181	0.154
SRP000637	Baker's Yeast	0.503	0.850	0.186	0.179
SRP028552	Baker's Yeast	0.964	0.928	0.383	0.382
SRP028614	Baker's Yeast	0.855	0.923	0.267	0.263
SRP033499	Baker's Yeast	0.573	0.904	0.204	0.194
SRP075766	Baker's Yeast	0.951	0.935	0.694	0.671
SRP010374	*C.elegans*	0.799	0.798	0.224	0.222
SRP014427	*C.elegans*	0.826	0.868	0.343	0.334
SRP026198	*C.elegans*	0.828	0.846	0.322	0.316
SRP056647	*C.elegans*	0.690	0.716	0.141	0.135
SRP028243	Drosophila	0.693	0.679	0.109	0.098
SRP045475	Drosophila	0.561	0.667	-0.019	-0.020
SRP076919	Drosophila	0.667	0.667	-0.025	-0.034
SRP108999	Drosophila	0.769	0.818	0.363	0.360
SRP010679	Human	0.878	0.878	0.421	0.404
SRP029589	Human	0.651	0.765	0.234	0.223
SRP063852	Human	0.919	0.919	0.522	0.498
SRP098789	Human	0.932	0.922	0.526	0.514
SRP102021	Human	0.891	0.900	0.427	0.417
SRP003554	Mouse	0.900	0.919	0.542	0.526
SRP062407	Mouse	0.910	0.936	0.588	0.568
SRP078005	Mouse	0.951	0.944	0.603	0.591
SRP091889	Mouse	0.939	0.924	0.509	0.497
SRP115915	Mouse	0.854	0.863	0.372	0.361
ERP007231	Rat	0.879	0.863	0.403	0.388
SRP045777	Rat	0.511	0.722	0.176	0.173
SRP056012	Rat	0.787	0.738	0.264	0.247
SRP010040	Zebrafish	0.670	0.668	0.136	0.108
SRP023942	Zebrafish	0.826	0.918	0.512	0.502
SRP034750	Zebrafish	0.894	0.937	0.660	0.649

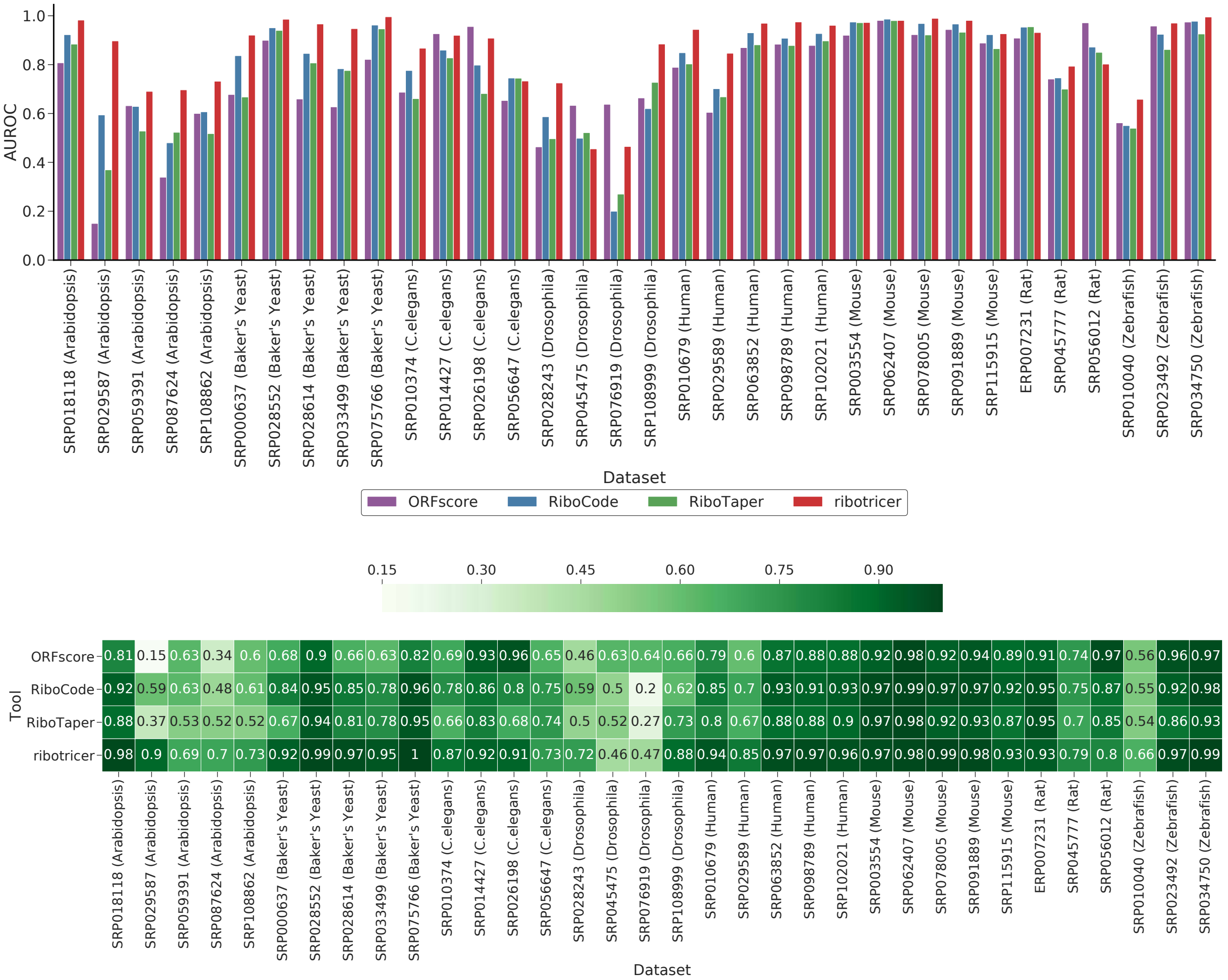


Figure 3.35

Distribution of area under ROC (AUROC) across multiple species. For each Ribo-seq and RNA-seq pair in a dataset, area under ROC was calculated for exon level classification using Ribotricer, Ribotaper, RiboCode and ORFScore.

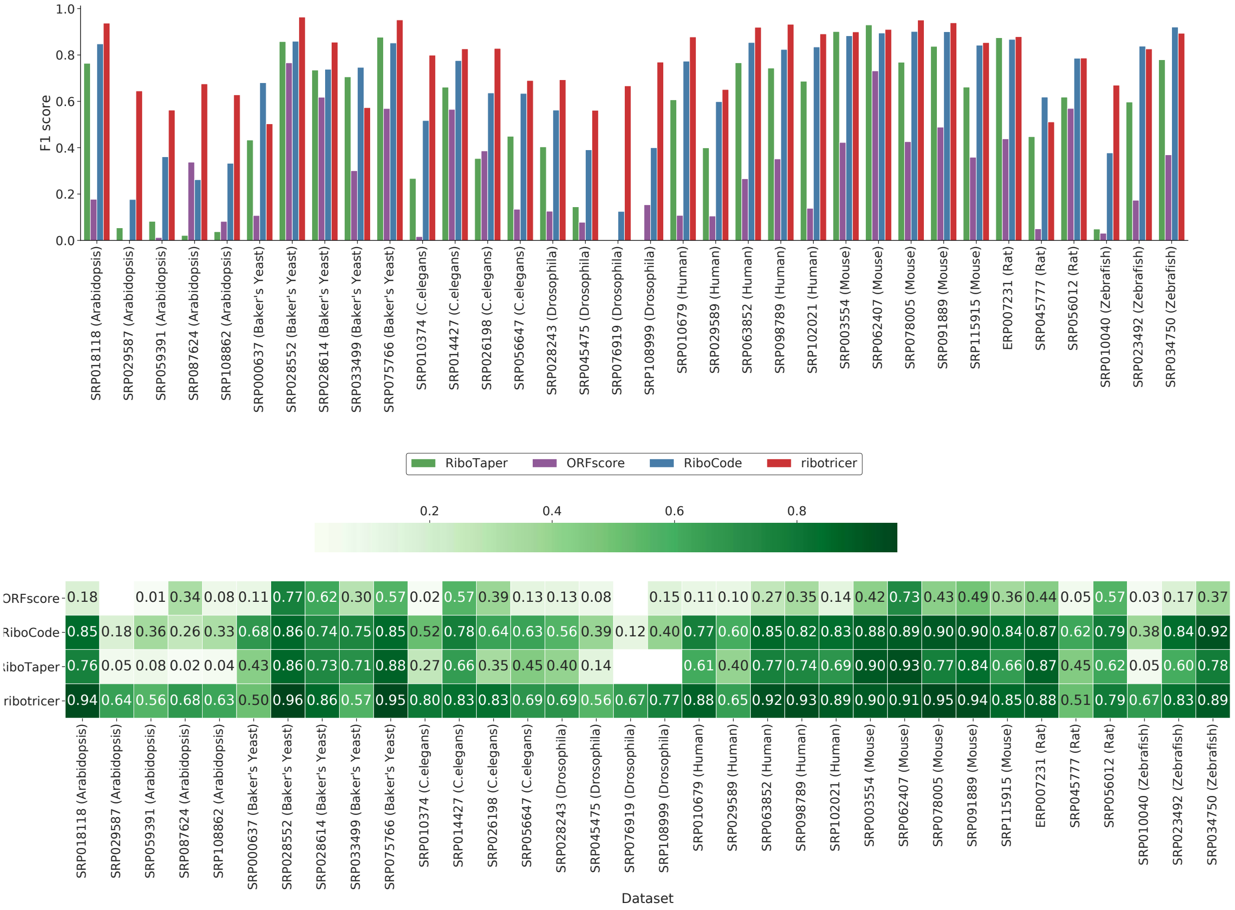


Figure 3.36
Distribution of F1 scores across species using species-specific cutoff. For each species two datasets were used to learn the cutoff score of ribotricer for that species.

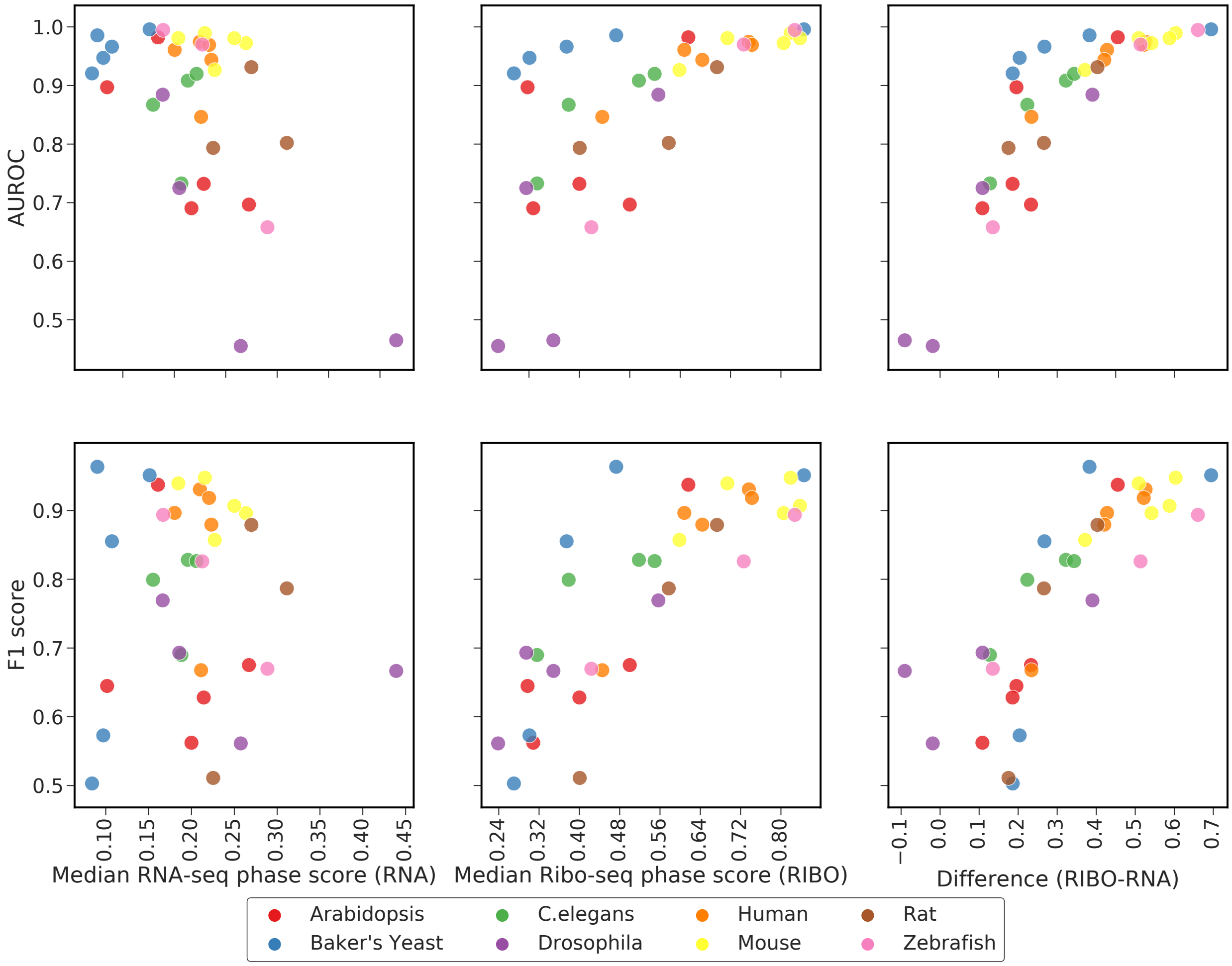


Figure 3.37

Performance of ribotricer at AUROC and F1 scores metrics across species at different median phase scores of RNA-seq and Ribo-seq samples using species-specific cutoff. For each dataset, median phase score was calculated for both RNA-seq and Ribo-seq samples for the same list of candidate ORFs.

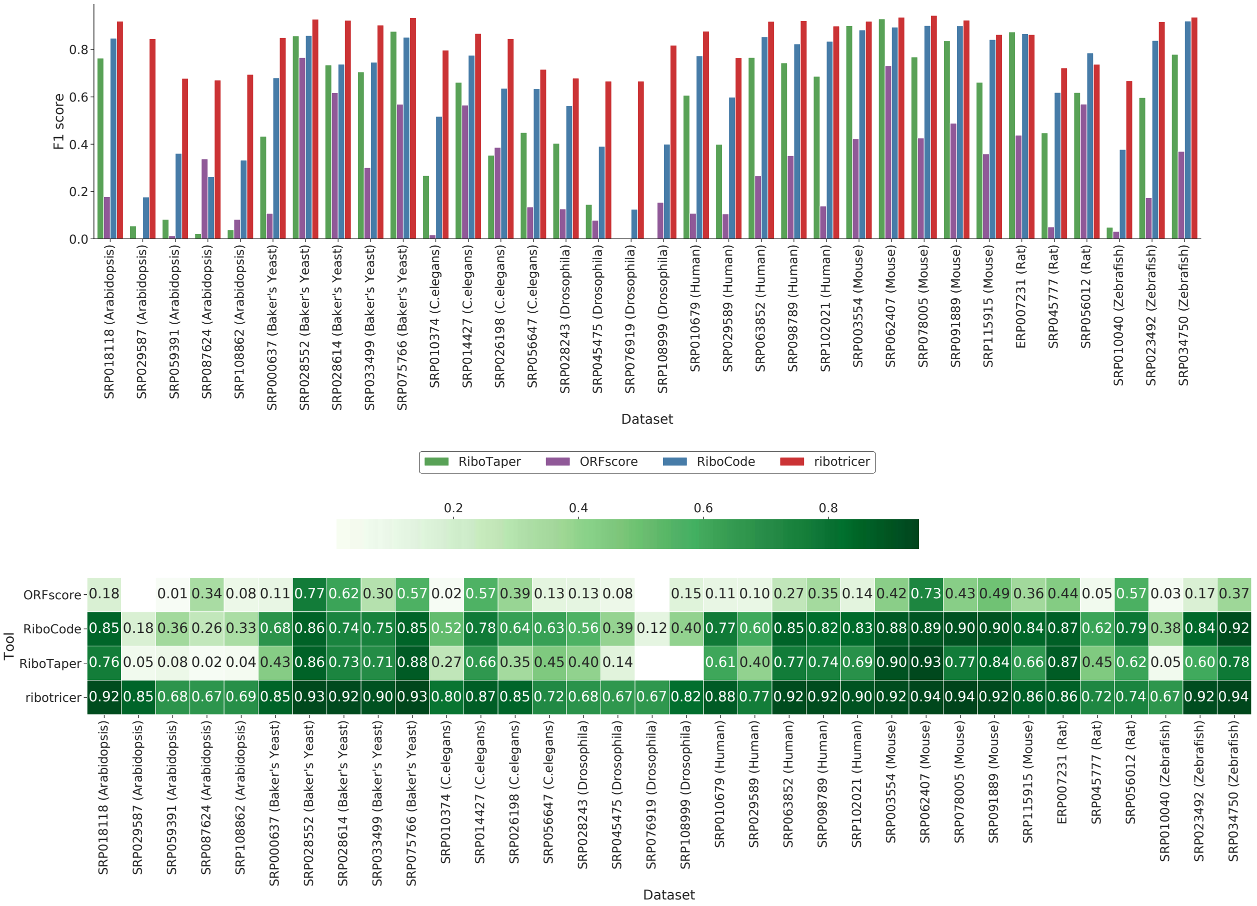


Figure 3.38

Distribution of F1 scores across species using dataset-specific cutoff. For each dataset, the cutoff was learned by determining the median phase score difference between Ribo-seq and RNA-seq profiles by sampling one-third of the total protein-coding transcripts $n_{\text{bootstrap}} = 10000$ times.

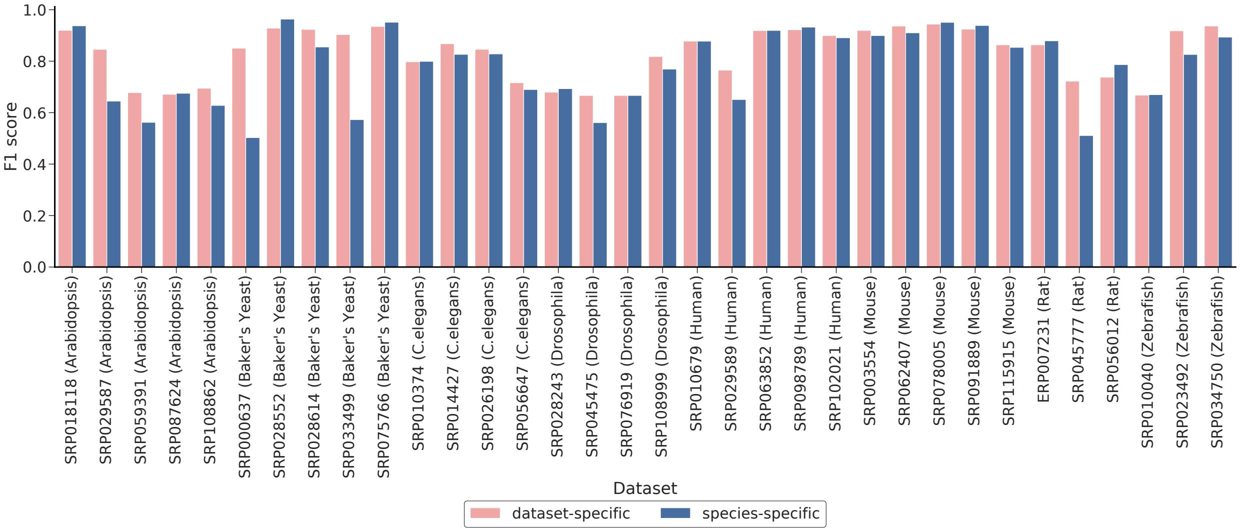


Figure 3.39
Difference in performance of ribotricer using species-specific or dataset-specific cutoffs. Species-specific cutoffs were learned by maximizing the F1 scores for two datasets per species while dataset-specific cutoffs were learned per dataset using the median difference of phase score of Ribo-seq and RNA-seq protein coding profiles.

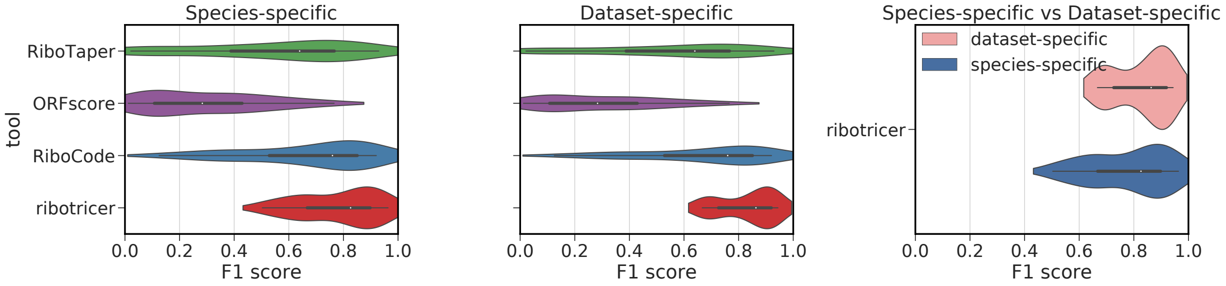


Figure 3.40
Summarized performance of ribotricer using species-specific and dataset-specific strategies. Species-specific cutoffs were learned by maximizing the F1 scores for two datasets per species while dataset-specific cutoffs were learned per dataset using the median difference of phase score of Ribo-seq and RNA-seq protein coding profiles. Species-specific and dataset-specific cutoffs only apply to ribotricer.

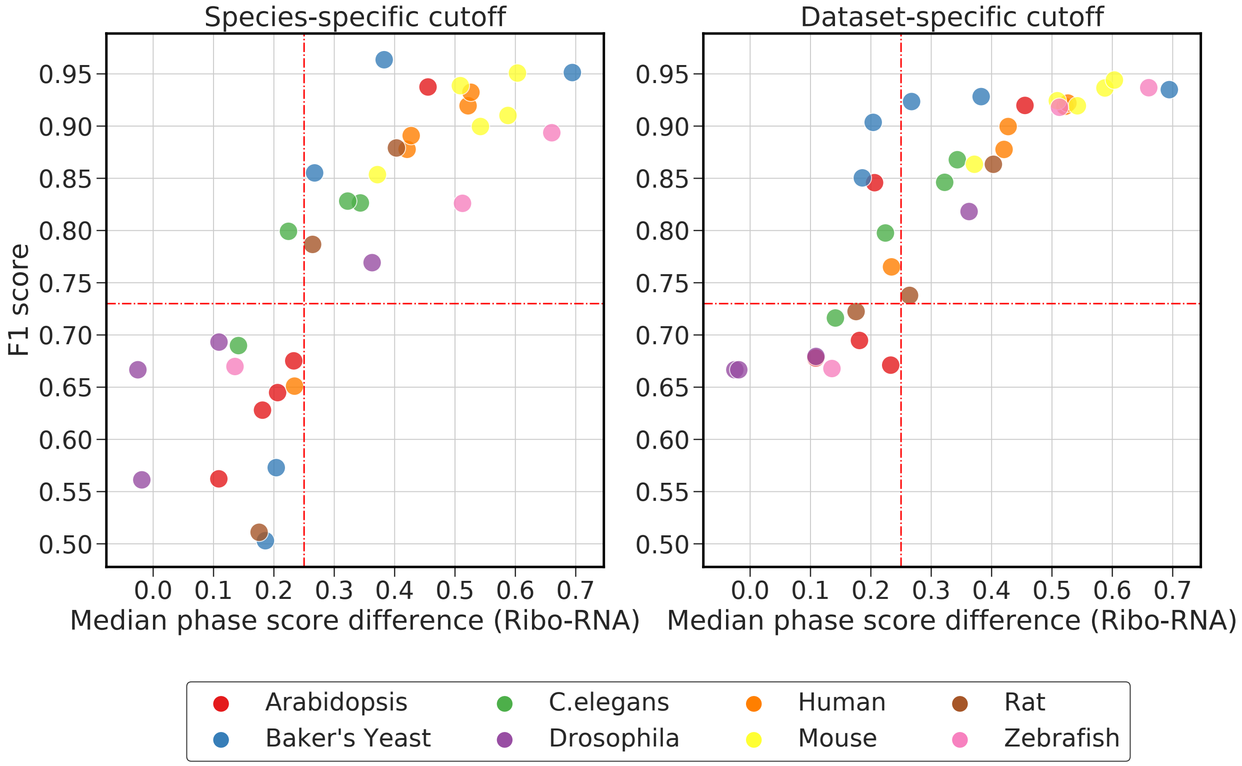


Figure 3.41
Distribution of ribotricer's F1 scores with respect to median phase score difference of Ribo-seq and RNA-seq, using species-specific and dataset-specific cuoffs. Species-specific cutoffs were learned by maximizing the F1 scores for two datasets per species while dataset-specific cutoffs were learned per dataset using the median difference of phase score of Ribo-seq and RNA-seq protein coding profiles. The dashed red lines indicate a median difference of 0.25 between Ribo-seq and RNA-seq phase scores results in a F1 score of 0.73 and above.

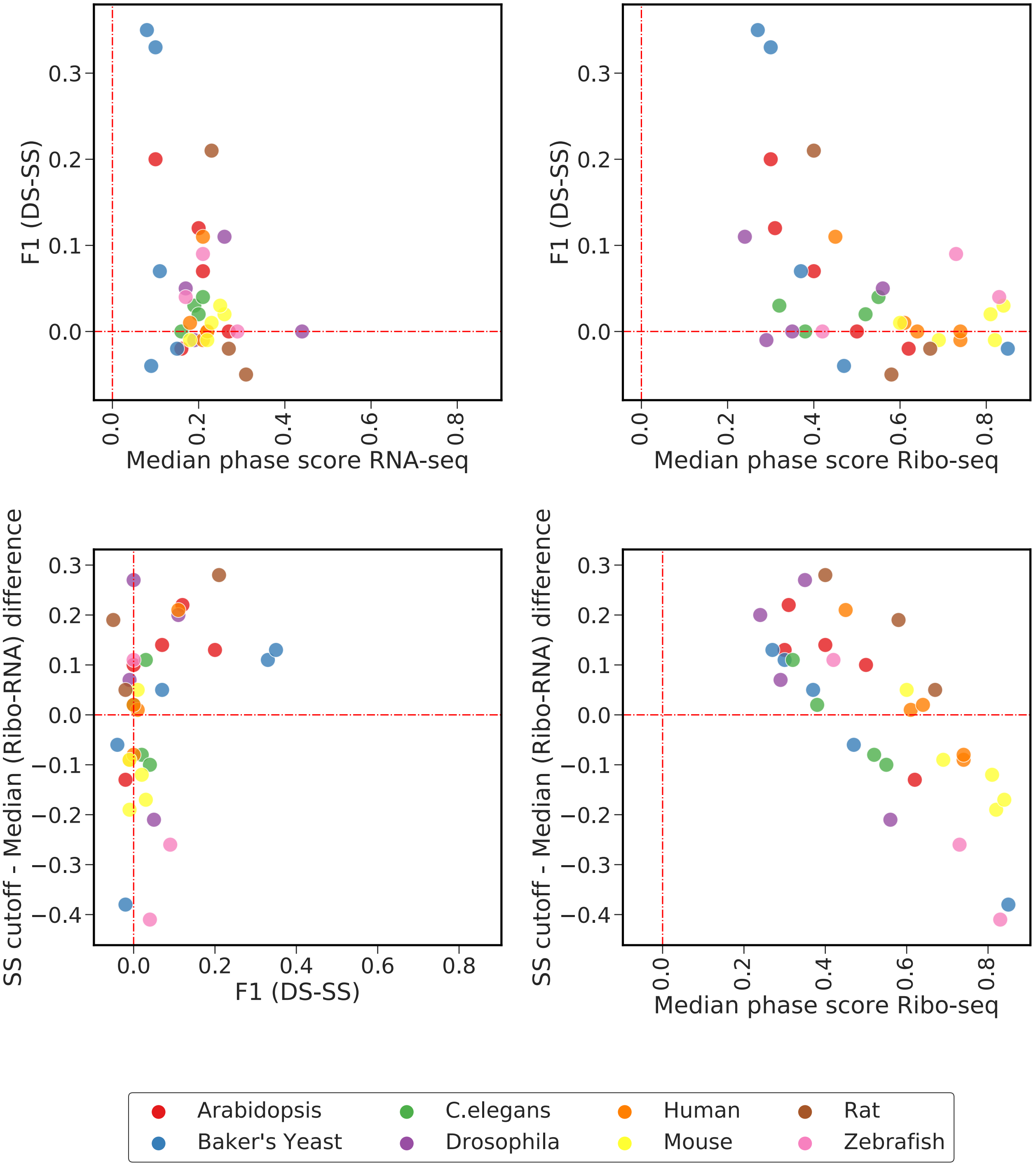


Figure 3.42
Effect of Ribo-seq and RNA-seq phase scores on species-specific and dataset-specific based F1 performance. F1 (DS-SS) indicates difference in F1 scores using species-specific (SS) or dataset-specific (DS) cutoff. Each single data point represents one dataset. Median phase scores were calculated using all the candidate ORF profiler of either RNA-seq or Ribo-seq sample.

learned empirically by using Ribo-seq and RNA-seq samples from two datasets and maximizing the F1 score by treating the Ribo-seq profiles of CCDS/CDS regions as ground true positive and the corresponding RNA-seq profiles as true negatives (Figure 3.34; Table 3.5). However, this approach is only best suited for species where there are multiple datasets available. For a new species where there are only few or none datasets available and hence the cutoff cannot be learned empirically, we recommend using the median score difference between the profiles of annotated CDS regions of a Ribo-seq and the corresponding RNA-seq sample. This strategy is also used by RibORF [34] which tunes the parameters of its model by selecting one-third of the CDS profiles as true positives. We followed this strategy of using the median phase score difference as the phase score cutoff for each of the four species: *C. albicans*, chimpanzee, macaque and *S. pombe*. Except for *S. pombe*, all other species have only one public dataset available to the best of our knowledge (Table S1).

We first generated candidate ORF list for each species using ribotricer over transcripts with annotated CDS regions. Phase scores were then calculated for each RNA-seq and Ribo-seq sample over these CDS annotated candidate ORFs (Figure 3.45). The median differences in Ribo-seq and RNA-seq phase scores for *C. albicans*, chimpanzee, macaque and *S. pombe* is summarized at the end of Table 3.7. We used these differences as species-specific cutoffs for benchmarking ribotricer against other methods.

Ribotricer results in the best AUROC for all the four species with the difference between ribotricer and the second best method statistically significant in all the cases (Figure 3.43; Table 3.8). It is worth mentioning that the AUROC metric is not dependent on the choice of the learned cutoff. Furthermore, ribotricer is also the best method using the F1 score metric (Figure 3.44; Table 3.9).

We recommend using the species-specific cutoffs for all the species as listed in Table 3.6. For any new species, we recommend using median phase score differences on ribotricer generated candidate ORFs over CDS annotated transcripts between Ribo-seq and RNA-seq samples (Figure 3.45). This can be determined by `ribotricer` itself, using the `learn-cutoff` subcommand.

3.5 Using ribotricer

In order to use `ribotricer`, the following three files are required:

- **GTF:** genome annotation file in GTF format (ENSEMBL/Gencode/others)
- **FASTA:** reference genome file in FASTA format
- **BAM:** alignment file in BAM format

Henceforth, we use the boldface acronyms above to refer to these files as such.

3.5.1 Preparing candidate ORFs list

ribotricer prepares a candidate list of ORFs given a GTF and FASTA file. For any species, given a reference and a fixed version of GTF, this step only needs to be done once. Ribotricer by default searches for ORFs defined by an 'AUG' start and an in-frame stop codon ('UAG', 'UAA', and 'UGA') and are a minimum of 60 nucleotides long. It is possible to expand the definition of ORF by supplying a list of all start codons using the `--start_codons` parameter. It is also possible to change the minimum length of an ORF by using the `--min_orf_length` option. If multiple potential in-frame start codons exist upstream of a stop codon, we always choose AUG if it exists, otherwise, we take the most upstream one as the start codon.

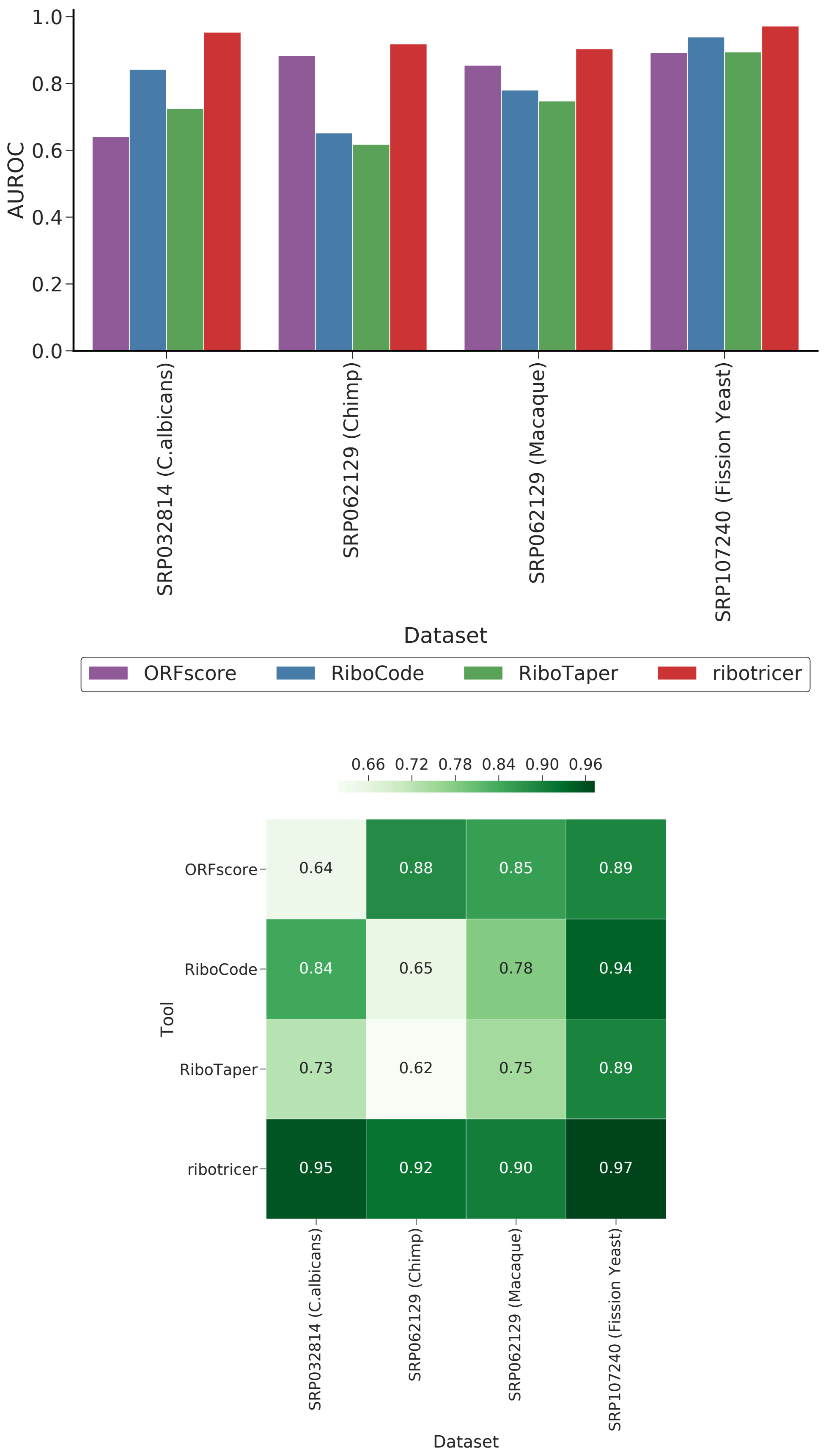


Figure 3.43

Distribution of area under ROC in the independent datasets. For each Ribo-seq and RNA-seq pair in a dataset, area under ROC was calculated for exon level classification using Ribotricer, Ribotaper, RiboCode and ORFScore.

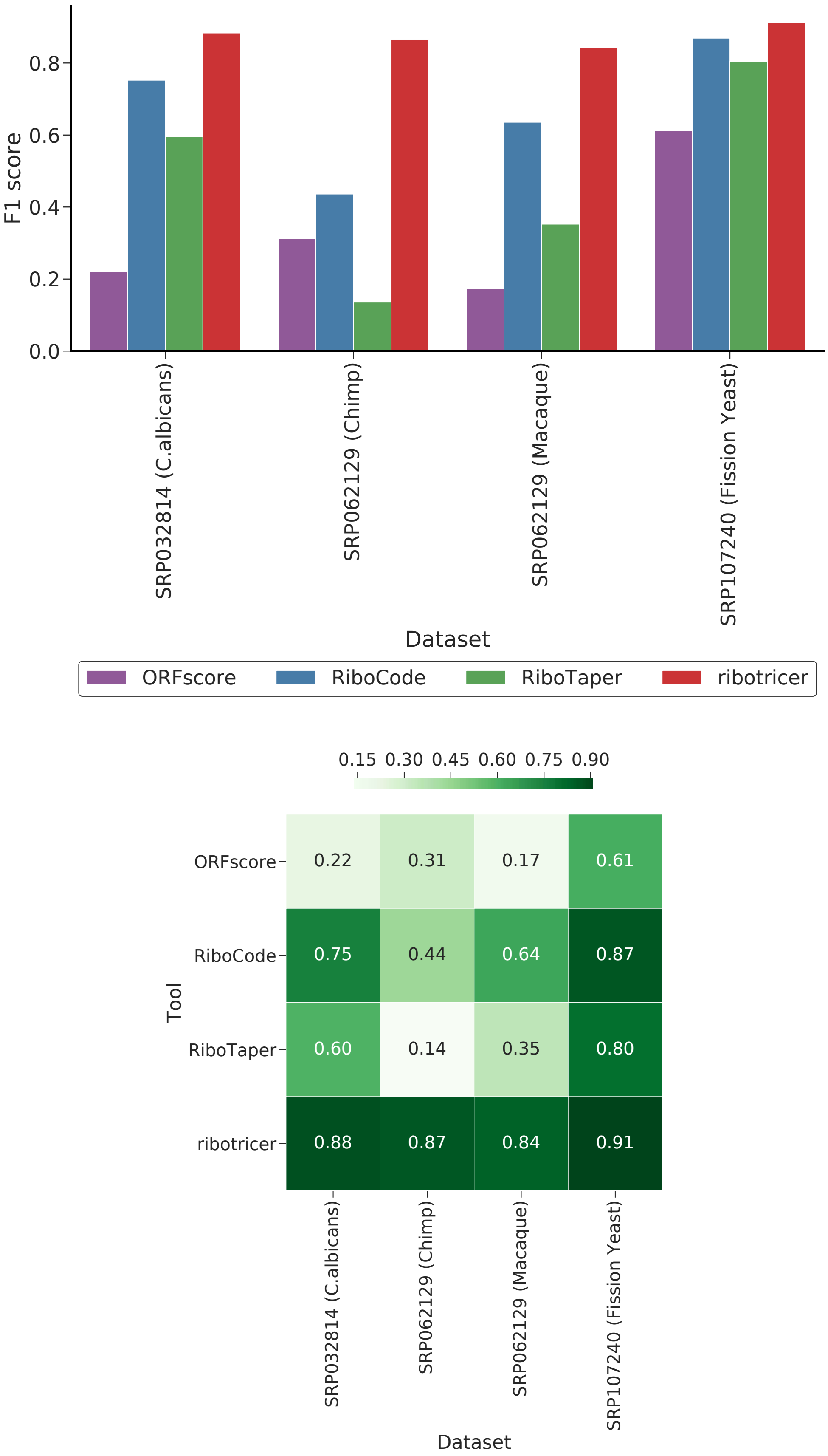


Figure 3.44

Distribution of F1 scores in the independent datasets. For each Ribo-seq and RNA-seq pair in a dataset, F1 score was calculated for exon level classification using Ribotricer, Ribotaper, RiboCode and ORFScore.

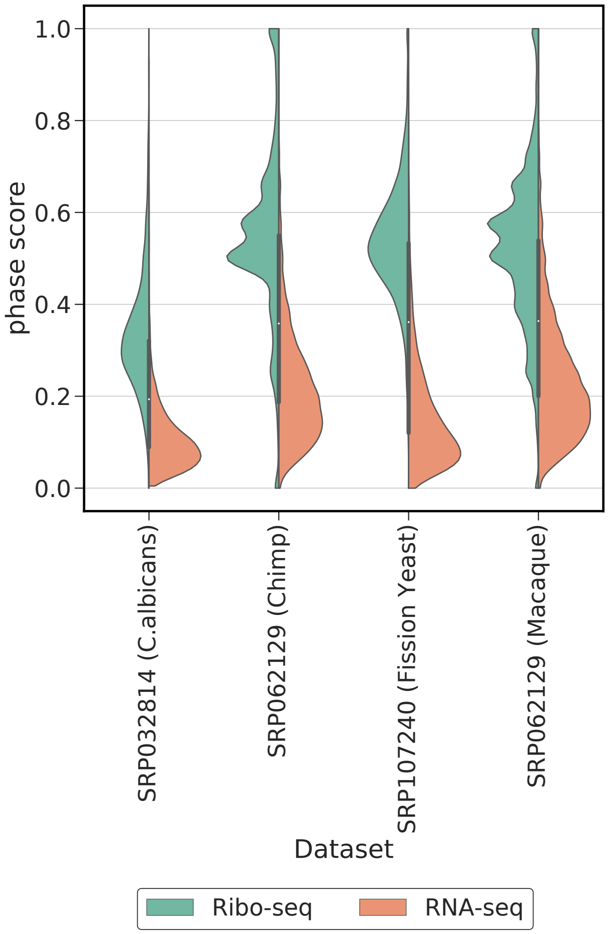


Figure 3.45
Distribution of ribotricer's phase scores for RNA-seq and Ribo-seq samples in the independent datasets. For each dataset phase scores were calculated for all candidate ORFs. For human and mouse, Ribo-seq CCDS profiles were treated as true positive and the corresponding RNA-seq profile was treated as true negative. For all other species Ribo-seq profile of annotated CDS regions were treated as true positive and the corresponding RNA-seq profile treated as true negative.

```
ribotricer prepare-orfs --gtf {GTF} \
                         --fasta {FASTA} \
                         --prefix {RIBOTRICER_INDEX}
```

The command above will create a list of candidate ORFs at the RIBOTRICER_INDEX location.

For this study, we used a total of ten codons with a maximum of one nucleotide difference from "ATG" as potential start codons including ATA, ATC, ATT, AAG, ACG, AGG, ATG, CTG, GTG, TTG. Note that we use 'T' as a nucleotide here instead of 'U' as the reference FASTA almost always contains DNA sequences.

3.5.2 Detecting actively translating ORFs using ribotricer

Ribotricer's ORF list as created above can then be used along with the BAM to define the translation status of these ORFs:

```
ribotricer detect-orfs --bam {BAM} \
                        --ribotricer_index {RIBOTRICER_INDEX}_candidate_ORFs.tsv \
                        --prefix {OUT_PREFIX}
```

For each ORF in the candidate ORFs list, ribotricer calculates the phase score on the read profiles after performing read length appropriate offset shifts. These offsets are determined by maximizing the cross-correlation of these profiles with the profile for the most abundant read length. Additionally, ribotricer automatically infers the sequencing protocol (forward/reverse) and only uses unique mapping reads that conform to the strand orientation in the GTF. For example, a read uniquely mapping to a gene defined on the negative strand for a forward stranded protocol, will be discarded.

In order to assign 'non-translating' or 'translating' status, ribotricer, by default, uses a cutoff threshold of 0.428. ORFs with phase score above 0.428 are marked as translating as long as they have at least five codons with non-zero read count. Ribotricer does not take coverage into account for predicting an ORF to be translating or not-translating. Apart from these two criteria, there is no other requirement for an ORF to be active. Though, a region with higher overall coverage as defined by number of reads per unit codon might be a more confident 'hit' for active translation, our method is designed to find evidence of active translation based on the qualitative pattern of "high-low-low" and hence our rankings are purely based on phase scores.

The default cutoff (0.428) was learned using public human and mouse Ribo-seq datasets, where the gap between Ribo- and RNA-seq phase scores is the highest amongst other species (Table 3.7) and hence, it is a conservative cutoff for detecting active translation. We provide a list of species-specific recommended cutoffs (Table 3.6), optimized for F1 score based performance.

The main output of the above command is a tab separated file consisting for each candidate ORF, its translation status, the corresponding transcript and gene and the ORF type. Different ORF types defined by ribotricer are described below:

- **annotated:** CDS annotated in the provided GTF file
- **super_uORF:** upstream ORF of the annotated CDS, not overlapping with any CDS of the same gene
- **super_dORF:** downstream ORF of the annotated CDS, not overlapping with any CDS of the same gene
- **uORF:** upstream ORF of the annotated CDS, not overlapping with the main CDS

- **dORF:** downstream ORF of the annotated CDS, not overlapping with the main CDS
- **overlap_uORF:** upstream ORF of the annotated CDS, overlapping with the main CDS
- **overlap_dORF:** downstream ORF of the annotated CDS, overlapping with the main CDS
- **novel:** ORF in non-coding genes or in non-coding transcripts of coding genes

3.5.3 Filtering actively translating ORFs using multiple criteria

In order to assign 'non-translating' or 'translating' status, ribotricer by default uses a cutoff threshold of '0.428'. ORFs with phase score above '0.428' are marked as translating as long as they have at least five codons with non-zero read count. By default, ribotricer does not take coverage or count information explicitly into account for predicting an ORF to be translating or not-translating. However, this behavior can be changed by following filters:

- `--min_valid_codons` (default=5): Minimum number of codons with non-zero reads for determining active translation
- `--min_valid_codons_ratio` (default=0): Minimum ratio of codons with non-zero reads to total codons for determining active translation
- `--min_reads_per_codon` (default=0): Minimum number of reads per codon for determining active translation
- `--min_read_density` (default=0.0): Minimum read density (total reads/length) over an ORF total codons for determining active translation

For each of the above filters, an ORF failing **any** of the filters is marked as 'non-translating'.

For example, to ensure that each ORF has at least 3/4 of its codons non-empty, we can specify

`--min_valid_codons_ratio` to be 0.75:

```
ribotricer detect-orfs --bam {BAM} \
                       --ribotricer_index {RIBOTRICER_INDEX}_candidate_ORFs.tsv \
                       --prefix {OUTPUT_PREFIX}
                       --min_valid_codons_ratio 0.75
```

It might also often be desired to have some minimum density of reads over an ORF. The read density here is defined as the ratio of total number of reads over an ORF to its length. For example to ensure that each 'translating' ORF has at least a read density of 10, we will specify `--min_read_density` to be 10.

```
ribotricer detect-orfs --bam {BAM} \
                       --ribotricer_index {RIBOTRICER_INDEX}_candidate_ORFs.tsv \
                       --prefix {OUTPUT_PREFIX}
                       --min_read_density 10.0
```

The above filters can be combined to give ORFs that have high read density as well as have have reads present over most of the codons in the profile. Note that increasing the value of any of the four filters will usually result in a smaller list of ORFs marked 'translating'.

3.5.4 Downstream ranking and filtering

It is also possible to filter actively-translating ORFs after running ribotricer. Ribotricer produces a tab separated file with columns that include read-density, number and ratio of valid codons to total codons in the ORF besides the phase score. As such, filtering can be performed downstream using awk or any other programming language. Here we provide an example of filtering and sorting the output of a ribotricer run using Python using the pandas library:

Listing 3.1
Filtering ORFs using python. The function returns a filtered list of translating ORFs which have a read density of at least 2.5; a total read count of atleast 50; and the ratio of non-empty codons to total codons atleast 0.75.

```
import pandas as pd
def filtered_df(df):
        df_filtered = df.loc[df.status=='translating']
        df_filtered = df.loc[(df['read_density']>=2.5) & \
                             (df['read_count']>=50) & \
                         (df['valid_codons_ratio']>=0.75)]
    df_sorted = df_filtered.sort_values(by=['phase_score',
                                            'read_density'],
                                        ascending=[False,
                                                   False])
    return df_sorted


# read ribotricer output
ribotricer_output_df = pd.read_csv('/path/to/translating_ORFs.tsv', sep='\t
# filter and sort ribotricer output
```

```
ribotricer_filtered_df = filtered_df(ribotricer_output_df)
```

3.5.5 Learning cutoff empirically from data

Ribotricer can learn cutoff empirically from the data. Given at least one Ribo-seq and one RNA-seq BAM file, ribotricer learns the cutoff by running one iteration of the algorithm on the provided files with a pre-specified cutoff (`--phase_score_cutoff`, default: 0.428) and then uses the generated output to find the median difference between Ribo-seq and RNA-seq phase scores of only candidate ORFs with `transcript_type` annotated as `protein_coding`:

```
ribotricer learn-cutoff --ribo_bams ribo_bam1.bam,ribo_bam2.bam \
--rna_bams rna_1.bam \
--prefix ribo_rna_prefix \
--ribotricer_index {RIBOTRICER_ANNOTATION}
```

By default, ribotricer searches for ORFs that are at least 60 nucleotides or 20 codons long to build the candidate list but this minimum length can be set to a user-defined value. We arrived at the default value of 20 codons by performing a simulation using the Ribo-seq profiles of genes with total codons > 100 and with at least 50% non-empty codons. In the simulation, we randomly sampled $10 - 100$ codons and generated a "downsampled" profile. The mean absolute difference between the original phase score calculated using the full length profile versus the "downsampled" profile with 20 or more codons is smaller than 0.05 and does not change after increasing the number of codons (Figs. 3.24 and 3.25).

Ribotricer enables discovery of both short and long ORFs that will deepen our understanding of translational regulation across various biological contexts. We envision ribotricer's phase score to become a commonly used quality control metric for assessing the quality of Ribo-seq datasets, especially for novel datasets in species where no prior datasets exist.

Chapter 4

Translational landscape of morphological changes in *Candida albicans*

4.1 Introduction

Environmental human pathogens have evolved the ability to survive in diverse environmental conditions. Fungal diseases have remained an important public health problem since the early 1980s [138]. Candidemia and other forms of invasive candidasis remain one of the most prevalent mycoses [139] and *Candida albicans* (*C. albicans*) remains the most common species responsible for it [140]. *C. albicans*is a fungal pathogen that inhabits the mucosal surfaces of most healthy individuals as human commensals. Though mostly asymptomatic, *C. albicans*is an opportunistic pathogen known to causes disease in individuals with a debilitated immune system or a disruption in the host's microbiome [141]. It normally manifests as a commensal in the human gastrointestinal tract, oral and vaginal cavities. Immunocompromised individuals such as organ transplant patients, patients undergoing cytotoxic or immunosupressive therapies, and HIV/AIDS patients are particularly at higher risk of infection [142]. Candida infections are associated with considerable mortality and morbidity [139, 143, 144]. *C. albicans*'s virulence arises because of its ability to 1) form true hyphae 2) resist phagocytosis 3) adhere to the host surfaces, and 4) secrete proteinase [141].

C. albicans are aerobic yeasts yet can grow anaerobically [145] and have a diploid genome. It adheres to the host cells epithelial surface via the molecular adhesins located in its cellular envelope. They are capable of undergoing a reversible morphological transition from single budding yeasts to a continuously branching filaments via a transitory psuedohyphal state. This ability of *C. albicans* to undergo reversible filamentous transition is one of the key reasons responsible for its virulence [146]. The yeast to filament transition in *C. albicans* occurs in a variety of conditions, but most importantly in the presence of serum at body temperature (37°C). Understanding the transcriptional and translational mechanisms behind this morphological transition can provide a better understanding of its pathogenesis. It can also help in the development of new anti-fungal drugs that can prevent infections in immunocompromised patients. Here, we provide a comprehensive study of the changes in transcriptional and translational landscape using deep sequencing Ribo-seq and RNA-seq in yeast and filamentous like growth conditions.

4.2 Results

4.2.1 Deep RNA sequencing and Ribosome profiling of *C. albicans* under yeast and filamentous like growth conditions

In order to study the transcriptional and translational landscape changes involved in yeast to filamentous transition, we performed deep sequencing of mRNA and ribosome bound mRNA fragments in *C. albicans* cells growing at 30° C and 37°C in the presence of serum (Figure 4.2). While cells at 30°C undergo yeast like growth, the cells at 37°C in the presence of serum show filamentous growth. From all experiments, we obtained a total of 246.2 million reads with a median of 6.25 million reads for Ribo-seq and 20.5 million reads for RNA-seq experiments (Figure 4.1).

4.2.2 Changes in the transcriptional landscape

In order to study changes that occur at the transcriptional level during yeast to filamentous transition, we performed differential expression analysis between the two conditions. Using DESeq2 [147], we identified a total of 498 genes that are up-regulated ($\log_2$ fold change > 1; adjusted p < 0.05) and 385 that are down-regulated ($\log_2$ fold change < -1; adjusted p < 0.05). A gene ontology analysis reveals, that while the up-regulated genes show enrichment in filamentous growth, nuclear periphery, and cellular bud neck besides other terms (Figure 4.4A). On the other hand, the down-regulated genes are involved in translation, ribosome and rRNA synthesis (Figure 4.4B). Table 4.1 highlights some of the genes that are transcriptionally induced during the morphological transition, have virulence-related properties and show significantly reduced translational efficiency (TE).

4.2.3 Translational landscape changes

In order to identify changes at the translational landscape, we sequenced the ribosome protected fragments (RPFs) in the two conditions following the protocol of Ingolia *et al.* [148]. RPFs are enriched in the range of 28-32 nucleotides as expected (Figure 4.2A). There is high correlation

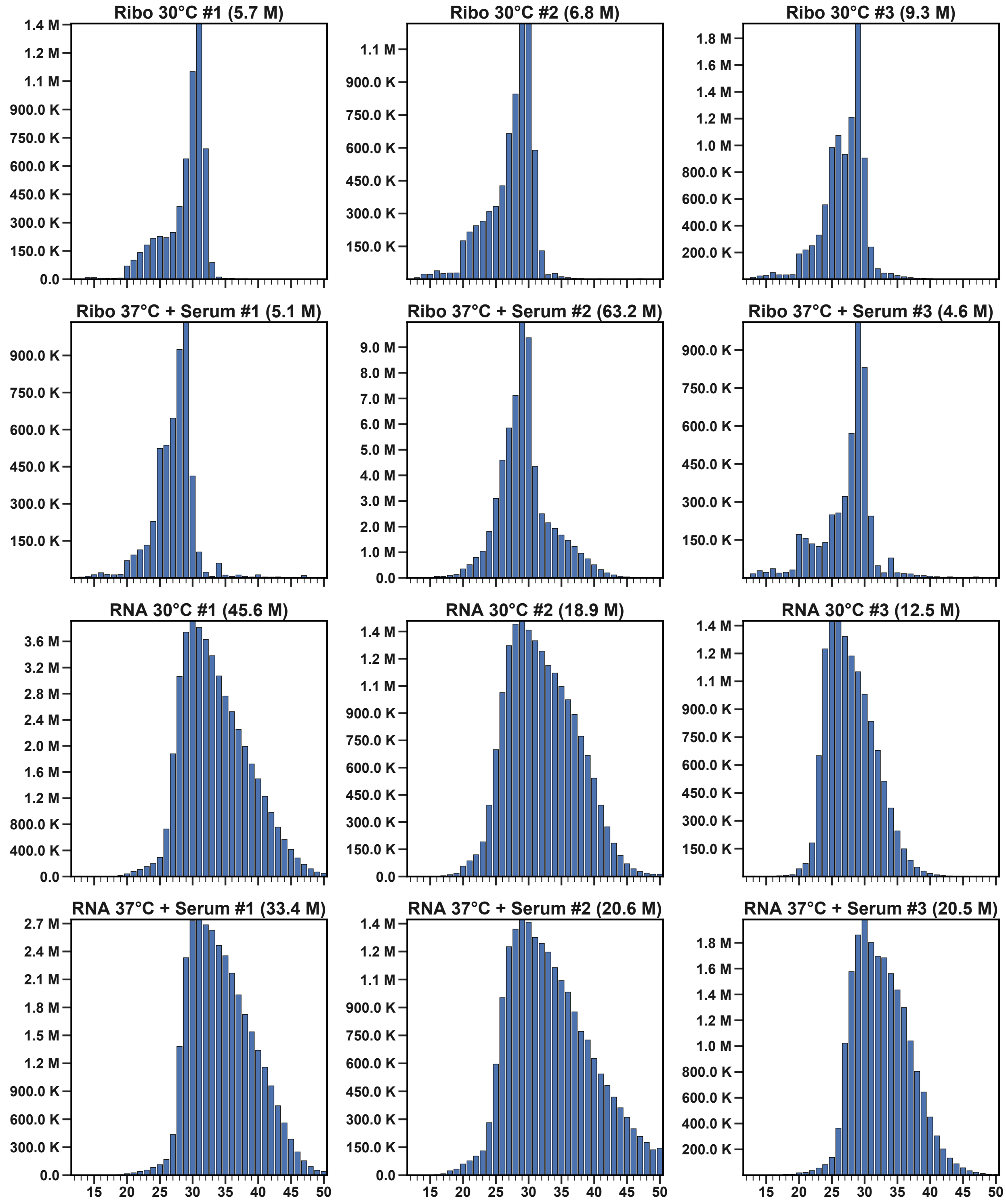


Figure 4.1

Distribution of total reads segregated by read length in Ribo-seq and RNA-seq experiments. Total read counts are displayed in parenthesis.

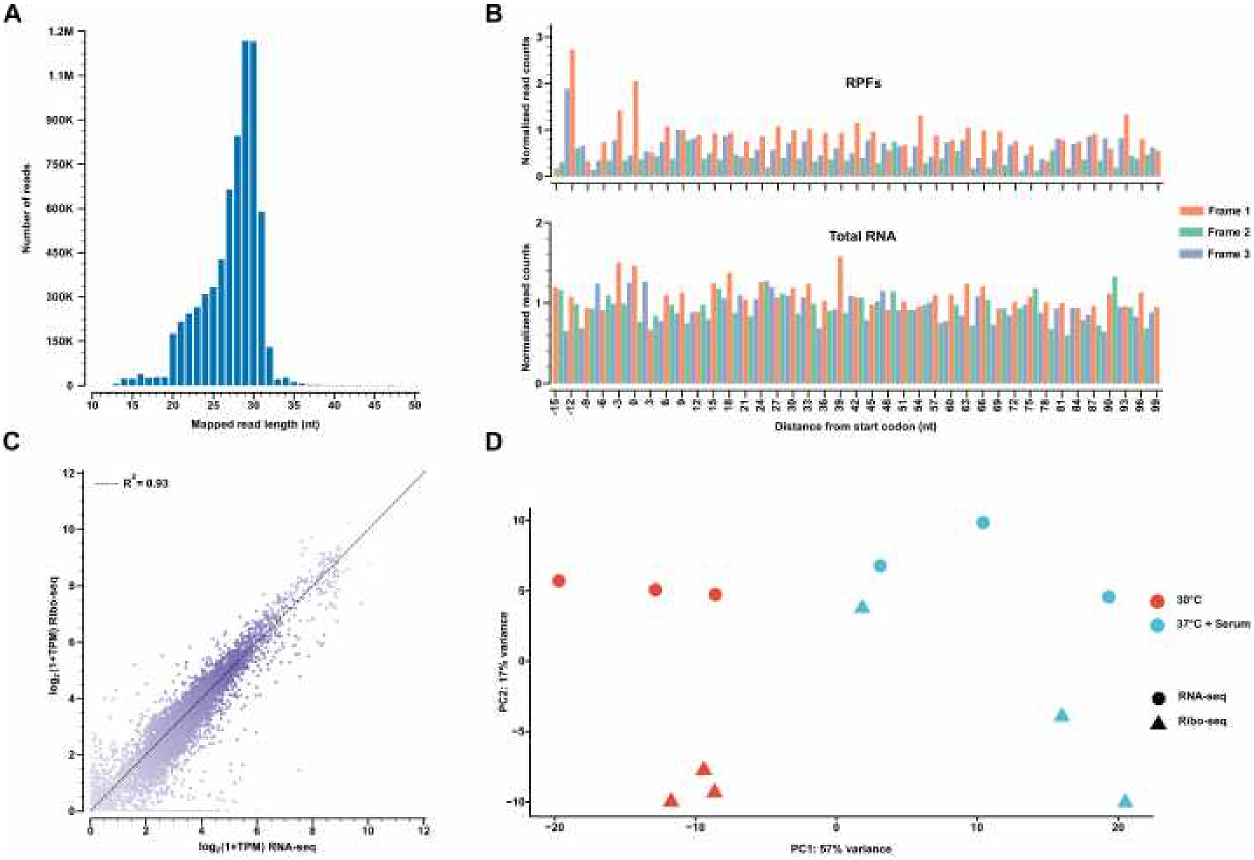


Figure 4.2

Ribosome profiling in *Candida albicans*. A) Read length distribution in a 30 ° C sample. B) Principle component analysis of Ribo-seq and RNA-seq samples in $30°$ C and $37°$ + Serum samples. C) Correlation between read counts measured as mapped transcripts per million (TPM) between a RNA-seq and the corresponding Ribo-seq sample. D) Metagene plot of a ribosome protected fragments (RPF) and RNA-seq fragments.

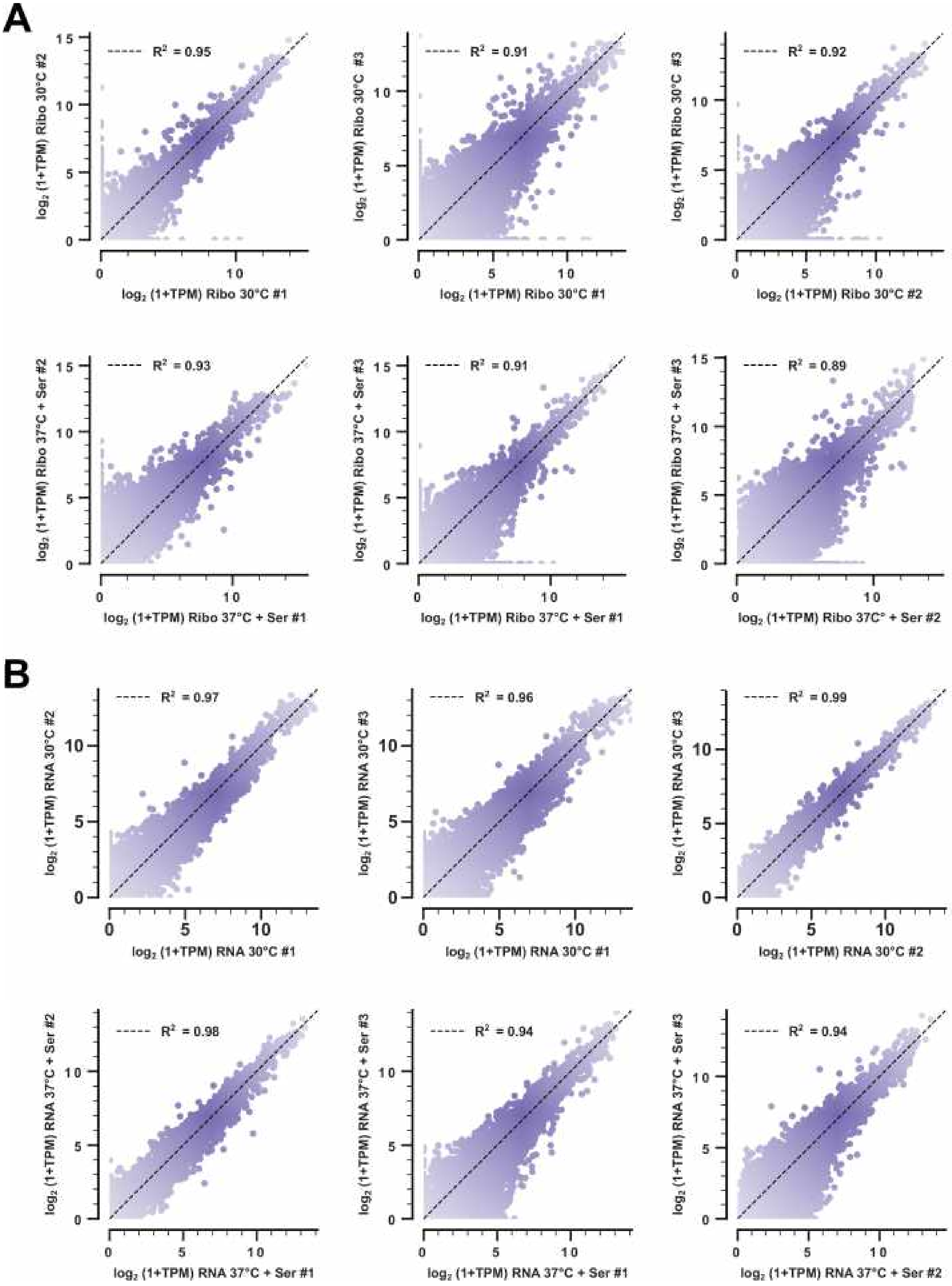


Figure 4.3

Correlation between replicates across conditions. Each row indicates pairwise correlation between the three replicates belonging to A) $30°$C and B) $37°$ C + Serum conditions across RNA-seq and Ribo-seq.

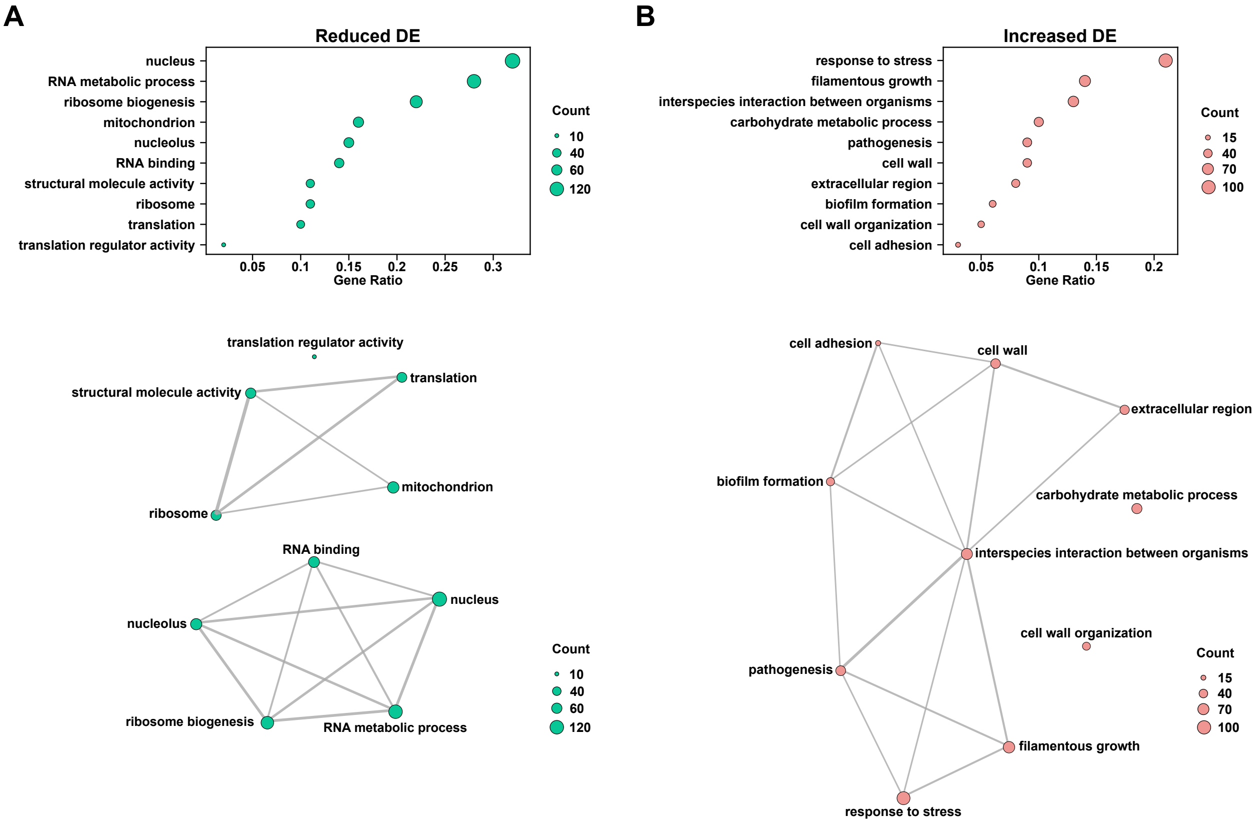


Figure 4.4

Differential expression analysis between 37° C + Serum and 30° C samples. Gene ontology of A) down-regulated and B) up-regulated genes.

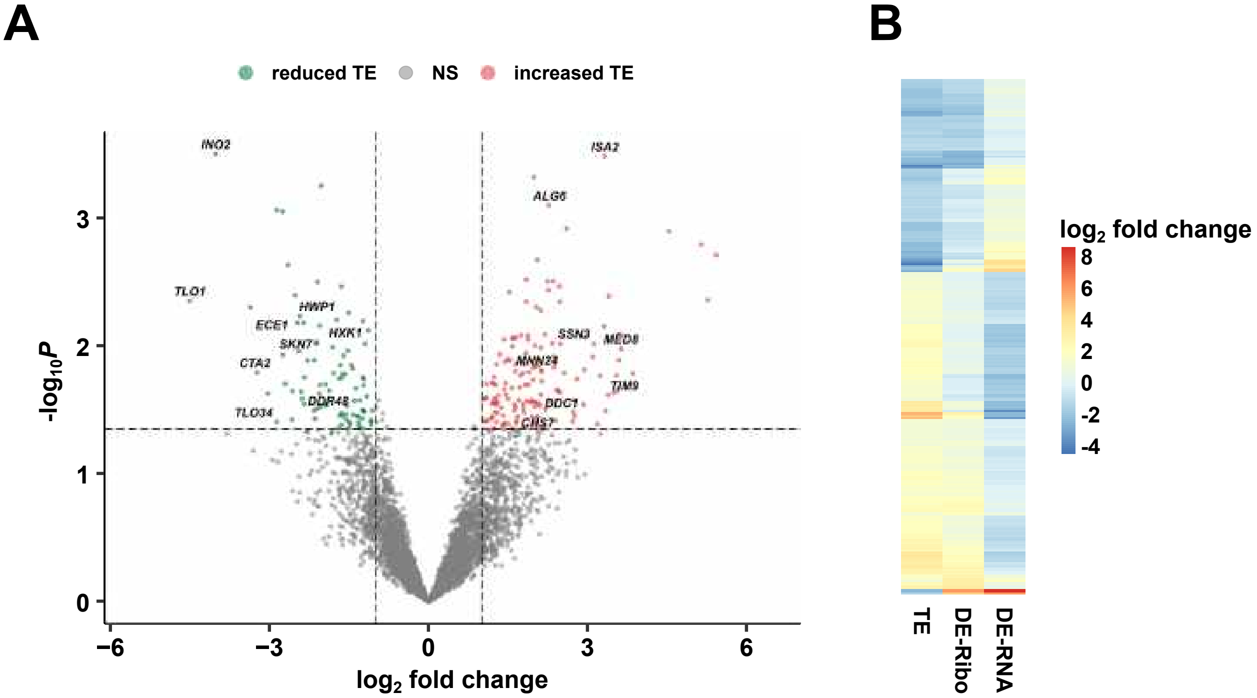


Figure 4.5

Differential translational efficiency between $37°C$ + Serum and $30°$C samples. A) Volcano plot showing the fold change in translational efficiency (TE) and the associated p-value. Differential TE genes are defined with the criteria $\log_2$ fold change > 1 (red, increased TE) or < -1 (green, reduced TE) and $P < 0.05$. B) Heatmap showing fold changes at the TE, RNA-seq and Ribo-seq levels for the genes highlighted in A.

Gene name	Ref. #	Description	Fold change (mRNA)	Fold change (TE)
C3_07980C	orf19.6192	protein of unknown function; Plc1-regulated	7.6	-109.1
TLO1	orf19.7544	member of TLO gene family	15.0	-22.8
C3_05990C	orf19.7380	protein of unknown function	17.8	-13.9
TLO34	orf19.2661	member of TLO gene family	21.4	-7.3
ECE1	orf19.3374	candidalysin, cytolytic peptide toxin essential for mucosal infection	388.0	-5.5
HWP1	orf19.1321	hyphal wall protein, adhesin, host transglutaminase substrate mimic	362.0	-5.4
SKN7	orf19.971	Putative response regulator in phophorelay signal transduction; required for H2O2 resistance	3.7	-4.9
DDR48	orf19.4082	immunogenic stress-associated protein	22.2	-4.5
HXK1	orf19.2154	GlcNAc kinase; required for hyphal growth and virulence	4.2	-4.3
RBT1	orf19.1327	cell wall protein similar to Hwp1; required for virulence	64.4	-2.7
SFU1	orf19.4869	GATA-type transcriptional regulator of iron-responsive genes; promotes gastrointestinal commensalism	2.0	-2.3
SFL2	orf19.3969	transcriptional regulator of morphogenesis	2.0	-2.3
SLK19	orf19.6763	Alkaline-induce plasma membrane protein important for cell wall; required for virulence	2.2	-2.1

Table 4.1
Genes transcriptionally induced during the *C. albicans* morphological transition and involved in pathogenesis and/or virulence-related properties show significantly reduced translational efficiency (TE)

(> 0.8, p-value < 0.05) between Ribo-seq and RNA-seq normalized read counts (Figure 4.2B and 4.3) while the metagene plots are periodic indicating the samples are representative of actively translation (Figure 4.2D).

We used ribotricer [149] to filter out actively translating fragments for all Ribo-seq experiments. Ribotricer exploits the periodicity information in Ribo-seq data to separate out the actively-translating from non-actively translating fragments. In order to identify changes in the translational landscape, we focused on genes showing differential translational efficiency. Translational efficiency is a measure of the rate of mRNA translation into proteins and can be approximated as the ratio of Ribo-seq to the corresponding RNA-seq normalized read counts. We used riborex [150] to identify genes that are differentially translationally efficient during the morphological transition. Gene ontology analysis reveals the genes showing higher translational efficiency are involved in pathogenesis and hyphael growth (Figure 4.5). Some of the key genes showing increased translational efficiency are highlighted in Table 4.2 while Table 4.1 highlights some key genes that are transcriptionally induced but show reduced translational efficiency.

4.2.3.1 Re-annotating the *C. albicans* transcriptome

The transcriptome annotation made available through the Candida Genomes Database (CGD) (http://www.candidagenome.org/) only contains the genomic coordinates of exons and coding domain sequences. The 5' and 3' untranslated regions (UTRs) are currently not annotated in the annotation file (GTF) made available through CGD website. However, Bruno *et al.* [151] provide coordinates for both 5' UTRs and 3' UTRs for Assembly 21.

We expanded the current annotation of exons and coding sequences by also adding UTR coordinates made available by Bruno *et al.* [151]. Since the annotation were made using

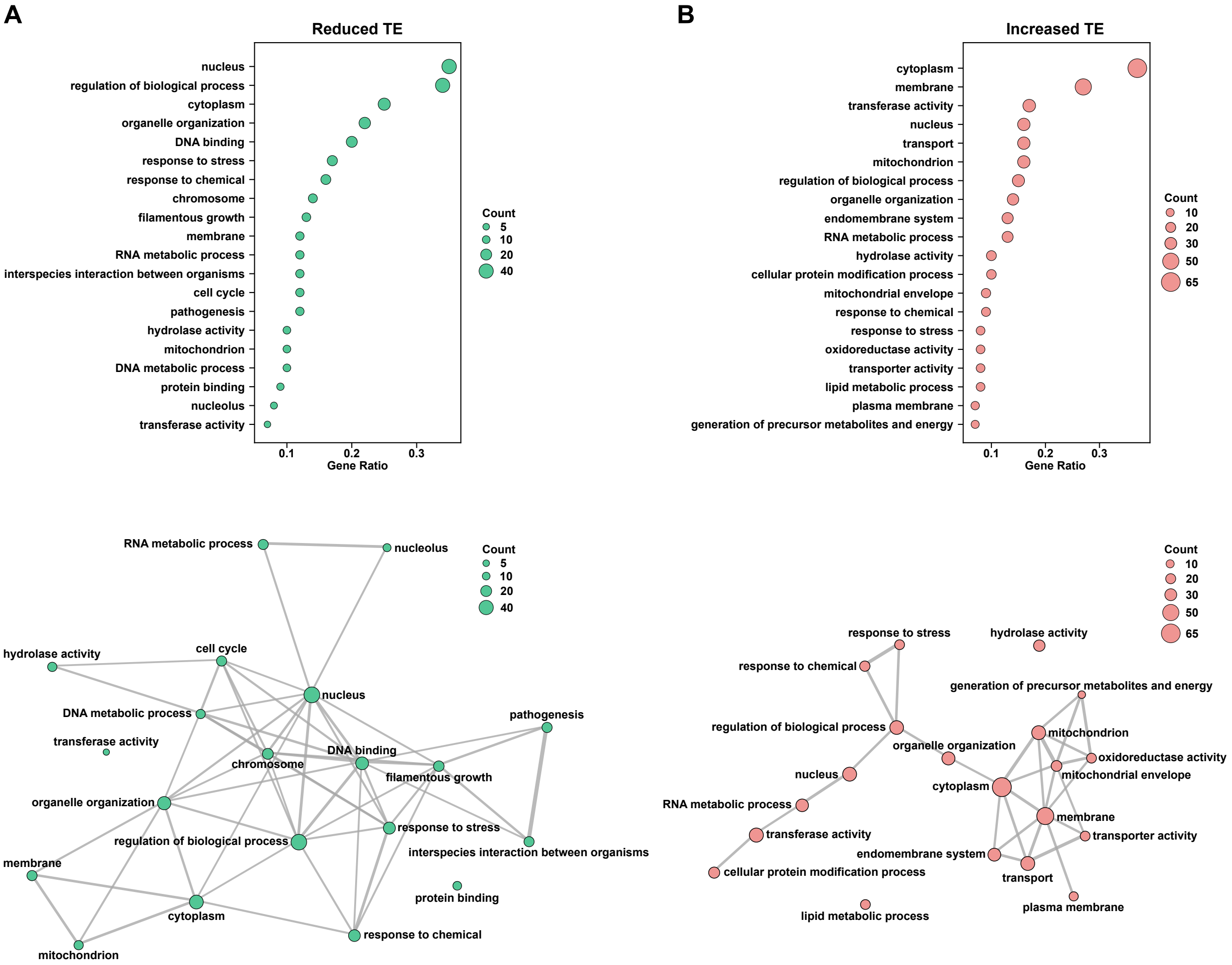


Figure 4.6
Gene ontology analysis of TE genes A) Genes with reduced TE B) Genes with increased TE

Gene name	Ref. #	Description	Fold change (TE)
AAP1	orf19.2810	putative amino acid permease; fungal-specific	40.5
FET3	orf19.4213	multicopper oxidase	33.8
SAP98	orf19.852	GPI-anchored aspartic endopeptidase	32.7
CDC14	orf19.4192	Protein involved in exit from mitosis and morphogenesis	26.2
SAP7	orf19.756	pepstatin A-insensitive secreted aspartyl protease	22.5
HGT20	orf19.1587	putative glucose transporter of the major facilitator superfamily	13.3
MED8	orf19.4497	ortholog of RNA Polymerase II Mediator complex component	12.3
GNP3	orf19.7565	putative high-affinity glutamine permease; fungal-specific	9.8
UEC1	orf19.4646	protein required for oral epithelial cell damage, hyphal growth and stress resistance	8.7
MAL31	orf19.3981	putative high-affinity maltose transporter	8.5
SSU1	orf19.7313	protein similar to S. cerevisiae Ssu1 sulfite transporter; important for filamentous growth	5.9
LIP6	orf19.4823	secreted lipase	5.8
BMT8	orf19.860	putative β-mannosyltransferase	5.7
ALG6	orf19.1843	putative glucosyltransferase involved in cell wall mannan biosynthesis	4.8
HGT5	orf19.6005	putative glucose transporter of the major facilitator superfamily	4.7
PSA2	orf19.4943	mannose-1-phosphate guanyltransferase	4.6
MNN24	orf19.1995	α-1,2-mannosyltransferase; required for normal cell wall mannan content	4.3
CHS7	orf19.2444	Protein required for chitin synthase III activity	4.1
HYR1	orf19.4975	GPI-anchored hyphal cell wall protein; important for resistance to killing by neutrophils, azoles	4.1
HMS1	orf19.921	hLh domain Myc-type transcription factor required for morphogenesis	3.6
TPO5	orf19.151	putative polyamine transporter	3.5
OPT8	orf19.5770	oligopeptide transporter	3.5
CEK1	orf19.2886	ERK-family protein kinase; required for yeast-hyphal switch, mating efficiency and virulence	3.5
PTC8	orf19.4698	predicted type 2C protein phosphatase; required for hyphal growth	3.1
RPD3	orf19.2834	histone deacetylase; regulates white-opaque switching	3.1
BEM1	orf19.4645	protein required for budding, hyphal growth and virulence	2.6
RAX2	orf19.3765	plasma membrane protein involved in establishment of bud sites and linear direction of hyphal growth	2.4
SMF12	orf19.2270	manganese transporter	2.4

Table 4.2
Selected genes showing significantly increased translational efficiency (TE) during the *C. albicans* morphological transition.

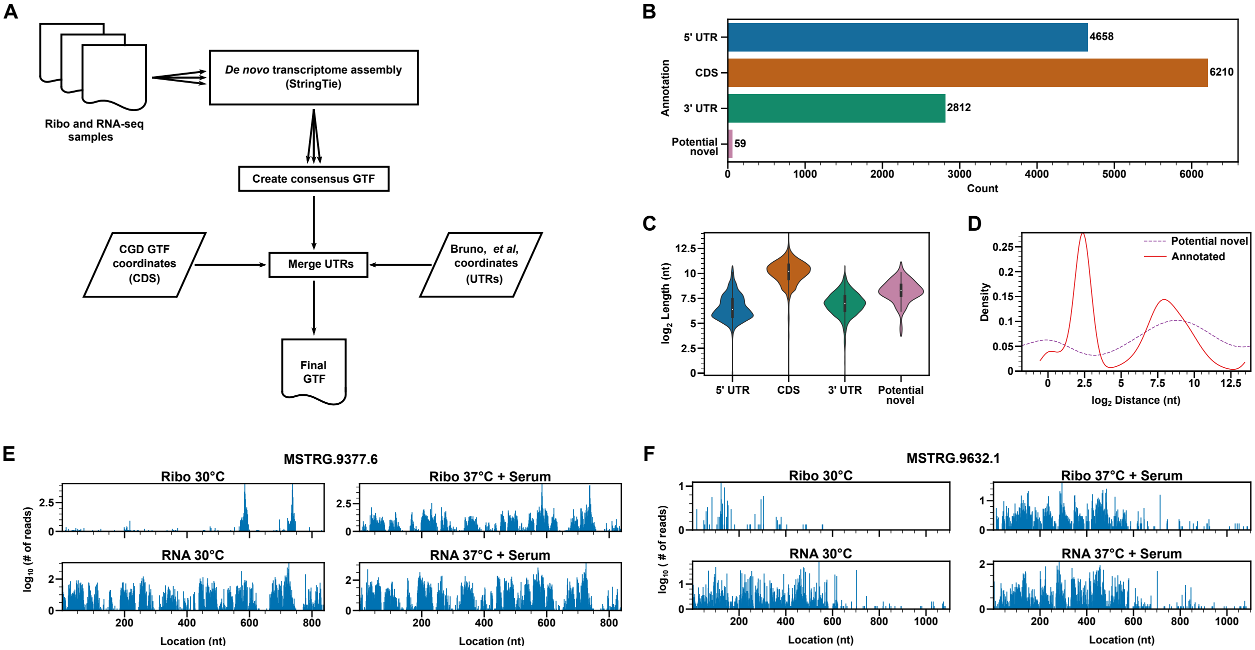


Figure 4.7
Ribo-seq enables novel ORF discovery. A) Workflow for re-annotating transcriptome of *Candida albicans.* B) Distribution of regions in the final annotation as obtained from workflow in A. Novel indicates exons that were not annotated in the original annotation. C) Distribution of sizes of 5' UTR, 3'UTR, CDS and novel exons. D) Distribution of distances of novel exons (purple, dashed) to known exons and known exons to other known exons (red). E) A novel exon with active transcription and translation in both $30°$C and $37°$C + Serum samples. F) A novel exon with active transcription with a strong Ribo-seq signal at a stop codon.

Assembly 21 while the current Assembly is 22, we lifted over the coordinates to the latest assembly using the liftOver [152] tool. The chain file was obtained from CGD. The lifted over candidates were further filtered to ensure there was no overlap between the existing coding domain sequences. This resulted in a total of 9084 5' UTRs, 5464 3' UTRs for a total of 12389 coding domain sequences for the diploid assembly (Figure 4.7B).

4.2.4 Detecting novel exons and potential ORFs

We hypothesized that using our deep-sequenced Ribo-seq and RNA-seq samples we could potentially discover “novel” exons and open reading frames given that the transcriptome annotation of *C. albicans* is currently incomplete. In order to discover novel exons, we performed a guided *denovo* assembly using our RNA-seq and Ribo-seq samples across the two conditions (Figure 4.7A). In particular, we used StringTie [153] along with the annotation file from CGD (r27 GTF) as the guide annotation (`-G ref.GTF`). For each input, StringTie outputs a new GTF which represents a super-set of transcripts and exons annotated in the guide GTF consisting of additional exons that are unannotated in the guide GTF. We then created a consensus catalogue of novel exons by looking for overlapping regions of the novel exons from all the new GTFs. Using our strategy we were able to discover additional 71 exons (Figure 4.7) that were previously unannotated.

The novel exons have smaller size distribution as to the previously annotated with the median length of novel exons being 315 nucleotides while that of previously annotated exons around 1187 nucleotides (Figure 4.7C). It is possible that the novel exons remained unannotated because they are smaller in size and hence harder to detect through any technique that relies on relative abundance of transcripts for annotating regions. It was also likely that these novel exons were essentially an extension of the previously known exons which could happen if the previously annotated exons were shorter than the actual exons. In order to rule out such a scenario, we calculated the distance of these novel exons to the previously unannotated exons. The distribution of these distances is similar to the distance of any known exon to any other known exon (Figure 4.7D).

Next, we used ribotricer to detect translation in novel exons. All of the 71 exons found have a potential open reading frame (ORF). Though all the ORFs show an accumulation of RPFs, ribotricer is able to find evidence of active translation in only one of them (Figure 4.8D).

4.2.5 Detecting pausing sites

Translation elongation is the most dynamic stage of translation when the ribosome scans one codon at a time and decodes it, resulting in a new amino acid that gets added onto the nascent peptide chain. Translation progresses at the speed of $\sim$ 6 amino acids per second [154, 74], but this speed is not constant. Ribosomes may slow down or speed up at certain codons during elongation to regulate protein synthesis. There is emerging evidence for the elongation process to be regulatory. In particular, it has been shown to be critical for early development [155], functioning of neurons [156, 157] and cancer [158]. Ribo-seq [24] provides nucleotide-level resolution of the ribosomes attached to mRNA. It has been used to demonstrate the slow elongation rates at proline [74, 64] , translation inhibition caused by a drug compound inducing sequence dependent stalling [1] and heat shock induced pausing [28].

We developed a new method to identify transcriptome-wide stalling sites using Ribo-seq data. Briefly, the method relies on smoothing the Ribo-seq profile and then locates peak pileups in the smooth profile (See methods). Applying this method to our data revealed stalling in two genes: an adhesion protein ALS1 and the 60S ribosomal protein RPL11 (Figure 4.8). Both the pausing sites are reproducible across all the three replicates of both the conditions.

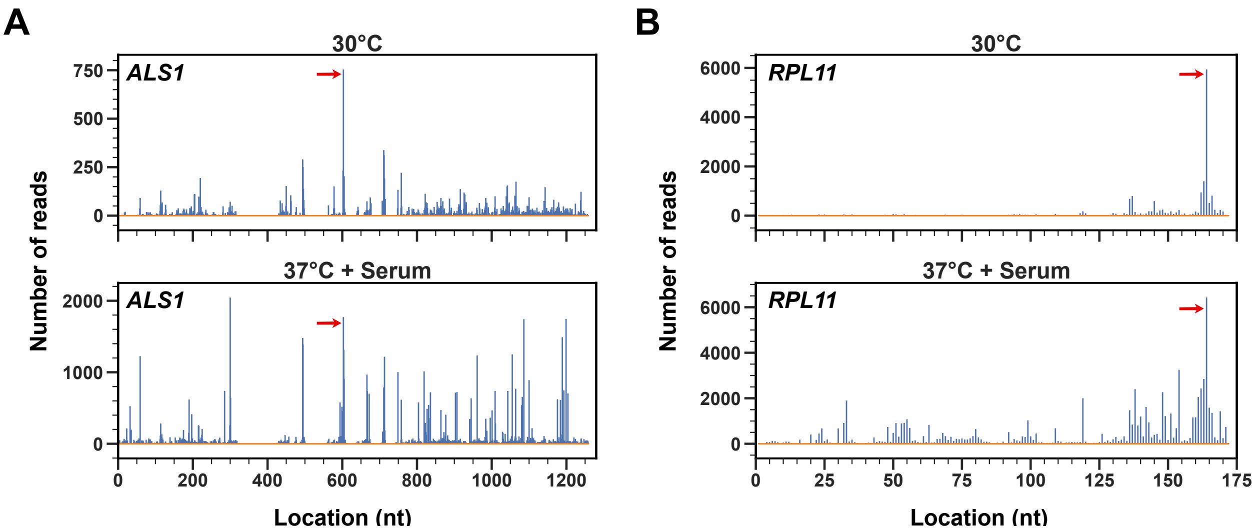


Figure 4.8

Pausing in *Candida albicans*. A) A reproducible pausing site in both 30°C and 37°C $+$ Serum conditions. A) ALS1 and B) RPL11.

4.3 Methods

4.3.1 RNA-seq and Ribo-seq data analysis

The quality of raw sequences reads from RNA-Seq and Ribo-Seq datasets were assessed using FastQC [159]. Adaptor sequences and low-quality score (phred quality score < 5) bases were trimmed from RNA-Seq and Ribo-Seq datasets with TrimGalore (v0.4.3) [160]. Trimmed sequences from RNA-seq and Ribo-seq were both mapped using STAR (v.2.5.2b) [95] using Assembly 22 fasta as the reference and GTF r27 from CGD allowing a mismatch of at most two positions. All the reads mapping to rRNA and tRNA sequences were filtered out before downstream analysis.

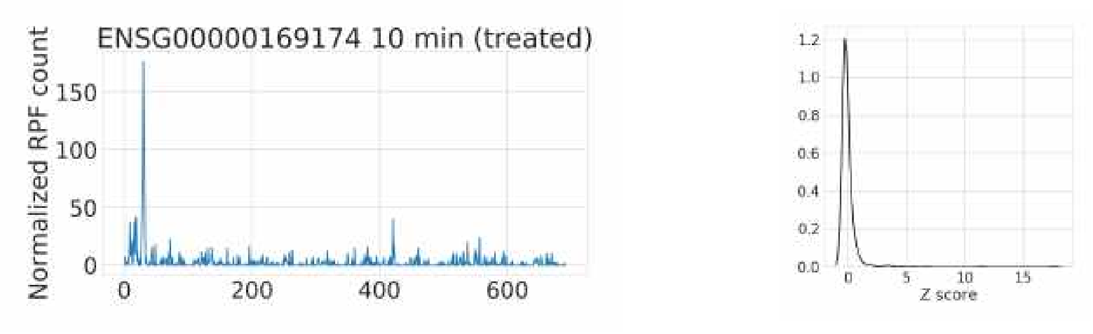
4.3.2 Differential expression and translational efficiency analysis

Differential expression analysis was performed using DESeq2 [147]. We filtered out genes with at least one read per replicate before doing the library size normalization and running the moderated t-statistic test. Genes are said to be differentially expressed if their absolute fold-change on $\log_2$ scale is at least 1 and the FDR adjusted p-value is at least 0.05. Gene ontology analysis was performed using clusterProfiler [161] using GO slim ontology file available from CGD.

Differential translational efficiency was performed using riborex [150]. Only genes that had a a read count of at least one per replicate were used as input to riborex. We define genes to be exhibiting differential translational efficiency if their absolute fold-change on $\log_2$ scale is at least 1 and the non-adjusted p-value is less than 0.05.

4.3.3 Detecting pausing sites

In Ribo-seq data, stalling site appears as sharp peaks [58]. However, the heterogeneity and the sparsity in data deems this task particularly challenging. A näive approach of identifying pausing sites using Z-scores [60] results in too many high-positives. In scenarios where ribosomal pausing leads to queuing up of ribosomes near the start codon site, the downstream profile might end up with fewer ribosomes due to some ribosomes getting dropped off at the paused site [28]. The $Z-$score approach ignores the trend effects that could arise from such scenarios besides ignoring the heterogeneity of the data. Moreover, the thresholds are arbitrarily defined (25 in [74] and 2 in [162]). In our experiments with real data, the $Z-$score distribution of the normalized RPF counts appears to be long-tailed (Figure 4.9b).



(a)
RPF profile of PCSK9

(b)
$Z-$scores

Figure 4.9
RPF profile of PCSK9 gene from data in Lintner *et al.* [1]. (a) The pausing signal observed at around 65^{th} codon in the form of a sharp peak is a *true* pausing site. (B) The associated $Z-$ score has a long tailed distribution, thus arbitrary thresholds leads to lot of false positives.

Given the heterogeneous nature of read counts in Ribo-seq profile, we considered smoothing the read counts based on adjacent read counts, the motivation being reducing the heterogeneity by pooling observations. The smoothed counts can then be passed onto a *peak finding* method such as even the $Z-$score based approach, provided the resulting distribution is near gaussian.

We propose using a de-noising approach using Savitzky–Golay filter [163] that has been applied to a wide range of problems involving signals of similar nature [164]. Savitzky-Golay filter finds a low-degree polynomial fit over adjacent points by the method of linear least squares. Post filtering, we use $Z-$scores to identify peaks at sub-codon resolution.

Savitzky-Golay filter is a digital filter that acts as a low-pass filter for smoothing the data. It increases the signal to noise ratio without distorting the signal overall. This is achieved by convolution, where in sub-sets of adjacent data points are fitted with a low-degree polynomial by the method of linear least squares. An analytical solution exists for finding the solution to least-squares problem if the data points are equally spaced in the form of "convolution coefficients".

Let $Y_{1:T} = \{y_1, y_2 \dots, y_T\}$ represent the profile of read counts over T codons. On treating them with m convolution coefficients the transformed points are

$$y'_t = \sum_{j=\frac{1-m}{2}}^{\frac{m-1}{2}} C_j y_{t+i} \qquad\qquad \frac{m-1}{2} \leq t \leq T - \frac{m-1}{2}$$

The coefficients C_j are analytically derived as follows. Consider a modified variable z such that given codons $\{1, 2, \dots, T\}$ such that $z = t - \bar{t}$ where $\bar{t} = \frac{(T+1)}{2}$.

Fitting a polynomial of degree k,

$$y = a_0 + a_1 z + a_2 z^2 + \dots + a_k z^k.$$

The coefficients a_i are solved using least-squares approximation,

$$\mathbf{a} = (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T y,$$

where i^{th} row of $\mathbf{J}$ is given by $(1, z_i, z_i^2, \dots, z_i^k)$. Hence the coefficients C_j above are given by $\mathbf{C} = (\mathbf{J}^T \mathbf{J})^{-1} \mathbf{J}^T$.

For each possible ORF and the corresponding Ribo-seq profile as obtained from ribotricer, we apply the Savitzky-Golay filter to this profile. Peaks are then called such that the called site has a signal to noise ratio above 2.5 where the noise is estimated by fitting a single variance parameter for the entire profile. For each such peak the corresponding p-value is calculated a gaussian distribution whose mean and variance are empirically estimated from the given profile. The p-values are corrected for multiple testing using Benjamini-Hochberg procedure [165].

Chapter 5

Integrated analyses of early responses to radiation in glioblastoma

5.1 Introduction

Glioblastoma is the most common intracranial malignant brain tumor with an aggressive clinical course. Standard of care entails maximally safe resection followed by radiotherapy with concomitant and adjuvant temozolomide. Nonetheless, the median overall survival remains approximately 16 months [166, 167], and the recent addition of tumor-treating fields to the standard of care has only increased median overall survival to 20.5 months [166]. Recurrence occurs in part because glioblastoma uses sophisticated cellular mechanisms to repair DNA damage from double-stranded breaks caused by ionizing radiation, specifically homologous recombination and non-homologous end-joining. Thus, the repair machinery confers a mechanism for resistance to radiation therapy. Ionizing radiation can also cause base damage and single-strand breaks, which are repaired by base excision and single-strand break repair mechanisms, respectively [168]. A comprehensive analysis of molecular mechanisms driving resistance to chemotherapy and radiation is required to surpass major barriers and advance treatments for glioblastoma.

The Cancer Genome Atlas (TCGA) was instrumental in improving the classification and identification of tumor drivers [169], but its datasets provide limited opportunities to investigate radiation response. Thus, studies using cell and murine models are still the best alternatives to evaluate radiation response at the genomic level. The list of biomarkers associated with radiation resistance in glioblastoma is still relatively small. Among the most relevant are FOXM1 [170, 171], STAT3 [171], L1CAM [172], NOTCH1 [173], RAD51 [174], EZH2 [175], CHK1/ATR [176], COX-2 [177], and XIAP [178]. Dissecting how gene expression is altered by ionizing radiation is critical to identify possible genes and pathways that could increase radio-sensitivity. A few genomic studies [179, 180, 181] have explored this question, but these analyses were restricted to describing changes in transcription.

Gene expression is regulated at multiple levels, and RNA-mediated mechanisms such as splicing and translation are particularly relevant in cancer biology. A growing number of inhibitors against regulators of splicing and translation are being identified [182]. Splicing alterations are a common feature across cancer types and affect all hallmarks of cancer [183]. Numerous splicing regulators display altered expression in glioblastoma (e.g. PTBP1, hnRNPH, and RBM14) and function as oncogenic factors [184]. Importantly, a genome-wide study using patient-derived models revealed that transformation-specific depended on RNA splicing machinery. The SF3b-complex protein PHF5A was required for glioblastoma cells to survive, but not neural stem cells (NSCs). Moreover, genome-wide splicing alterations after PHF5A loss appear only in glioblastoma cells [185]. Translation regulation also plays a critical role in glioblastoma development. Many translation regulators such as eIF4E, eEF2, Musashi1, HuR, IGF2BP3, and CPEB1 promote oncogenic activation in glioblastoma, and pathways linked to translation regulation (e.g., mTOR) promote cancer phenotypes [186].

To elucidate expression responses to radiation, we conducted an integrated study in U251 and U343 glioblastoma cell lines covering transcription (mRNAs and lncRNAs), splicing, and translation. We determined that the downregulation of FOXM1 and members of the E2F family are likely the major drivers of observed alterations in cell cycle and DNA replication genes upon radiation exposure. Genes involved in RNA regulatory mechanisms were particularly affected at the transcription, splicing, and translation levels. In addition, we identified several oncogenic factors and genes associated with poor survival in glioblastoma that displayed increased expression upon radiation exposure. Importantly, many have been implicated in radio-resistance, and therefore, their inhibition in combination with radiation could increase therapy efficacy.

5.2 Methods

5.2.1 Cell culture and radiation treatment

U251 and U343 cells were obtained from the University of Uppsala (Sweden) and maintained in Dulbecco's Modified Eagle Medium (DMEM, Hyclone) supplemented with 10% fetal bovine serum, 1% Penicillin/Streptomycin at 37°C in 5% $\mathrm{CO_2}$-humidified incubators and were sub-cultured twice a week. Cells were plated after appropriate dilution, and ionizing radiation treatment was performed on the next day at a dose of 5 Gray (Gy). A cabinet X-ray system (CP-160 Cabinet X-Radiator; Faxitron X-Ray Corp., Tucson, AZ) was used. After exposure to ionizing radiation, cells were cultured for 1 and 24 hours (hrs).

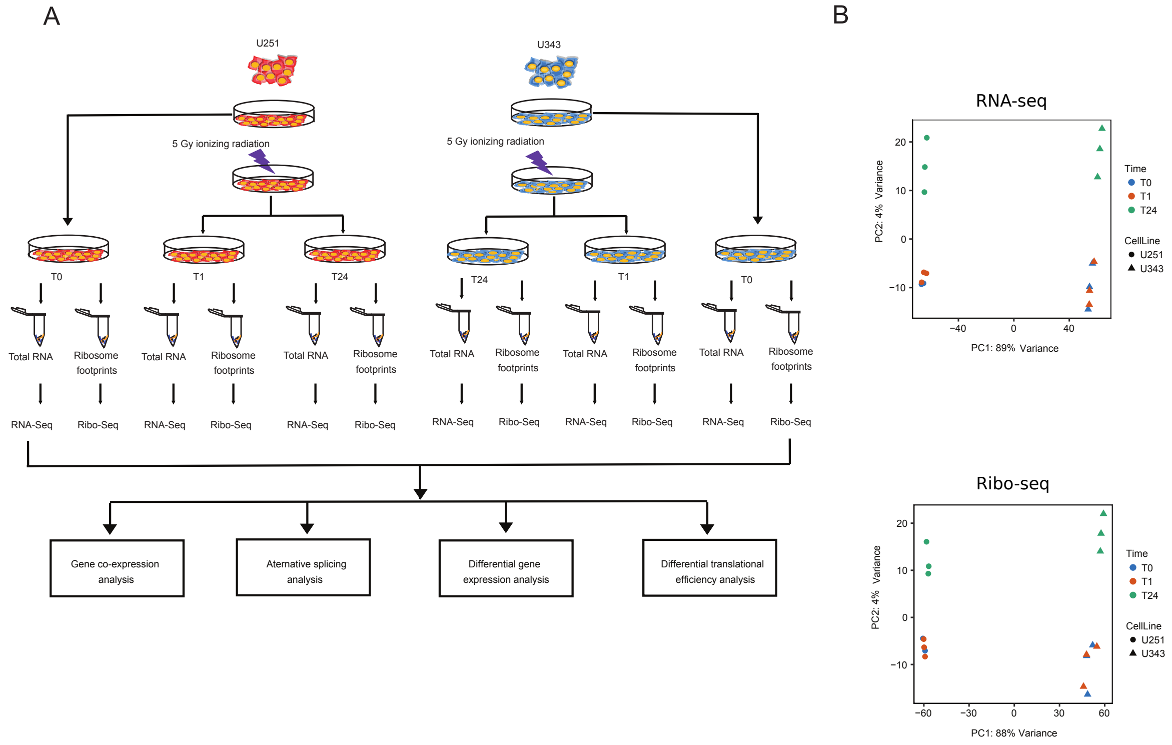


Figure 5.1

Experimental design and Principal Component Analysis of RNA-Seq and Ribo- Seq data from glioblastoma cell lines. A) Schematic representation of experimental protocol followed for radiation exposure of glioma cell lines, and sample preparation for sequencing the RNA and ribosome footprints. B) Principal component analyses performed on normalized log- transformed read counts of RNA-Seq and Ribo-Seq datasets.

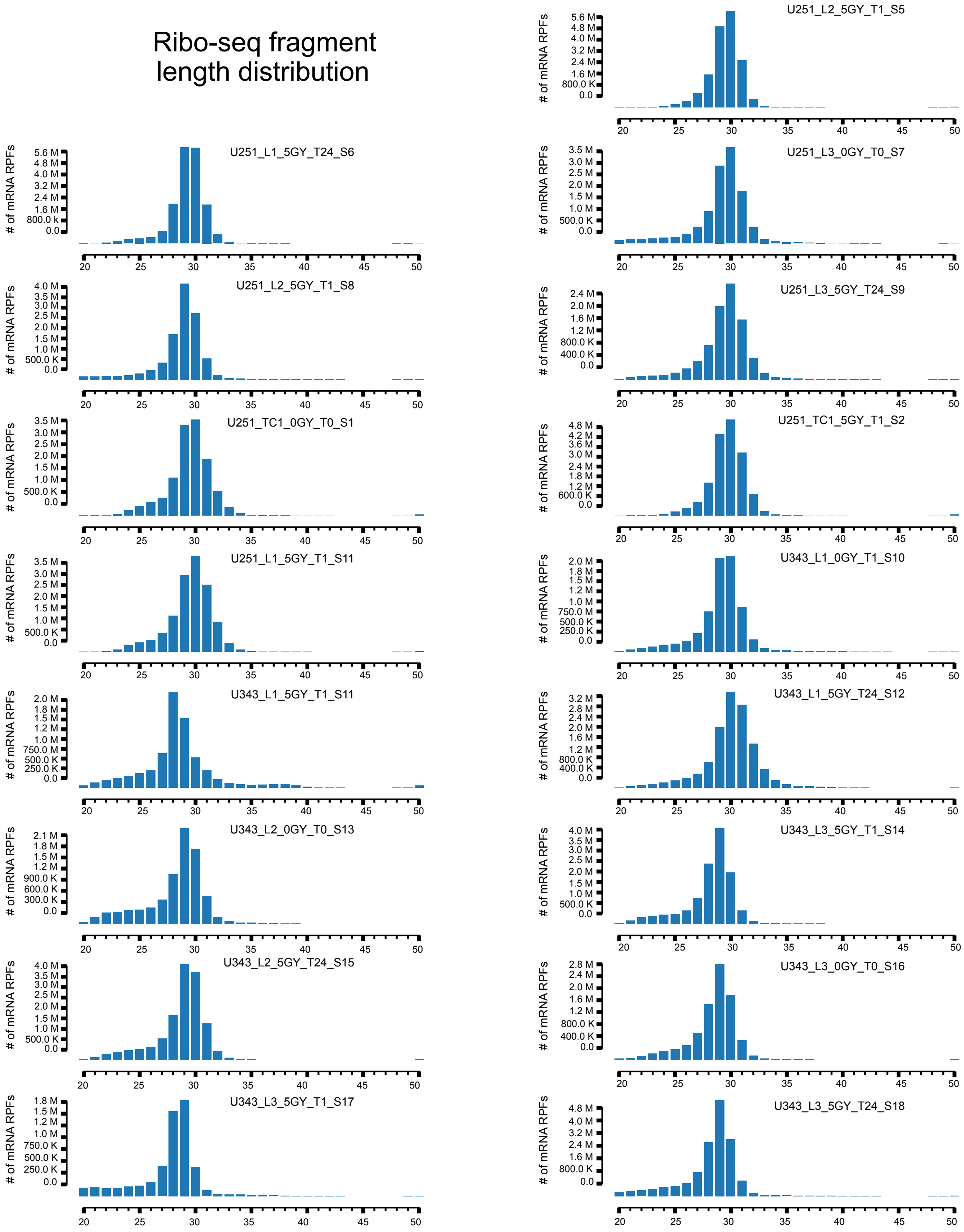


Figure 5.2

Fragment length distribution of ribosome footprints of glioblastoma cell lines. Fragment lengths distribution was obtained using ribotricer.

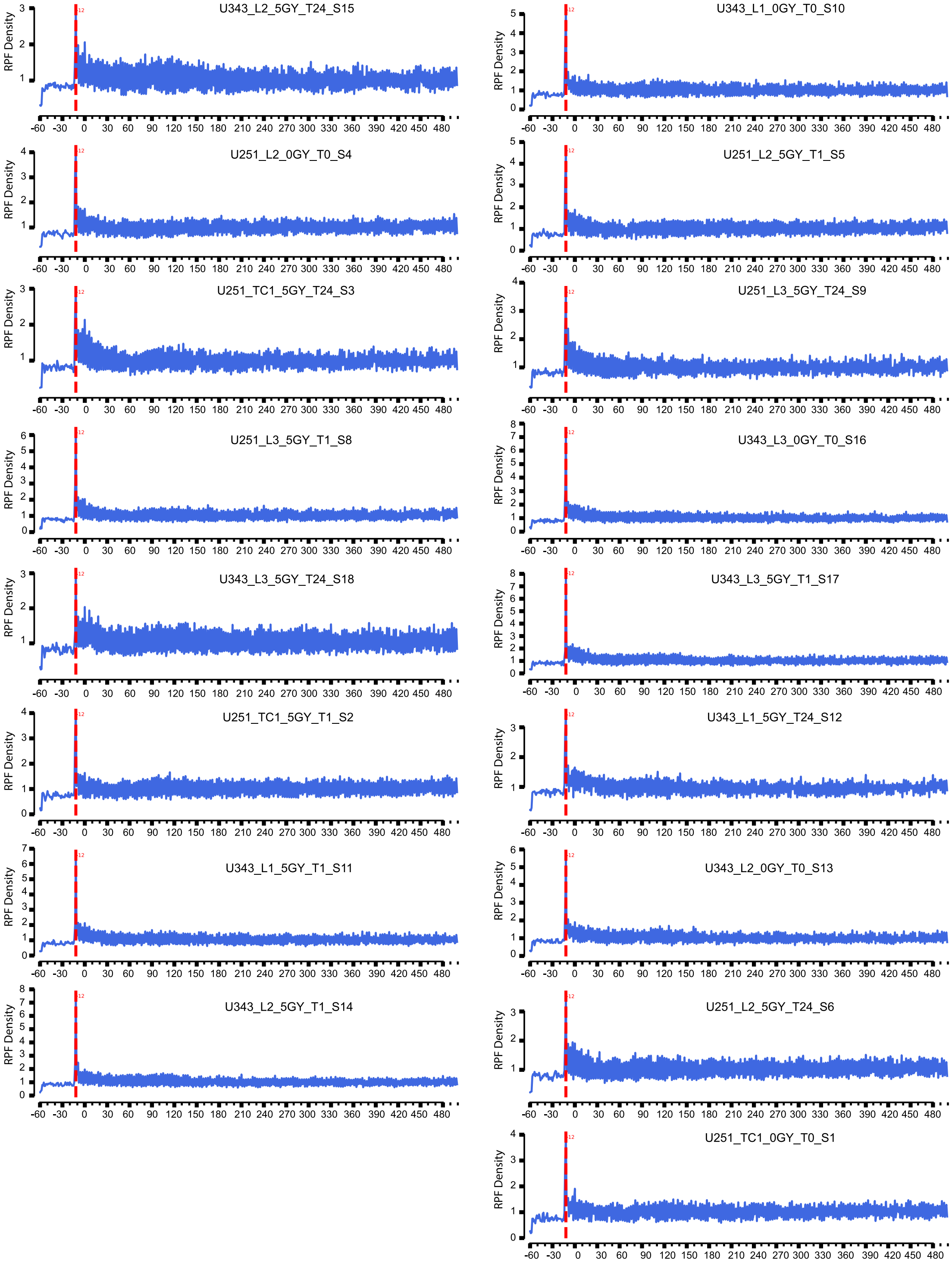


Figure 5.3
The ribosome density profiles of glioblastoma cell lines. The metagene distribution was obtained using ribotricer.

5.2.2 RNA preparation, RNA-seq and Ribosome Profiling (Ribo-seq)

RNA was purified using a GeneJet RNA kit from Thermo Scientific. The TruSeq Ribo Profile (Mammalian) kit from Illumina was used to prepare material for ribosome profiling (Ribo-seq). RNA-seq and Ribo-seq samples were prepared according to Illumina protocols and sequenced at UTHSCSA Genome Sequencing Facility.

5.2.3 Overall strategy to identify gene expression alterations upon radiation

To identify the most relevant expression alterations in the early response to radiation, we analyzed samples from U251 and U343 cells collected at 0 (T0), 1 (T1), and 24 (T24) hours post-radiation. To capture the progressive dynamics of expression alterations, we compared T0 to T1 samples and T1 to T24 samples. Our strategy to identify the most relevant alterations in expression with maximal statistical power was to combine all samples and use a design matrix with cell type defined as a covariate with time points (Figure S1).

5.2.4 Sequence data pre-processing and mapping

The quality of raw sequences reads from RNA-Seq and Ribo-Seq datasets were assessed using FastQC [159]. Adaptor sequences and low-quality score (phred quality score < 5) bases were trimmed from RNA-Seq and Ribo-Seq datasets with TrimGalore (v0.4.3) [160]. The trimmed reads were then aligned to the human reference genome sequence (Ensembl GRCh38.p7) using STAR aligner (v.2.5.2b) [95] with GENCODE [187] v25 as a guided reference annotation, allowing a mismatch of at most two positions. All the reads mapping to rRNA and tRNA sequences were filtered out before downstream analysis. Most reads in the Ribo-seq samples mapped to the coding domain sequence (CDS). The distribution of fragment lengths for ribosome footprints was enriched in the 28-30 nucleotides range, as expected (Figure S2). The ribosome density profiles exhibit high periodicity as within the CDS, as expected since ribosomes traverse three nucleotides at a time (Figure S3). The periodicity analysis was performed using ribotricer [149]. The number of reads assigned to annotated genes included in the reference genome was obtained by htseq-count [188].

5.2.5 Differential gene expression analysis

For differential expression analysis, we performed counting over exons for the RNA-seq samples. For translational efficiency analyses, counting was restricted to the CDS. A Principal Component Analysis (PCA) was then performed on RNA-Seq and Ribo-Seq data from U251 and U343 cells. Most variation was explained by the cell type along the first principal component, and radiation time-related changes were captured along the second principal component (Supplementary Figure S1B). Differential gene expression analysis was performed by employing the DESeq2 package [147], with read counts from both U251 and U343 cell samples as inputs. We adjusted p-values controlling for the false discovery rate (adjusted p-value) using the Benjamini and Hochberg (BH) procedure [165]. Differentially expressed genes were defined with an adjusted p-value < 0.05

5.2.6 Weighted gene co-expression network analysis

Weighted Gene Co-expression Network Analysis

(WGCNA) [189] uses pairwise correlations on expression values to identify genes significantly co-expressed across samples. We used this approach to identify gene modules with significant co-expression variations as an effect of radiation. The entire set of expressed genes, defined here as those with one or higher transcripts per million higher (TPM), followed by variance stabilization) from U251 and U343 samples were clustered separately using the signed network strategy. We used the Z_{summary} [190] statistic as a measure of calculating the degree of module preservation between U251 and U343 cells. Z_{summary} is a composite statistic defined as the average of the density and connectivity based statistic. Thus, both density and connectivity are considered for defining the preservation of a module. Modules with $Z_{\text{summary}} > 5$ were considered as significantly preserved. The expression profile of all genes in each co-expression module can be summarized as one "eigengene". We used the eigengene-based connectivity (kME) defined as the correlation of a gene with the corresponding module eigengene to assess the connectivity of genes in a module. The intramodular hub genes were then defined as genes with the highest module membership values (kME $>$0.9). All analysis was performed using the R package WGCNA. The protein-coding hub genes were then selected for gene ontology enrichment analysis.

5.2.7 Translational efficiency analysis

We used Riborex [150] to perform differential translational efficiency analysis. The underlying engine selected was DESeq2 [147]. DESeq2 estimates a single dispersion parameter per gene. However, RNA-Seq and Ribo-Seq libraries can have different dispersion parameters owing to different protocols. We estimated the dispersion parameters for RNA-Seq and Ribo-Seq samples separately and found them to be significantly different (mean difference $= 0.04$, p-value $<$ $2.2e-16$). This leads to a skew in translational-efficiency p-value distribution since the estimated null model variance for the Wald test is underestimated. To address this issue, we performed a p-value correction using fdrtool [191] that re-estimates the variance using an empirical bayes approach.

5.2.8 Alternative splicing analysis

Alternative splicing analysis was performed using rMATS [192]. All reads were trimmed using cutadapt [193] with parameters (`-u -13 -U -13`) to ensure trimmed reads had equal lengths (138 bp). rMATs was run with default parameters in paired end mode (-t paired) and read length set to 138 bp (`-len 138`) using GENCODE GTF (v25) and STAR index for GRCh38.

5.2.9 Gene ontology (GO) and pathway enrichment analysis

To classify the functions of differentially enriched genes, we performed GO enrichment, and the Reactome pathway [194] analysis using Panther [195]. For both analyses, we considered terms to be significant if BH adjusted p-values weree < 0.05, and fold enrichment is > 2.0. Further, we used REVIGO [2] to reduce redundancy of the enriched GO terms and visualize the semantic clustering of the identified top-scoring terms. We used STRING database (v10) [3] to construct protein-protein interaction networks and determine associations among genes in a given dataset. The interactions are based on experimental evidence procured from high-throughput experiments text-mining, and co-occurrence. Only high-confidence (0.70) nodes were retained.

5.2.10 Expression correlation analysis

Gene expression correlation analysis was done using Gliovis [4] using glioblastoma samples (RNAseq) from the TCGA. To select correlated genes, we used Pearson correlation, $R > 0.3$, and p-value < 0.05. A list of genes affecting survival in glioblastoma was downloaded from GEPIA [196]. A list of long non-coding RNAs (lncRNAs) implicated in glioma development was obtained from Lnc2Cancer [197]. Drug-gene interactions were identified using the Drug-Gene Interaction Database [198].

5.2.11 Availability of data and materials

The processed data of read abundance matrices is available through GEO accession GSE141013. Scripts for differential expression analysis and translational efficiency analysis are available at `https://github.com/saketkc/2019_radiation_gbm`

- Table S1
- Table S2
- Table S3
- Table S4
- Table S5
- Table S6
- Table S7
- Table S8

- Table S9

- Table S10

5.3 Results

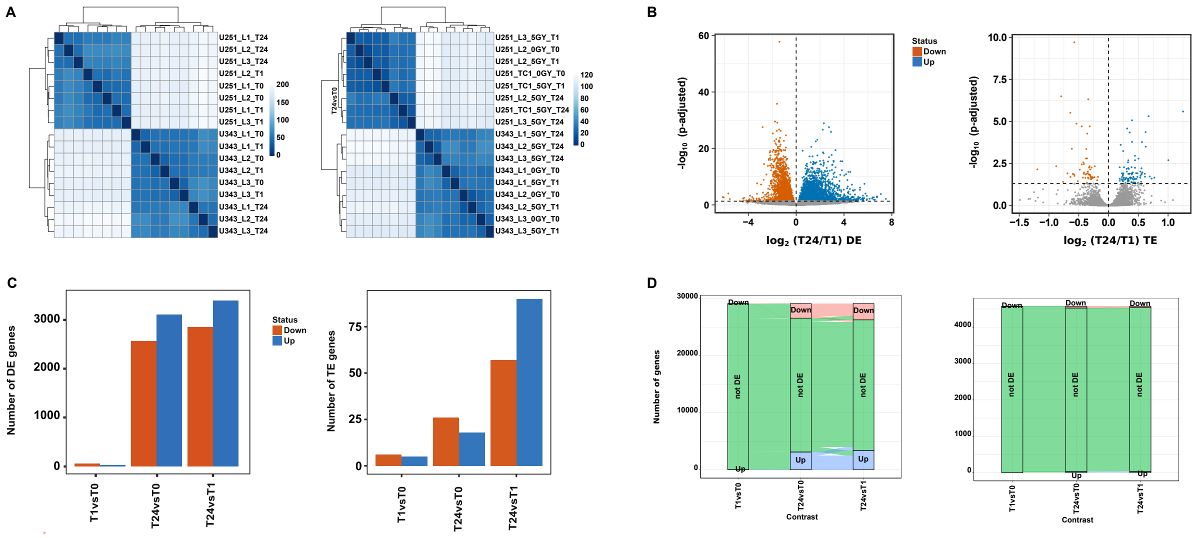


Figure 5.4

Global view of glioblastoma cell lines transcription and translation profiles after radiation. A) Differentially expressed genes (left) and the number of genes whose translation efficiency is differentially regulated (right) after radiation exposure at different time points. B) Volcano plots showing the expression and transition alterations of genes at 24 hr compared to 1 hr after radiation exposure. Blue dots indicate upregulated genes (adjusted p-value < 0.05, $\log_2$ fold change < 0), and orange dots indicate downregulated genes. (adjusted p-value < 0.05, $\log_2$ fold change < 0). C) Sizes of gene modules found in U251 and U343 cell lines. D) Preservation Median Rank and Z_{summary} for all modules. A lower median rank indicates the module is preserved, and the corresponding modules in U251 and U343 cell lines share a high number of genes. A Z_{summary} score of 2-10 indicates weak preservation, while a Z_{summary} >10 indicates high preservation.

5.3.1 Changes in global transcriptome profile in response to radiation

We first conducted an integrated analysis to evaluate the early impact of radiation [1 hour (T1) and 24 hours (T24)] on the expression profile of U251 and U343 GBM lines. A relatively small number of genes displayed altered expression at T1 (Supplementary Figures S4A and S4B). Downregulated genes are mainly involved in transcription regulation and include 18 zinc finger transcription factors displaying high expression correlation in glioblastoma samples from TCGA (Table S1). Upregulated sets contain genes implicated in cell cycle arrest, apoptosis, and stress such as ZFP36, FBXW7, SMAD7, BTG2, and PLK3 (Table S1).

Since many alterations were observed when comparing the T1 vs. T24 time points (Table S1), we opted to focus on genes showing the most marked changes (log_2 fold-change > 1.0 or < -1.0 and adjusted p-value < 0.05) to identify biological processes and pathways most affected at the T24 time point. Top enriched GO terms and pathways among downregulated genes include chromatin remodeling, cell cycle, DNA replication, and repair (Figure 5.5A). Additionally, we identified several GO terms associated with mRNA metabolism, decay, translation, and ncRNA processing, suggesting active participation of RNA-mediated processes in radio-response (Figure 5.5B). Network analysis indicated the set of genes in these categories is highly interconnected (Figure 5.5C and Table S2).

To expand the expression analysis, we employed WGCNA [189] to identify gene modules with significant co-expression variation as an effect of radiation. All identified modules, along with the complete list of genes in each module, are shown in Supplementary Figures S4C and S4D and Table S3. Seven modules were identified ($Z_{\text{summary}} > 5$) as tightly regulated, independent

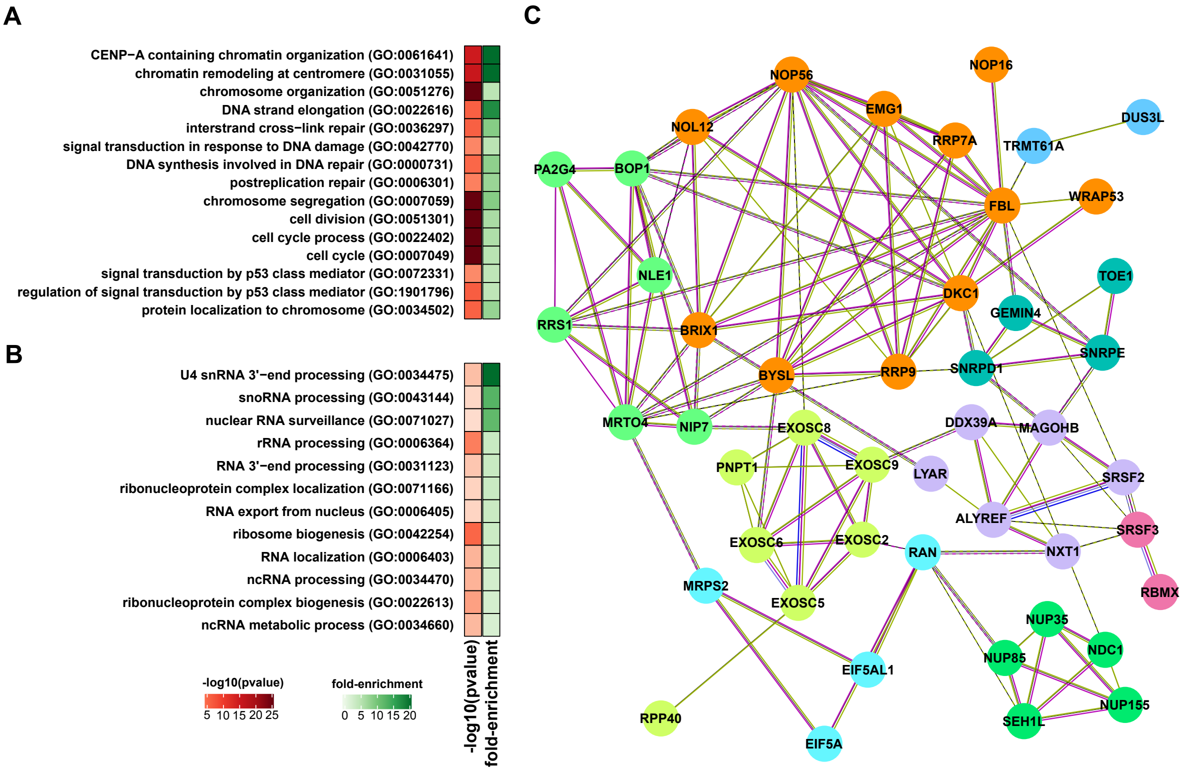


Figure 5.5
Characteristics of downregulated genes at 24 hours (T24) after radiation exposure in glioblastoma cell lines. A) Enriched gene ontology related to cell cycle, DNA replication, and repair among downregulated genes. B) RNA-related Gene Ontology (GO) terms enriched among downregulated genes summarized using REVIGO [2]. C) Protein-protein interaction network, according to STRING [3] showing downregulated genes associated with RNA-related functions. Gene clusters based on the strength of connection and gene function are identified by color. Lines colors indicate the type of association: light green indicates an association based on literature findings; blue indicates gene co-occurrence; magenta indicates experimental evidence.

of the cell line (Figure S4D). Among modules with the highest significant correlation (0.8, p-value$< 1e-7$), module 2 contains genes downregulated in T24, with many involved in cell cycle, metabolism mRNA metabolism, processing, splicing, and transport (Table S3), corroborating results described above.

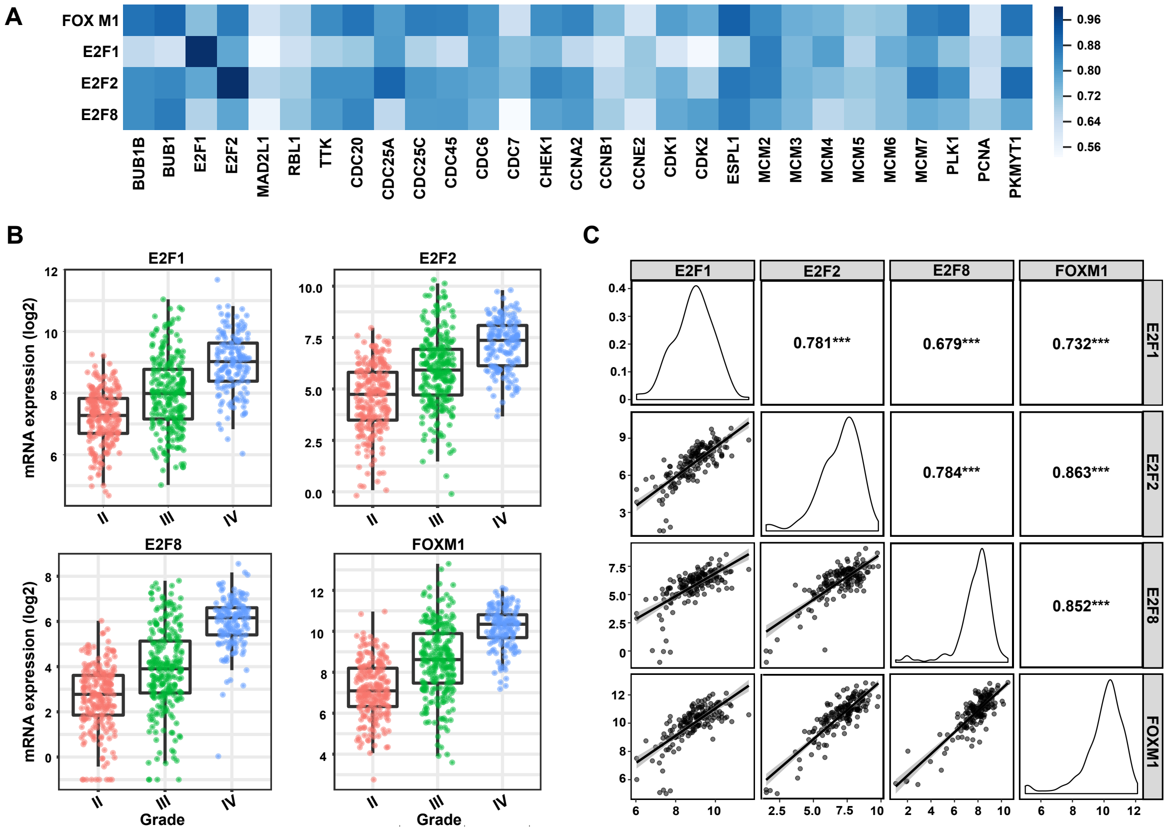


Figure 5.6

E2Fs and FOXM1 in glioblastoma. A) Correlation of E2F1, E2F2, E2F8, and FOXM1 with target genes involved in cell cycle. B) Expression levels of E2F1, E2F2, E2F8, and FOXM1 in gliomas grades II, III, and IV in TCGA samples. C) E2F1, E2F2, E2F8, and FOXM1 expression correlation in glioblastoma (TCGA samples) using Gliovis [4]. *** p-value < 0.0001.

Next, we investigated downregulated genes with the gene set enrichment analysis (GSEA) tool Enrichr [199] and conducted expression correlation analysis with Gliovis [4]. Based on their genomic binding profiles and effect of gene expression, FOXM1 and the E2F family of transcription factors emerged as potential regulators of a large group of cell cycle/DNA replication-related genes in the affected set (Figure 5.6A, Table S4). In agreement, E2F1, E2F2, E2F8, and FOXM1 displayed a significant decrease upon radiation. FOXM1 and E2F factors have been previously implicated in chromatin remodeling, cell cycle regulation, DNA repair, and radio-resistance [200, 201]. All four factors are highly expressed in glioblastoma with respect to low-grade glioma. Importantly, they display high expression correlation with a large set of downregulated genes implicated in cell cycle and DNA replication and among themselves in glioblastoma samples in TCGA (Figure 5.6B-C).

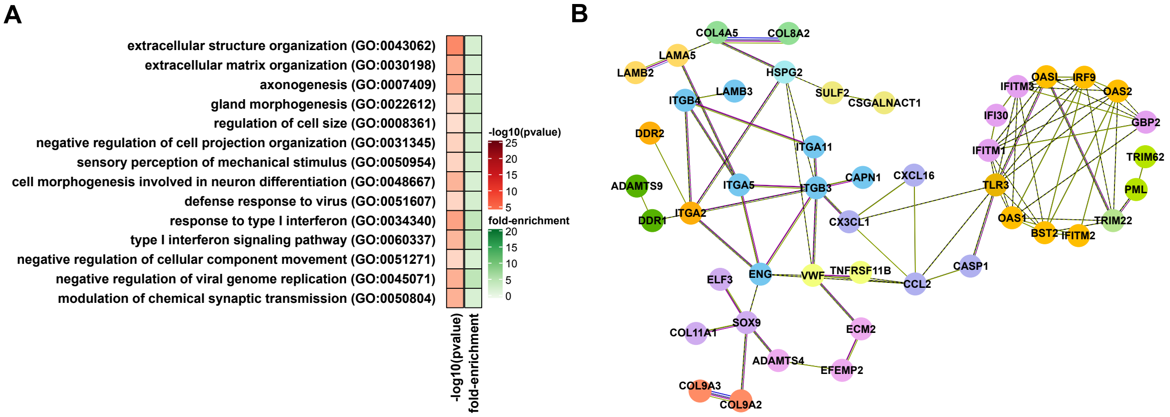


Figure 5.7
Global view of upregulated genes at T24 post-radiation in glioblastoma cells. A) Gene ontology analysis of upregulated genes B) Protein-protein interaction networks according to STRING [3] showing genes associated with extracellular matrix organization and response to interferon. Gene clusters based on the strength of connection and gene function are identified by color. Lines colors indicate type of association: light green, association based on literature findings; blue indicates gene co-occurrence; magenta indicates experimental evidence.

Upregulated genes at T24 are preferentially associated with the extracellular matrix receptor interaction pathway, extracellular matrix organization, axonogenesis, and response to type I interferon (Figure 5.7A, and Table S2). With respect to the extracellular matrix, we observed changes in the expression levels of several collagens (types II, IV, V, and XI), glycoproteins of the laminin family (subunits $\alpha, \beta, \mathrm{and} \gamma$), and also integrins (subunits α, and β) (Figure 5.7B, and Table S1). Collagen type IV is highly expressed in glioblastoma and implicated in tumor progression [202]. In addition, it has been observed that the activation of two integrins, ITGB3 and ITGB5, contributes to radio-resistance [203].

Radiation treatment also induced the expression of genes involved in neuronal differentiation and axonogenesis. Some key genes in these categories include SRC, VEGFA, EPHA4, DLG4, MAPK3, BMP4, and several semaphorins. These genes can have very different effects on glioblastoma development, with some factors activating oncogenic programs and others behaving as tumor suppressors. Similarly, type I interferon's effects on treatment are varied. For instance, interferon inhibited proliferation of glioma stem cells and their sphere-forming capacity and induced STAT3 activation [204]. On the other hand, chronic activation of type I IFN signaling has been linked to adaptive resistance to therapy in many tumor types [205].

Activation of oncogenic signals post-radiation could counteract treatment effects and later contribute to relapse. We searched the set of highly up-regulated genes post-radiation for previously identified radio-resistance genes in glioblastoma, oncogenic factors and genes whose high expression is associated with poor prognosis (Table S5). In Table 5.1, we list these genes according to their molecular function. Since several of these genes have never been characterized in the context of glioblastoma, our results open new opportunities to prevent radio-resistance and increase treatment efficiency. Importantly, there are inhibitors available against several of these proteins (Table S4).

Table 5.1
List of oncogenic factors, genes whose high expression is associated with poor survival and genes previously associated with radio-resistance in GBM that showed increased expression upon radiation. Genes are listed according to molecular function.

Function	Genes
Membrane protein	AQP1, ARHGEF2, BAALC, CSF1R, CSPG4, EPS8L2, ERBB3, FGFR4, FYN, GPM6A, ITGB3, JUP
Protein kinase	ANKK1, CDKN1A, CSF1R, ERBB3, FAM20C, FGFR4, FYN, IKBKE, MERTK, PDGFRB, SRC, TEC
Gene expression regulation	ARID3A, ASAH1, BCL3, BCL6, CBX7, CEBPB, ELF3, FAM20C, FEZF1, HOXA1, HOXB9, JUP, KDM5B, LMO1, LMO2, MACC1, MAF, MSI1, MUC1, NKX2-1, PML, PRDM6, RORC, SATB1, SREBF1, TP53BP1, ZMYM2
Enzymatic activity	ACSS2, AGAP2, APIP, ARHGEF2, C1R, CARD16, CD24, CDKN1A, CEBPB, CSF1R, CSPG4, CTSZ, CUL7, CYTH4, EPS8L2, ERBB3, FAM20C, FGFR4, FTH1, FUCA1, FYN, GHDC, IDO1, IKBKE, ITGB3, JUP, KDM5B, MCF2, MERTK, MFNG, MRAS, NKX2-1,PDE6G,PDGFRB, PML, QPRT, RRM2B, SERPINA5, SFN, SGSH, SRC, SREBF1, TEC, TGFB1, ZMYM2
Phosphotransferase	CSF1R, ERBB3, FAM20C, FYN, IKBKE, MERTK, PDGFRB, SREBF1, TEC
Cell surface receptor	BMP7, CSF1R, ERBB3, FGFR4, FYN, ITGB3, ITGB5, LRIG2, MCF2, MERTK, MFNG, PDGFRB, PRDM6, SRC, TEC, TRPM8
Metabolism regulation	ACSS2, APIP, BCL3, BCL6, BTG2, CEBPB,CRTC1, CSPG4, CTSZ, ELF3, FUCA1, IDO1, ITGB3, JUP, MAF, MFNG, NKX2-1, PARP3, PRDM6, PTGES, QPRT, RRM2B, SGSH, TGFB1, TP53BP1, TRPM8, USP9X, VEGFA, ZMYM2

5.3.2 Changes in lncRNA profile in response to radiation

lncRNAs have been implicated in the progression of glioblastoma [206], but their role in response to ionizing radiation is still poorly understood. We identified 161 lncRNAs with expression alterations in T1 vs. T24 comparisons. Analysis of this set with LnC2Cancer [197] identifieddentified several lncRNAs aberrantly expressed in cancer and with relevance to prognosis (Table S1). We also detected significant downregulation of MIR155HG, whose high expression is associated with glioma progression and poor survival [207]. Another downregulated lncRNA with relevance to prognosis is linc000152, whose increased expression has been observed in multiple tumor types [208, 208]. On the other hand, we observed a significant upregulation of two "oncogenic" lncRNAs, NEAT1 and FTX. NEAT1 is associated with tumor growth, grade, and recurrence rate in gliomas [209], while FTX promotes cell proliferation and invasion through negatively regulating miR-342-3p [210]. Thus, if further studies corroborate NEAT1 and FTX as players in radio-resistance, targeting these lncRNAs should be considered to improve treatment response.

5.3.3 Effect of radiation on splicing

Alternative splicing impacts genes implicated in all hallmarks of cancer [211] and is an important component of changes in expression triggered by ionizing radiation [212]. All types of splicing events (exon skipping, alternative donor, and acceptor splice sites, multiple exclusive exons, and intron retention) were affected similarly upon exposure to radiation (Table S7). At T24, we observed that transcripts associated with RNA-related functions (especially translation), showed the most splicing alterations. Affected transcripts encode ribosomal proteins, translation initiation factors, regulators of translation, and genes involved in tRNA processing and endoplasmic

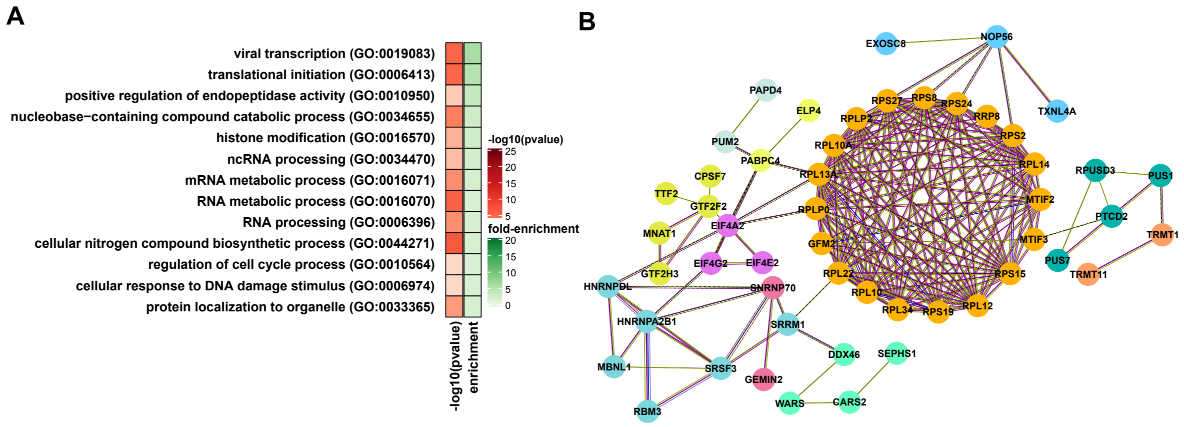


Figure 5.8

Impact of radiation on the splicing profile of glioblastoma cells. A) GO-enriched terms among genes showing changes in splicing profiles at T24. GO-enriched terms are summarized using REVIGO [2]. B) Protein-protein interaction networks according to STRING [3] showing genes associated with RNA-related functions whose splicing profiles displayed alterations at T24. Gene clusters based on the strength of connection and gene function are identified by color. Lines color indicate type of association: light green, an association based on literature findings; blue indicates gene co-occurrence; magenta indicates experimental evidence.

reticulum. Other enriched GO terms include mRNA and ncRNA processing, mRNA degradation, and modification. Catabolism is another process associated with several enriched terms, suggesting that splicing alterations in genes involved in catabolic routes could ultimately contribute to apoptosis (Figure 5.8A-B, and Table S7). Changes in the splicing profile are likely driven by an alteration in the expression of splicing regulators. In Table 5.2, we show a list of splicing factors displaying strong expression alterations. Among those previously connected to glioblastoma development upon radiation, LGALS3 is the most extensively characterized. LGALS3 is a galactosidase-binding lectin and non-classic RNA binding protein implicated in pre-mRNA splicing and regulation of proliferation, adhesion, and apoptosis; LGALS3 also is a marker of the early stage of glioma [213].

Table 5.2
Splicing regulators showing changes in expression 24 hours post-radiation. Factors showing an increase in the expression are shown in red, while factors showing a decrease in the expression are represented in blue.

Gene ID	Gene name	Function
AHNAK2	Protein AHNAK2	splicing regulation
ESRP1	Epithelial splicing regulatory protein 1	regulation of mRNA splicing
LGALS3	Galectin-3	signaling receptor binding
NOVA2	RNA-binding protein Nova-2	alternative splicing regulation
SNRPN	Small nuclear ribonucleoprotein N	spliceosomal snRNP assembly
ALYREF	THO complex subunit 4	RNA binding
DDX39A	ATP-dependent RNA helicase DDX39A	RNA helicase
GEMIN4	Gem-associated protein 4	rRNA processing
HNRNPL	Heterogeneous nuclear ribonucleoprotein L	alternative splicing regulation
LSM2	U6 snRNA-associated Sm-like protein LSm2	U6 snRNA-associated Sm-like protein
MAGOHB	Mago nashi homolog 2	exon-exon junction complex
PPIH	Peptidyl-prolyl cis-trans isomerase H	ribonucleoprotein complex binding
RBMX	RNA-binding motif protein, X chromosome	regulation of mRNA splicing
SNRPD1	Small nuclear ribonucleoprotein Sm D1	spliceosomal snRNP assembly
SNRPE	Small nuclear ribonucleoprotein E	spliceosomal snRNP assembly
SRSF2	Serine/arginine-rich splicing factor 2	regulation of mRNA splicing
SRSF3	Serine/arginine-rich splicing factor 3	regulation of mRNA splicing
TTF2	Transcription termination factor 2	transcription regulation

5.3.4 Differential translational efficiency

We used Ribo-seq [31] to identify changes in translation efficiency triggered by radiation. Translation, protein localization, and metabolism appear as top enriched terms among downregulated genes in T1 vs. T24 comparisons (Tables S8-S9). In particular, several ribosomal proteins, along with translation initiation factors and mTOR, showed a significant decrease in translation efficiency (Figure 5.9A-B). Overall, these results indicate repression of the translation machinery post-radiation exposure and its strong auto-regulation. Since changes in components of the translation machinery are occurring at all levels (transcription, splicing, and translation) at T24, we expect that major translational alterations take place in later stages of post-radiation.

In the upregulated set, we highlight three genes FTH1, APIP, and LRIG2 that could potentially counteract the impact of radiation (Table S10). FTH1 encodes the heavy subunit of ferritin, an essential component of iron homeostasis [214]. Pang *et al.*, 2016 [215] reported that H-ferritin plays an important role in radio-resistance in glioblastoma by reducing oxidative stress and activating DNA repair mechanisms. The depletion of ferritin causes down-regulation of ATM, leading to increased DNA sensitivity towards radiation. APIP is involved in the methionine salvage pathway and has a key role in various cell death processes. It can inhibit mitochondria-mediated apoptosis by directly binding to APAF-1 [216]. LRIG2 is a member of the leucine-rich and immunoglobulin-like domain family [217], and its expression levels are positively correlated with the glioma grade and poor survival. LRIG2 promotes proliferation and inhibits apoptosis of glioblastoma cells through activation of EGFR and PI3K/Akt pathway [218].

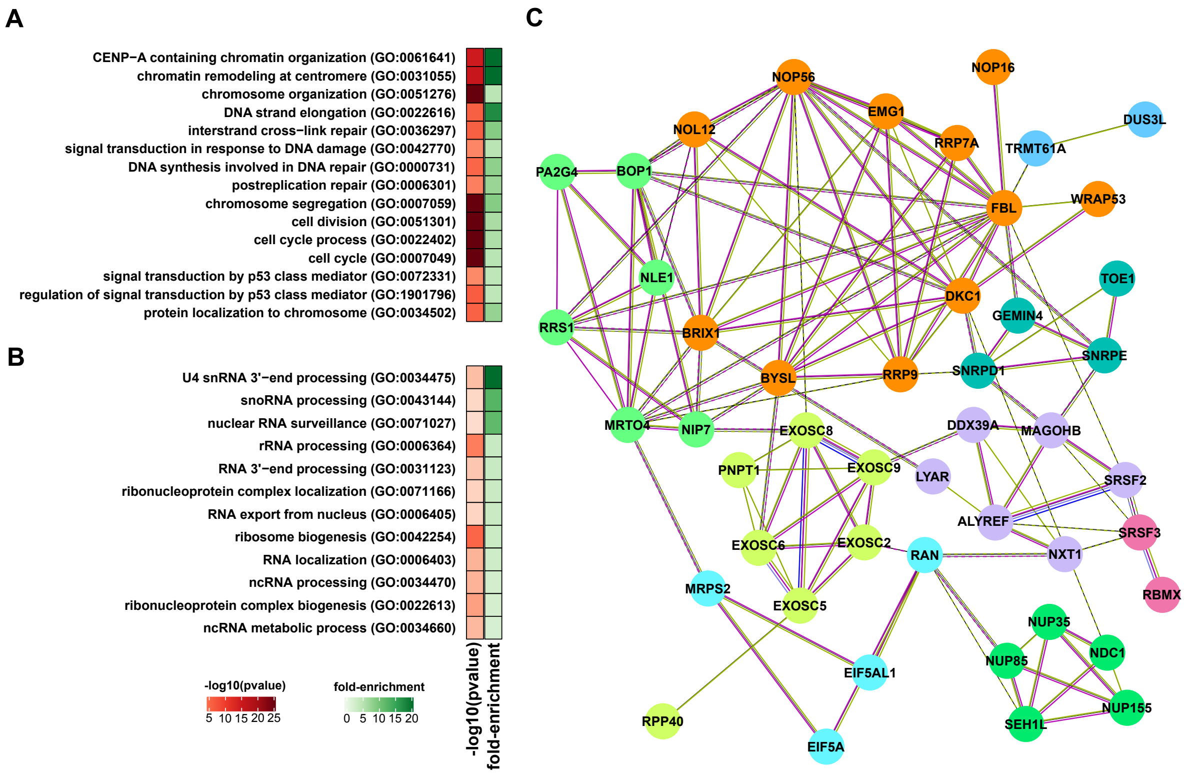


Figure 5.9

Impact of radiation on the translation profile of glioblastoma cells. A) GO-enriched terms among genes showing changes in translation efficiency at T24. GO-enriched terms are summarized using REVIGO [2]. B) Protein-protein interaction network, according to STRING [3] showing genes whose translation efficiency decreased at T24. Gene clusters based on the strength of connection and gene function are identified by color. Line colors indicate the type of association: light green, an association based on literature findings; blue indicates gene co-occurrence; magenta indicates experimental evidence.

5.3.5 Crosstalk between regulatory processes

Parallel analyses of transcription, splicing, and translation alterations in the early response to radiation provided an opportunity to identify crosstalk between different regulatory processes. The datasets showed little overlap, with just a few genes showing alterations in two different regulatory processes. However, we identified several shared GO terms when comparing the results of alternative splicing, mRNA levels, and translation efficiency (Table S10). These terms show two main groups of biological processes. The first group indicates that the expression of genes involved in DNA and RNA synthesis and metabolism is particularly compromised. The second group is related to translation initiation. Ribosomal proteins were particularly affected (Figures 5.8 and 5.9). There is growing support for the concept of specialized ribosomes. According to this model, variations in the composition of the ribosome due to the presence or absence of certain ribosomal proteins or alternative isoforms could ultimately dictate which mRNAs get preferentially translated [219]. Therefore, these alterations could later lead to translation changes of a specific set of genes.

5.4 Discussion

We performed the first integrated analysis to define global changes associated with the early response to radiation in glioblastoma. Our approach allowed the identification of “conserved” alterations at the transcription, splicing and translation levels and defined possible crosstalk between different regulatory processes. Alterations at the level of transcription were dominant, but changes affecting genes implicated in RNA mediated regulation were ubiquitous; they indicate that these processes are important components in radio-response and suggest that more robust changes in splicing and translation might take place later.

5.4.1 E2F1, E2F2, E2F8, and FOXM1 as major drivers

We observed marked changes in the mRNA levels of genes implicated in cell cycle, DNA replication, and repair 24 hours (T24) after radiation. Downregulation of several transcription factors, most of them members of the zinc finger family, was observed at one hour post-radiation. This group displays high correlations in expression within glioblastoma samples from TCGA, suggesting that they might work together to regulate gene expression. Unfortunately, most are poorly characterized, and the lack of information has prevented establishing further connections to changes in the cell cycle and DNA replication that we observed at T24.

GSEA and expression correlation analysis suggested that the downregulation of members of the E2F family is likely responsible for several of the expression changes we observed at T24. E2Fs have been defined as major transcriptional regulators of the cell cycle. The family has eight members that could act as activators or repressors depending on the context, and are known to regulate one another. They are upregulated in many tumors due to overexpression of cyclin-dependent kinases (CDKs), inactivation of CDK inhibitors, or RB Transcriptional Corepressor 1 (RB1) and are linked to poor prognosis. Alterations in E2F genes can induce cancer in mice [220, 221]. Specifically, we found that three E2F members showed decreased expression upon radiation: E2F1, E2F2, and E2F8, all of which have been previously implicated in glioblastoma development.

E2F1 is probably the best-characterized member of the E2F family. Besides its known effect on cell cycle regulation and DNA replication, it is also a positive regulator of telomerase activity, binding the TERT promoter [222]. Recent studies show that lncRNAs and miRNAs function in an antagonistic fashion to regulate E2F1 expression, ultimately affecting cell proliferation, glioblastoma growth, and response to therapy [223, 224, 225]. E2F2 has been linked to the maintenance of glioma stem cell phenotypes and cell transformation [226, 227]. Several tumor suppressor miRNAs (let7b, miR-125b, miR-218, and miR-138) decrease the proliferation and growth of glioblastoma cells by targeting E2F2 [226, 228, 229, 230]. Although still poorly characterized in the context of glioblastoma, E2F8 drives an oncogenic phenotype in glioblastoma. Its expression is modulated by HOXD-AS1, which serves as a sponge and prevents the binding of miR-130a to E2F8 transcripts [231]. FOXM1 is another potential regulator of the group of cell cycle and DNA replication genes affected by radiation. FOXM1 is established as an important player in chemo- and radio-resistance and a contributor to glioma stem cell phenotypes [170, 171, 232, 233, 234, 235, 236, 237]. FOXM1 and E2F protein have a close relationship and share target genes [238]. Additionally, FOXM1- and E2F2-mediated cell cycle transitions are implicated in the malignant progression of IDH1 mutant glioma [239].

E2F and FOXM1 targeting could be considered as an option to increase radio-sensitivity. Since the development of transcription factor inhibitors is very challenging, an alternative to be considered is the use of BET (bromodomain and external) inhibitors. BET is a family of proteins that function as readers for histone acetylation and modulates the transcription of oncogenic programs [240]. Recent studies in glioblastoma with a new BET inhibitor, dBET6, showed promising results and established that its effect on cancer phenotypes comes via disruption of the transcriptional program regulated by E2F1 [241].

5.4.2 RNA processing and regulation as novel categories in radio-response

Besides the expected changes in expression of cell cycle, DNA replication and repair genes, radiation affected preferentially the expression of genes implicated in RNA processing and regulation. Additionally, we identified a co-expression module containing multiple genes associated with translation initiation, rRNA and snoRNA processing, RNA localization, and ribonucleoprotein complex biogenesis.

Many regulators of RNA processing are implicated in glioblastoma development, and splicing alterations affect all hallmarks of cancer [242]. Radiation-induced changes in the splicing patterns of oncogenic factors and tumor suppressors such as CDH11, CHN1, CIC, EIF4A2, FGFR1, HNRNPA2B1, MDM2, NCOA1, NUMA1, RPL22, SRSF3, TPM3, APC, CBLB, FAS, PTCH1, and SETD2. We also observed changes in expression of four RNA processing regulators previously identified in genomic/functional screening for RNA binding proteins contributing to glioblastoma phenotypes: MAGOH, PPIH, ALYREF, and SNRPE [243].

5.4.3 Potential new targets to increase radio-sensitivity and prevent relapse

Activation of oncogenic signals is an undesirable effect of radiation that could influence treatment response and contribute to relapse. We observed increased expression or translation and splicing alterations of a number of pro-oncogenic factors, genes whose high expression is associated with poor survival and genes previously implicated in radio-resistance.

Among genes with the most marked increase in expression upon radiation, we identified members of the Notch pathway (HES2, NOTCH3, MFNG, and JAG2). Notch activation has been linked to radio-resistance in glioblastoma, and Notch targeting improves the results of radiation treatment [244, 245]. We also identified several genes associated with the PI3K-Akt, Ras, and Rap1 signaling pathways that increased expression levels upon radiation exposure. Targeting these pathways has been explored as a therapeutic option in glioblastoma [246, 244]. Other oncogenic factors relevant to glioblastoma that had increased expression after radiation exposure include SRC, MUC1, LMO2, PML, PDGFRβ, BCL3, and BCL6.

Anti-apoptotic genes (BCL6, RRM2B, and IDO1) also showed increased expression upon radiation. BCL6 is a member of the ZBTB family of transcription factors, which functions as a p53 pathway repressor. The blockage of the interaction between BCL6 and its cofactors has been established as a novel therapeutic route to treat glioblastoma [247]. RRM2B is an enzyme essential for DNA synthesis and participates in DNA repair, cell cycle arrest, and mitochondrial homeostasis. The depletion of RRM2B resulted in ADR-induced apoptosis, growth inhibition, and enhanced sensitivity to chemo- and radiotherapy [248]. IDO1 is a rate-limiting metabolic enzyme involved in tryptophan metabolism that is highly expressed in numerous tumor types [249]. The combination of radiation therapy and IDO1 inhibition enhanced therapeutic response [250].

Among genes whose high expression correlates with decreased survival in glioblastoma, we identified several components of the "matrisome" and associated factors (FAM20C, SEMA3F, ADAMTSL4, ADAMTS14, SERPINA5, and CRELD1). The core of the "matrisome" contains ECM proteins, while associated proteins include ECM-modifying enzymes and ECM-binding growth factors. This complex of proteins assembles and modifies extracellular matrices, contributing to cell survival, proliferation, differentiation, morphology, and migration [251]. In addition, several genes of the proteinase inhibitor SERPIN family (SERPINA3, SERPINA12, SERPINA5, and SERPINI1) implicated in ECM regulation [252] were among those with high levels of expression upon radiation.

In conclusion, our results generated a list of candidates for combination therapy. Contracting the effect of oncogenic factors and genes linked to poor survival could increase radio-sensitivity and treatment efficiency. Importantly, there are known inhibitors against several of these proteins (Table S5). Moreover, RNA processing and translation were determined to be important components of radio-response. These additional vulnerable points could be explored in therapy, as many inhibitors against components of the RNA processing and translation machinery have been identified [253, 254].

Chapter 6

Ribopod: A database of uniformly processed Ribo-seq datasets

6.1 Introduction

Even after 20 years of the Human Genome Project, there appears to be a no consensus on the number of protein coding genes [255]. The ENSEMBL [256] gene prediction process is based on alignments of protein and cDNA sequences, which is claimed to produce a low rate of false positives. The consensus coding sequence project (CCDS) [101] represents a consensus model which is defined as protein-coding regions that agree at the start codon, stop codon and splice junctions such that the prediction meets certain quality assurance benchmarks. These benchmarks involve finding consensus between CCDS regions and SWISS-PROT [257] proteins and ensuring CCDS regions satisfy genomic conservation criterion. Ribo-seq data provided exact information on regions that are being actively translated. The availability of public Ribo-seq datasets across different physiological and pathological contexts across different species has made it possible to learn new biology, particularly of mechanisms that are shared across the conditions. This has motivated the development of Ribo-seq databases which provide access to processed Ribo-seq data.

There have been multiple attempts at creating a database for Ribo-seq studies including GWIPS-Viz [258], RPFdb [259], and HRPDViewer [260]. GWIPS-Viz [258] provides a database of Ribo-seq coverages across multiple species. It provides global aggregates of analyzed public datasets as UCSC genome browser tracks. HRPDViewer [260] allows visualization of transcript level profiles for different datasets. RPF-db [259] provides a visualization interface for read count summaries across multiple r egions of the transcriptome as a summary statistic for quality.

While largely useful, these databases do not systematically handle the intricacies of Ribo-seq data. Not all the fragments in Ribo-seq data represent active translation [85, 68] and hence any data downloaded from these datasets needs to be pre-processed to obtain actively translating fragments before performing any downstream analysis. We aim to bridge this gap by developing a new database, ribopod, that provides a visualization interface for assessing the quality scores of a public Ribo-seq dataset besides providing access to *de-noised* Ribo-seq profiles. Ribopod aims bridge the gap in leveraging the existing public datasets to understand translational regulation.

6.2 Ribopod: a database of de-noised Ribo-seq datasets

We developed a database of uniformly processed public Ribo-seq data while paying attention to the intricacies of Ribo-seq data. We use our method ribotricer to filter out non-active fragments and pay attention to the library protocols, track and store cell/tissue type, treatment and, other related metadata as obtained from NCBI's Sequence Reach Archive (SRA) or Gene Expression Omnibus (GEO).

6.2.1 Obtaining metadata

The NCBI Sequence Read Archive (SRA) is the primary archive of next-generation sequencing datasets including Ribo-seq datasets. However, methods to programmatically access this data are limited. We needed a way to automatically extract all metadata of each Ribo-seq project to minimize manual curation. We developed `pysradb` package [93] that provided a simple and user-friendly command-line interface for querying metadata and downloading datasets from. Metadata for all SRA projects in ribopod is obtained using pysradb. Retrieving metadata for any SRA accession is done using:

```
$ pysradb metadata SRP010679 --desc --expand
```

... [truncated]

run_accession	cell_line	sample_type	source_name	treatment
SRR403882	pc3	polya rna	pc3 human prostate cancer cells	vehicle
SRR403883	pc3	ribosome protected rna	pc3 human prostate cancer cells	vehicle
SRR403884	pc3	polya rna	pc3 human prostate cancer cells	rapamycin
SRR403885	pc3	ribosome protected rna	pc3 human prostate cancer cells	rapamycin
SRR403886	pc3	polya rna	pc3 human prostate cancer cells	pp242
SRR403887	pc3	ribosome protected rna	pc3 human prostate cancer cells	pp242
SRR403888	pc3	polya rna	pc3 human prostate cancer cells	vehicle
SRR403889	pc3	ribosome protected rna	pc3 human prostate cancer cells	vehicle
SRR403890	pc3	polya rna	pc3 human prostate cancer cells	rapamycin
SRR403891	pc3	ribosome protected rna	pc3 human prostate cancer cells	rapamycin
SRR403892	pc3	polya rna	pc3 human prostate cancer cells	pp242
SRR403893	pc3	ribosome protected rna	pc3 human prostate cancer cells	pp242

Ribopod' provides a visualization interface for visualizing the metagene plots, fragment length distribution and the phase scores generated by ribotricer. Each sample in an SRA accession is described separately on all the three metrics (Figure 6.1).

6.2.2 Workflow

After obtaining metadata and raw data from SRA, the raw sequences are mapped to the transcriptome using STAR [95] using a transcriptome annotation from ENSEMBL [256]. We

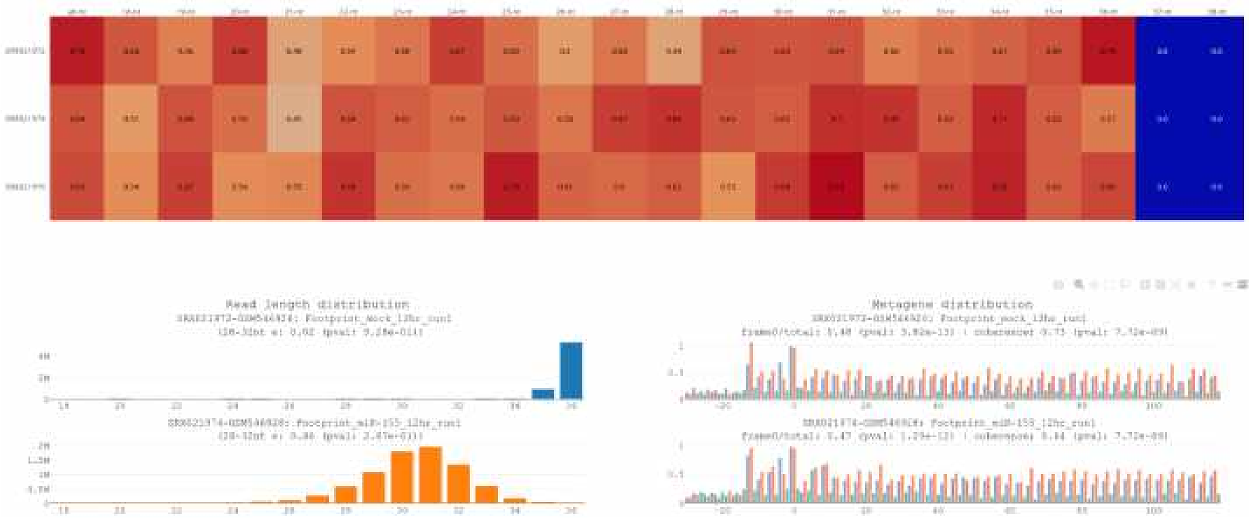


Figure 6.1
Ribopod's visualization interface. The top panel shows a heatmap of ribotricer phase scores for each fragment length. The bottom left shows the fragment length distribution plots for each sample with the corresponding metagene plot on the right. See http://ribopod.usc.edu.

allow for a maximum of two bases mismatches while mapping and only uniquely mapping reads are retained. Next, the uniquely mapped reads are processed through ribotricer to assess them for periodicity. Ribotricer outputs de-noised profiles retaining only the actively translating Ribo-seq reads (Figure 6.2).

6.3 Results

Ribotricer's phase score [149] can be used as a quality control metric to asses the quality of Ribo-sseq data. We processed multiple Ribo-seq projects from different species including human, mouse, rat, zebrafish, fruitfly, *C. elegans*, *S. cerevisiae*, and *S. pombe*. A detailed list of datasets and the associated data is available on this link: `https://bit.ly/ribopod-datasets`. The code used to analyze the datasets is available at `https://github.com/saketkc/re-ribo-smk` and the code for the database is available at `https://github.com/saketkc/ribopod`.

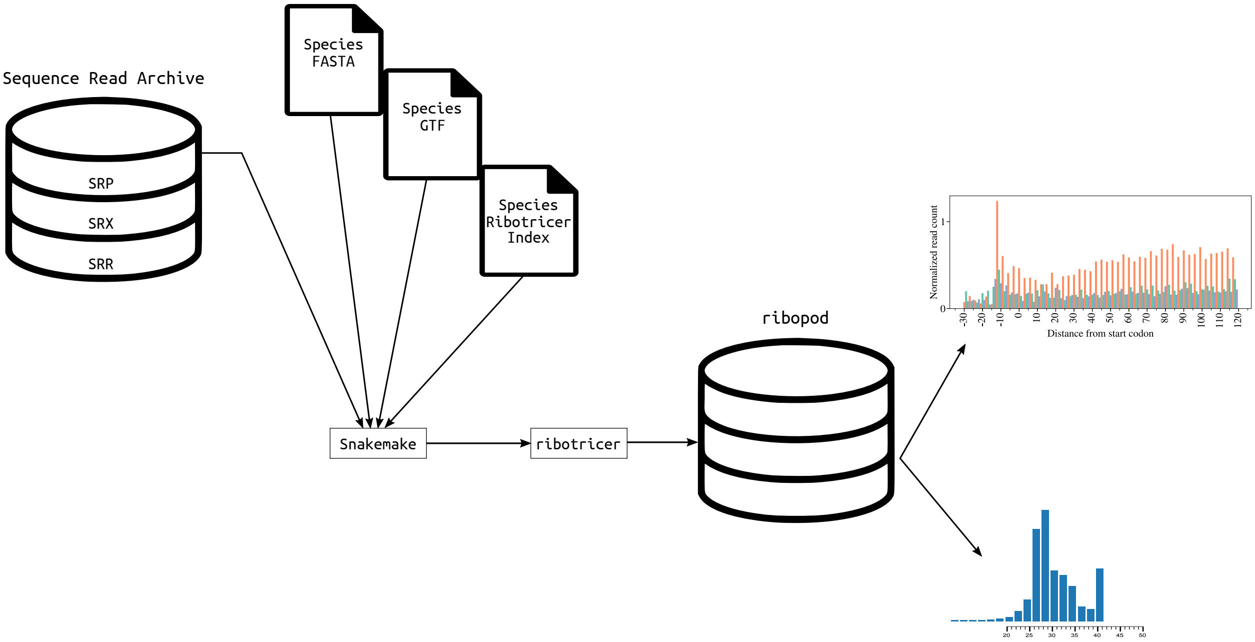


Figure 6.2
Workflow of ribopod

We analyzed hundreds of libraries from different species and assessed their phase score as calculated using ribotricer. As seen in Figure 6.3, the phase score of Ribo-seq datasets are highly variable. They also appear to be dependent on the species. The species-specific effect have two likely sources or origin: 1) biological difference 2) technological difference. However, it is difficulty to quantify their relative contribution. One striking observation of this effect is in *Drosophila* datasets in which all the Ribo-seq libraries seem to have a low phase-score overall. This effect is not necessarily related to a biological difference arising from *Drosophila* samples, since the *C. elegans* and Zebrafish, the two nearby species seem to have higher phase scores.

6.4 Conclusion

Ribopod provides a readily accessible database of processed Ribo-seq datasets that have been processed to filter only actively-translating fragments. Since each dataset is accompanied

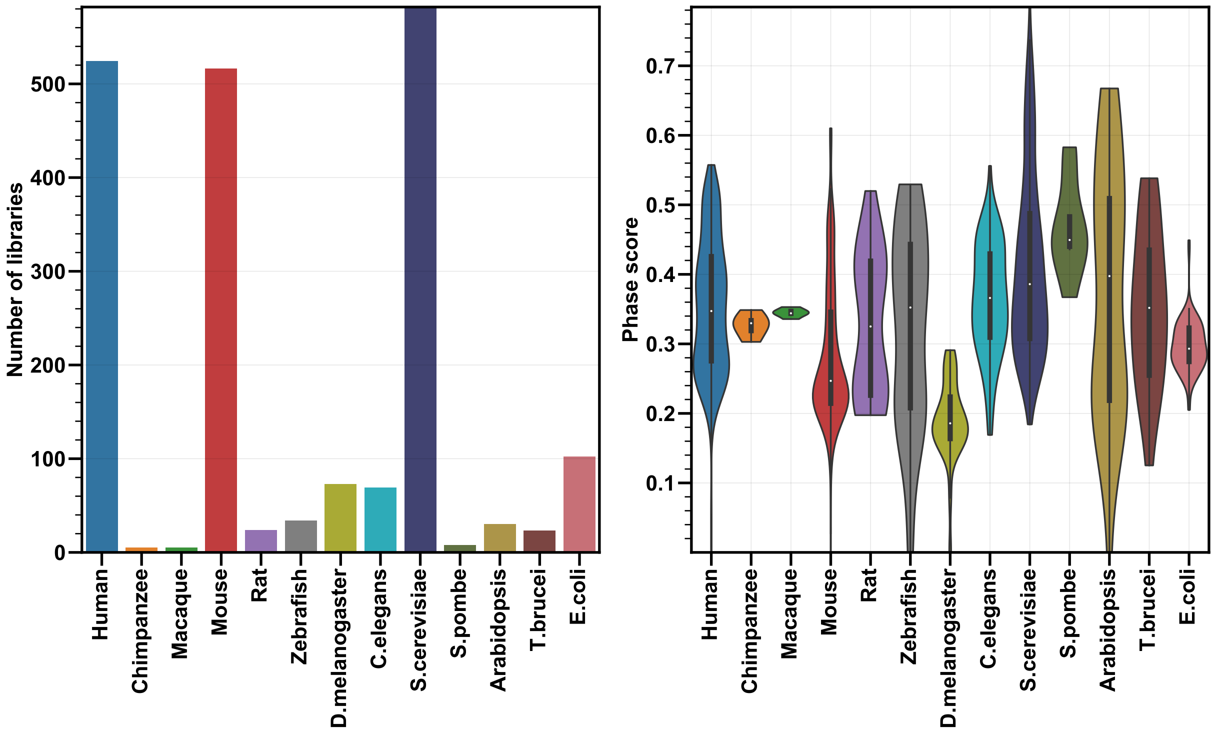


Figure 6.3
Phase score distribution for datasets present in ribopod. Phase scores were calculated over annotated protein-coding regions only.

with an associated phase score, ribopod can be used to perform novel analysis on high quality datasets. The availability of de-noised Ribo-seq profiles across both protein-coding and upstream open reading frames (uORFs) can serve as the key resource for discovering the role of uORFs in different physiological and pathological contexts.

Chapter 7

Conservation of uORF mediated regulation across species

7.1 Introduction

The term expression in "gene expression", a colloquial term in biology, refers to the process by which the gene synthesizes functional products, most often in the form of proteins. Gene expression is modulated by both transcriptional and post-transcriptional regulation. In all organisms, the synthesis of protein requires two key steps: transcription and translation. The expression itself is regulated at multiple steps spanning transcription, post-transcription, translation and post-translation [19, 20]. Upstream open reading frames (uORFs) are a major regulatory element located in the 5' leader sequences that play a role in regulating translation. uORFs are defined by a start codon and an in-frame stop codon and are located upstream of the main coding sequence in the 5' untranslated region (5' UTR) [261, 262, 27]. uORFs have been suggested to conform to Kozak sequence rules [57]. Kozak sequence rules define a 'favorable' context required for initiating translation. However, this hypothesis in the context of uORFs has also been contested

[263]. uORFs have been hypothesized to play a role of down-regulating the expression of corresponding CDS [264, 265, 266]. Around 50% transcripts in human, mouse, and Zebrafish have been hypothesized to have potential uORFs [267, 261, 262, 29].

In eukaryotes, the process of translation may itself regulate gene expression independent of the encoded peptide [268]. The uORFs are unlikely to code for functional peptide products because their peptide sequences are not conserved, even if the presence in 5' region itself might be conserved [265]. But they play dual roles of i) allowing downstream re-initiation and ii) capturing some fraction of the pre-initiation complex thus reducing the translation of protein-coding region. A well studied example of the first role is demonstrated by the first uORF in activating transcription factor 4 (ATF4) transcript that is constitutively translated, and then ribosomes re-initiate at either the second uORF or the CDS [79, 269]. Stress induces phosphorylation of the α subunit of eukaryotic translation initiation factor 2 (eIF2). The phosphorylated eIF2 inhibits the action of eIF2B which in turn attenuates the formation of eIF2-GTP-tRNA$_{\mathrm{i}}^{\mathrm{met}}$ ternary complex. This prevents recycling of eIF2 between different runs of the protein synthesis cycle and overall global translation initiation thus leading to an overall decrease in protein levels [270]. Paradoxically, in ATF4, the stress-induced phosphorylation of eIF2 increases the translation which in turn activates the integrated stress response (a pro-survival pathway). ATF4 has two conserved uORFs, uORF1 and uORF2. In unstressed condition, lower levels of eIF2α phosphorylation favor the ribosomes to re-initiate translation at uORF2. When phoshorylated, the formation of eIF2-GTP-tRNA$_{\mathrm{i}}^{\mathrm{met}}$ is sequestered which reduces the probability that the ribosomes terminating at uORF1 will be able to acquire enough concentration of this complex to re-initiate translation at uORF2. However, they they may acquire them in the process of scanning the downstream AUG corresponding to the CDS, thus up-regulating the translation. For the second example, where uORFs are inhibitory to the main CDS. uORF translation can thus affect the expression of downstream CDS in a global [266] or a local [271] setting.

uORFs have been known to play a role in myriad of human diseases [272, 273, 274, 79, 275]. In particular, loss of function mutations in uORF of kinases is known to be associated with malignancies [276], uORFs serve as a source of tumor antigen [277] and in neurodegenerative diseases [278]. Preotemics based analyses have also linked uORFs to lower protein levels [279]. While much is known about the regulatory role of uORFs in the context of individual genes [273, 274], a global analysis on their translation status across different physiological and pathological contexts is currently missing. We utilize public Ribo-seq studies across eight species including human, mouse, rat, zebrafish, fruitfly, *C. elegans*, *S. cerevisiae*, and *S. pombe* across different physiological and pathological contexts to discover uORFs that are always under active translation and analyze their sequence context and its conservation across the species.

7.2 Results

In order to decipher the regulatory role of uORFs and to asses if it is conserved across species we analyzed multiple public Ribo-seq datasets as made available through ribopod (Chapter 6). Our results are organized in two parts. In the first part we establish the regulatory activity of uORFs across three species (human, chimp, and macaque) using matched Ribo-seq and RNA-seq data. In the second part we analyze datasets from eight species including four vertebrates: human, mouse, rat, and zebrafish and four invertebrates: fruitfly, *C. elegans*, baker's yeast, and *S. pombe*. to establish the conservation of uORFs across species. The datasets spanning these species have been performed in a variety of physiological and pathological contexts. The total

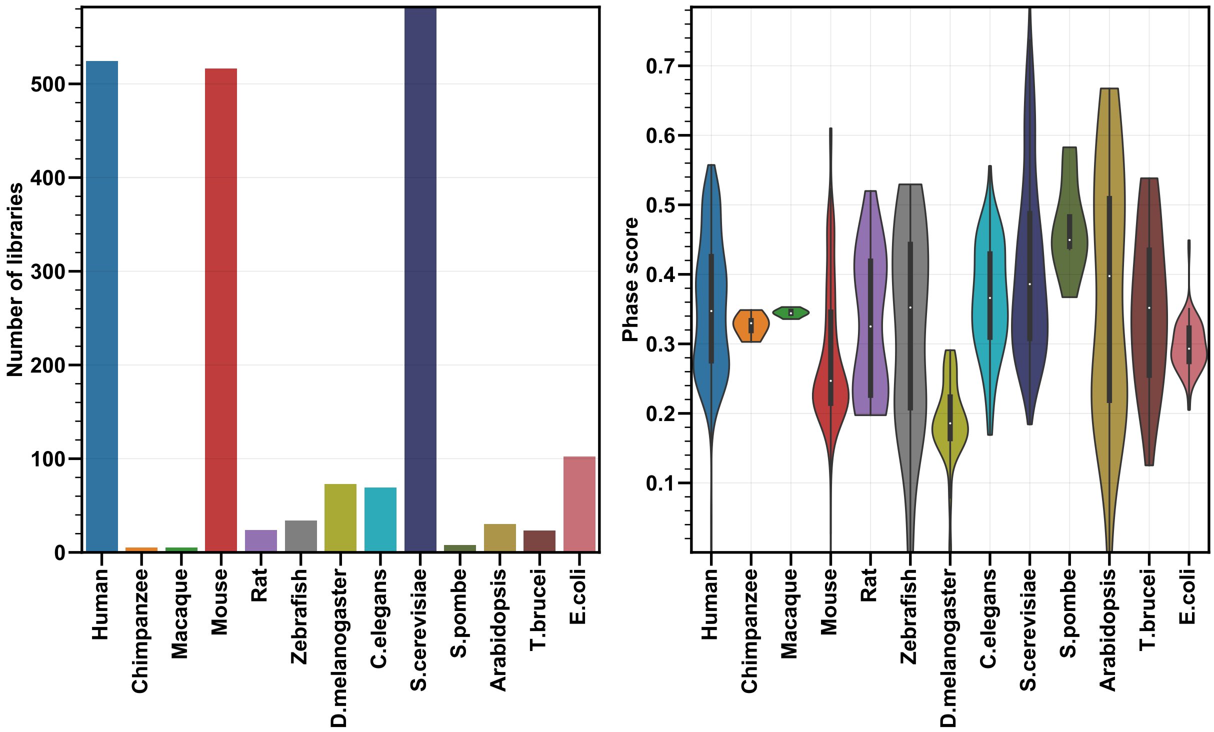


Figure 7.1
Distribution of phase scores and the number of libraries

number of libraries and their corresponding ribotricer generated phase scores [149] are summarised in Figure 7.1.

7.2.1 uORFs have a repressive effect on translational efficiency

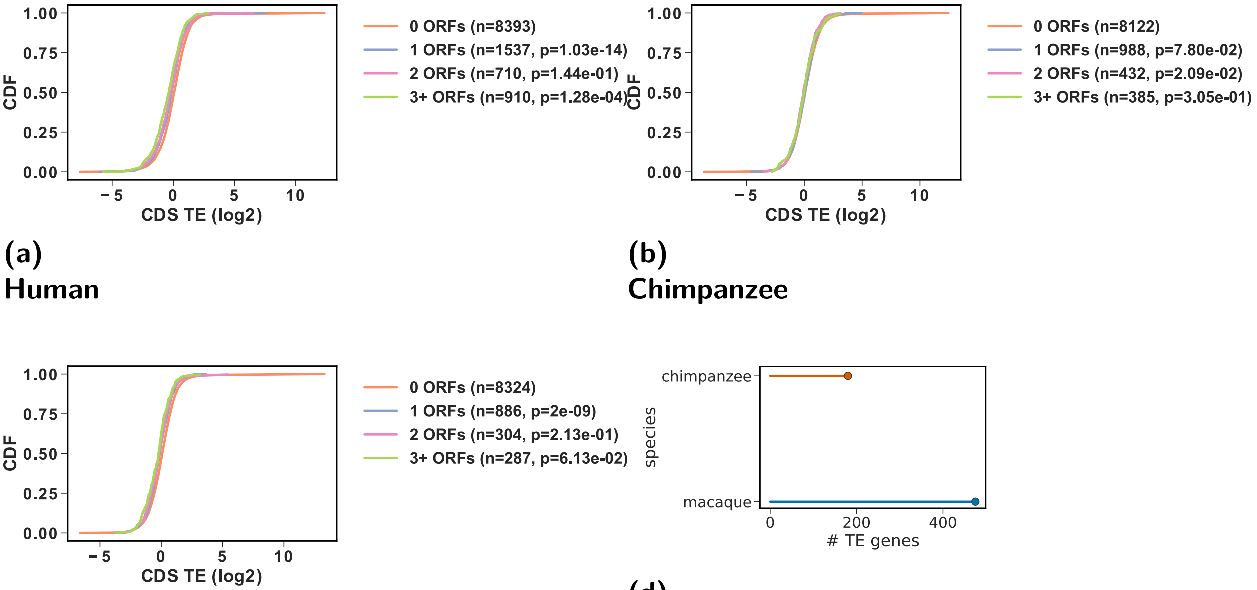
To first characterize the regulatory role of uORFs, we analyzed Ribo-seq data available from Wang *et al.* [280] (SRA accession SRP062129). The dataset consists of Ribo-seq performed in lymphoblastoids in human, macaque and chimpanzee. A matched RNA-seq study is available for this dataset from Khan *et al.* [281] (SRA accession SRP028612) which we utilize for calculating translational efficiency. Translational efficiency of a gene is reflective of the density of ribosomes attached to the mRNA for every transcript. We processed mapped Ribo-seq datasets through ribotricer and only retained actively-translating fragments. Translational efficiency was calculated as the ratio of Ribo-seq to RNA-seq counts over a gene. uORFs were also searched using ribotricer (See Methods).

To asses the regulatory impact of uORFs on translational efficiency, we compared the translational efficiency of all genes with different number of actively translating uORFs (0, 1, 2, 3, and above) and plotted its cumulative distribution across all the three species (Figure 7.2).

Genes with higher number of uORFs tend to have reduced translational efficiency as compared to transcripts with lesser number of uORFs across human, macaque and chimpanzee. We tested using a Mann Whitney test. Genes with no actively translating uORFs have significantly higher translational efficiency as compared to the genes with 1 or more ORFs.

Among the orthologous genes, human and chimpanzee have 180 translationally efficient genes ($p < 0.05$) while 475 genes are translationally efficient between human and macaque. These numbers are in line with our expectation based on phylogenetic relationship between the three species.

We also wanted to assess if the uORF mediated regulatory activity is conserved across these species. To this end, we compared the distribution of uORF translational efficiency to coding domain sequence (CDS) translational efficiency between human and chimpanzee and between human and macaque using one to one orthologous genes. The uORF activity is correlated across species with the correlation between human and chimpanzee as 0.45 (Figure 7.3a) and between human and macaque as 0.33 (Figure 7.3b), thus capturing the phylogenetic relationship between the two species.



(a)
Human

(b)
Chimpanzee

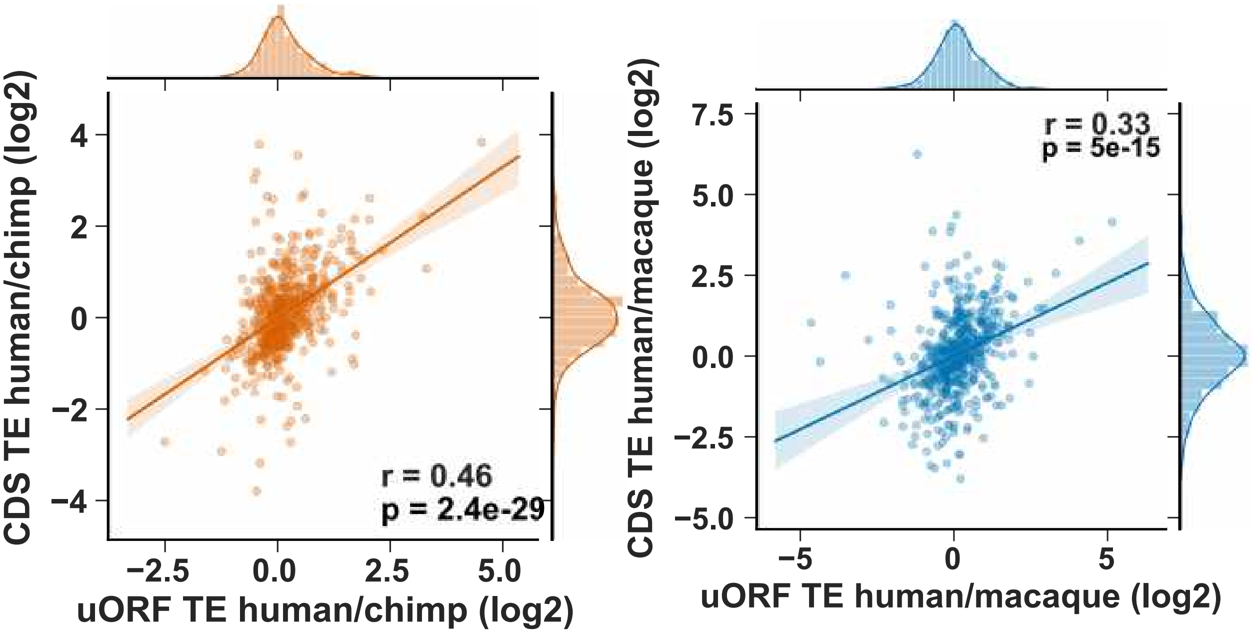
(c)
Macaque

(d)
Number of TE orthologous genes in chimpanzee and macaque

Figure 7.2
Higher number of actively translating uORFs result in down-regulation of the main ORF in (a) human (b) chimpanzee, and (c) macaque. (d) Number of differential translationally efficient genes in chimpanzee and macaque with respect to human.

7.2.2 uORFs are present across species

We characterized the prevalence of uORFs, by searching for "complete" ORFs in the 5' UTRs. A "complete" ORF is defined by the presence of a start codon (AUG) with an in-frame stop codon (UAG, UAA, and UGA). There is also evidence that translation can be initiated at non-AUG start codons in the uORFs can initiate at non-AUG start codons [282, 283]. However, unless the Ribo-seq experiment was prepared with a translation initiation inhibitor, it is difficult to determine the translation initiation site using Ribo-seq data alone. Thus, we took the more conservative approach of using only AUG start codons. We defined 'dORFs' (downstream ORFs) to be analogous of uORFs that are located in the 3' UTR. Not much is known about the regulatory role of the dORFs.



(a)
Chimpanzee

(b)
Macaque

Figure 7.3
Correlation between translational efficiency of CDS and uORF in orthologous regions of human, chimpanzee and macaque indicate the phylogenetic distances. Correlation values are indicated in the legend

We searched for start codons (AUG) in the 5' UTR and the first in-frame stop-codon to define the uORFs. Analogously, we searched for AUGs and the first in-frame stop codon in the 3' UTRs to define the dORFs. Since such uORFs and dORFs can overlap with coding domain sequences (CDS), we categorized them into different categories based on their overlap status (Table 7.1). For the rest of our discussion, we use 'uORFs' to refer to 'super_uORFs' which do not overlap with any CDS of any of the possible isoforms.

The relative distribution of dORFs is higher as compared to uORFs across species (Figure 7.4). Most genes have just one most upstream uORF or most downstream dORF present (Figure 7.5). The annotated CDS appear to be the longest regions, while the size distribution of uORFs and dORFs are similar and are around four times shorter than the annotated CDS (Figure 7.6).

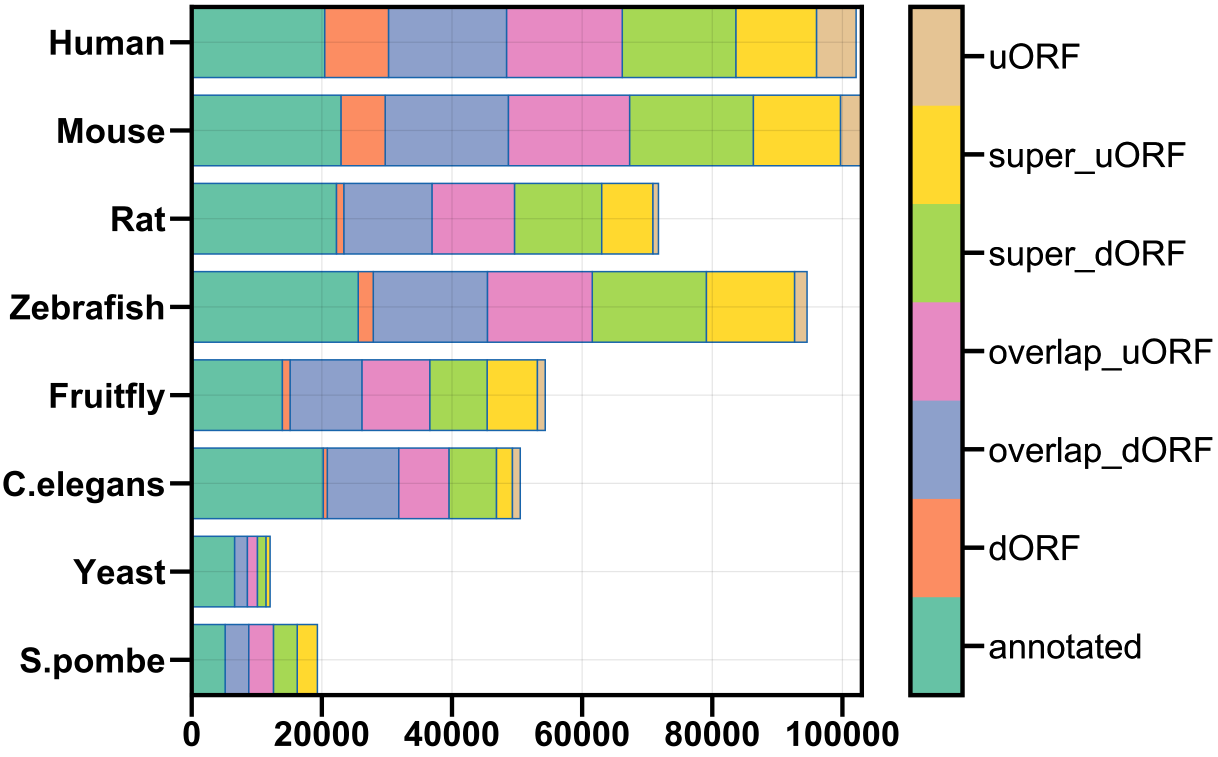


Figure 7.4
Distribution of number of known and potential ORFs. ORFs were searched using ribotricer [149]. super_uORF and super_dORF describe uORFs and dORFs that do not overlap with any CDS in any of the isoforms. The description of each ORF is given in Table 7.1.

ORF type	Description
annotated	CDS annotated in the provided GTF file
super_uORF	upstream ORF of the annotated CDS, not overlapping with any CDS of the same gene
super_dORF	downstream ORF of the annotated CDS, not overlapping with any CDS of the same gene
uORF	upstream ORF of the annotated CDS, not overlapping with the main CDS
dORF	downstream ORF of the annotated CDS, not overlapping with the main CDS
overlap_uORF	upstream ORF of the annotated CDS, overlapping with the main CDS
overlap_dORF	downstream ORF of the annotated CDS, overlapping with the main CDS
novel	ORF in non-coding genes or in non-coding transcripts of coding genes

Table 7.1
ORF types and their description. ORFs for each species were determined using ribotricer [149]

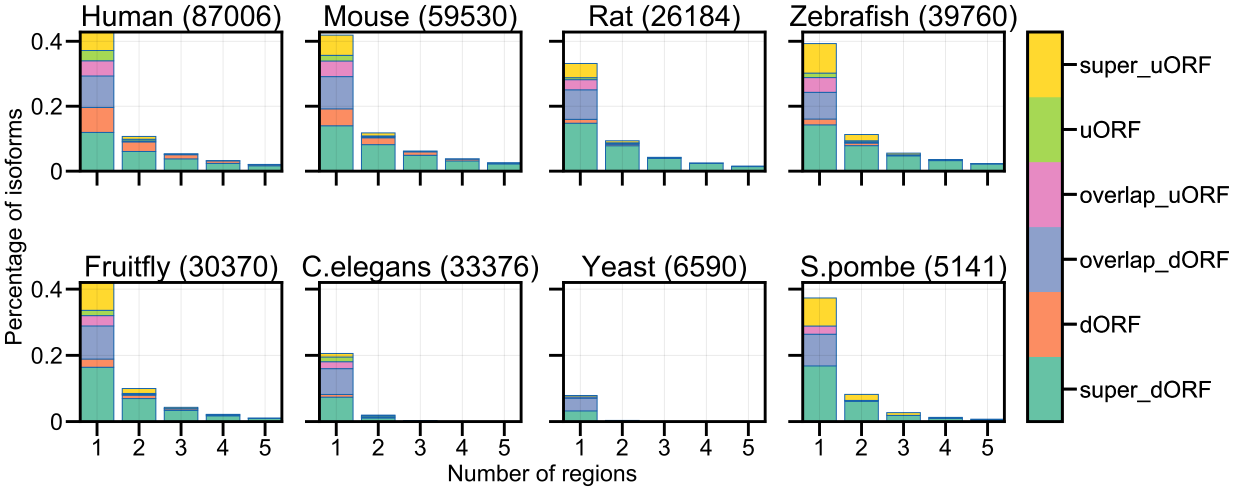


Figure 7.5
Distribution of number of different ORF types per annotated CDS. For each annotated CDS number uORFs and dORFs were searched using ribotricer by looking for a start codon and an in-frame stop codon in the 5'UTR and 3'UTR regions respectively.

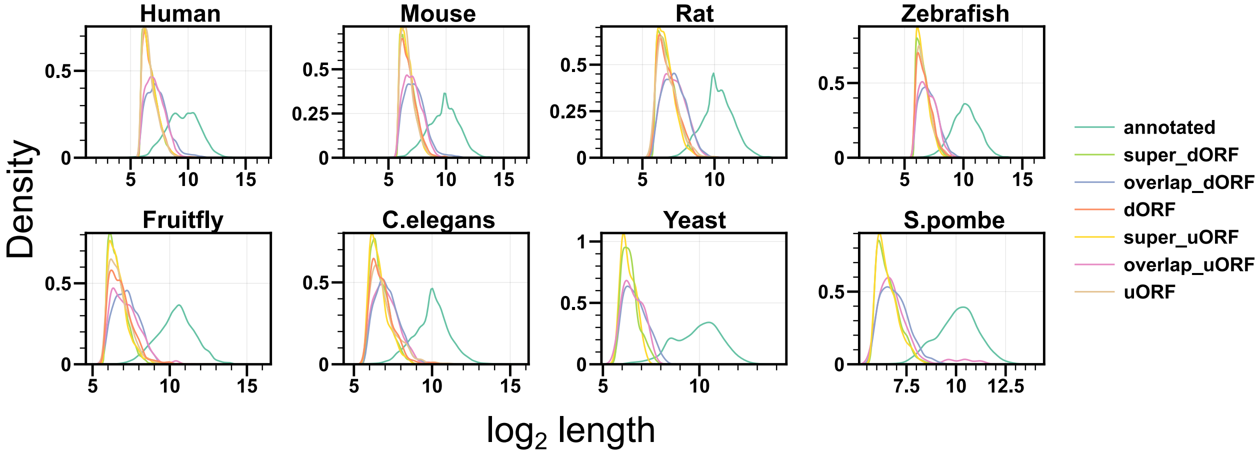


Figure 7.6
Distribution of lengths of different ORF types. Kernel density estimates of the length of each ORF type on log_2 scale across species.

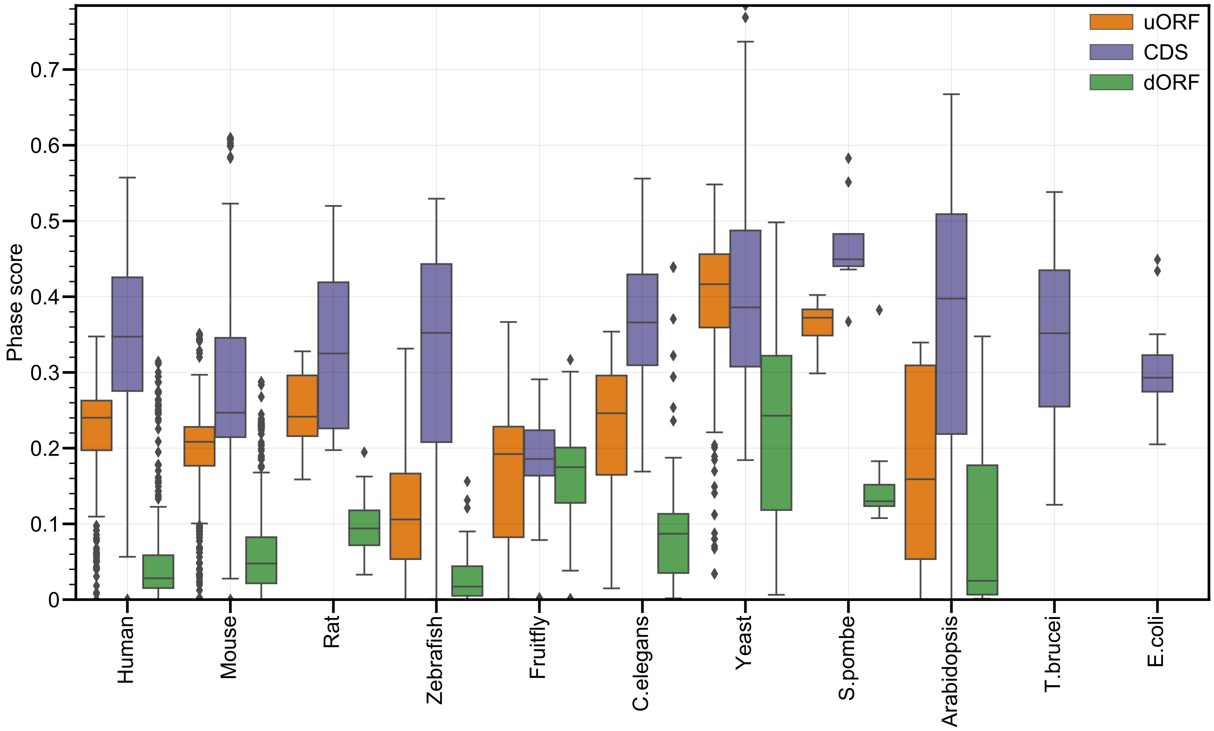


Figure 7.7
Distribution of phase scores in uORF, CDS and dORF. Phase scores were calculated using ribotricer across all the datasets available on bit.ly/ribopod-datasets.

7.2.3 More uORFs than dORFs show active translation

We calculated the phase score of each ORF generated through ribotricer [149]. The phase score is a reflective of the active-translation status of an ORF being translated based on the periodicity in its Ribo-seq profile. A higher phase score indicates active-translation. Across all the species, CDS has the highest phase scores followed by uORFs and dORFs (Figure 7.7). Though the abundance of dORFs is higher as compared to uORFs (Figure 7.4), the uORFs seem to be translated more often. There is a modest positive correlation $(r = 0.33, p < 1e - 6)$ between phase scores of CDS and uORFs implying a uORF translation does not necessarily shutdown the translation of the downstream CDS and a modest negative correlation $(r = -0.26, p < 1e - 6)$ between phase scores of CDS and dORFs (Figures 7.10 and 7.10).

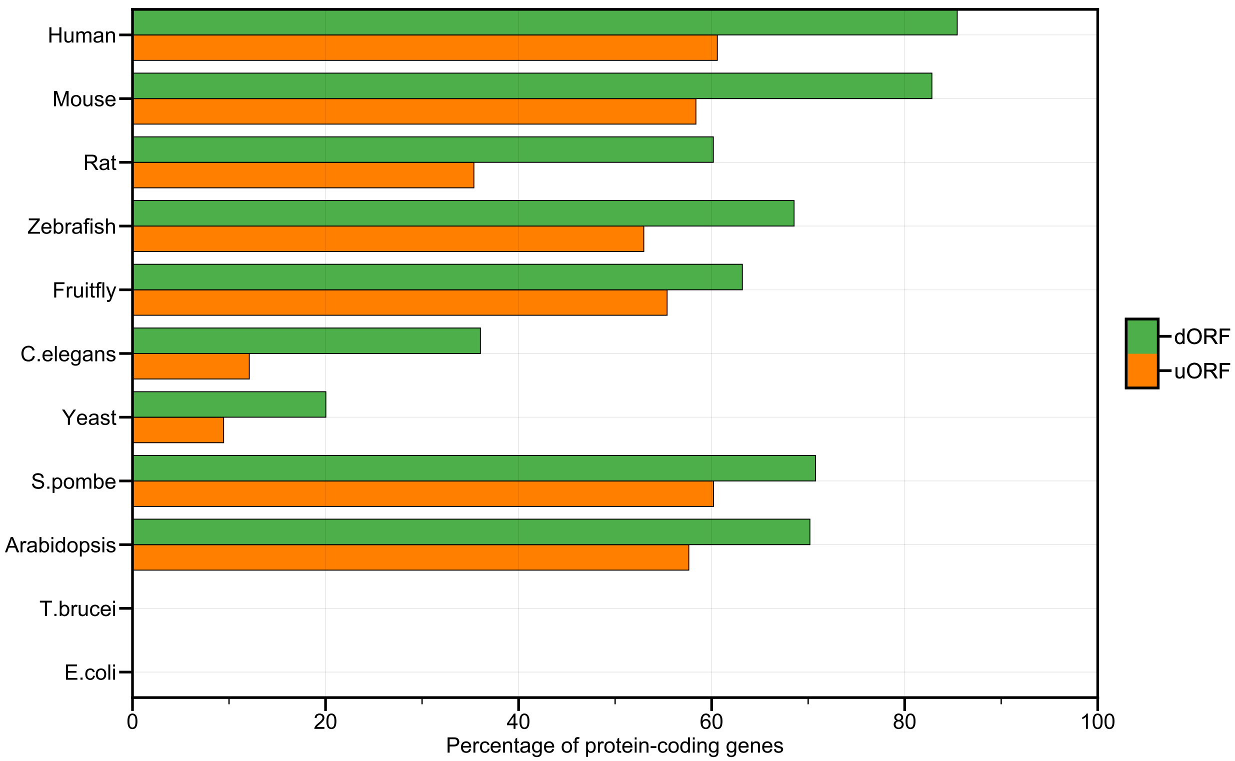


Figure 7.8
Distribution of percentage of uORFs and dORFs under active translation with respected to coding genes. Proportion of uORFs and dORFs with respect to annotated CDS regions that are almost always translating.

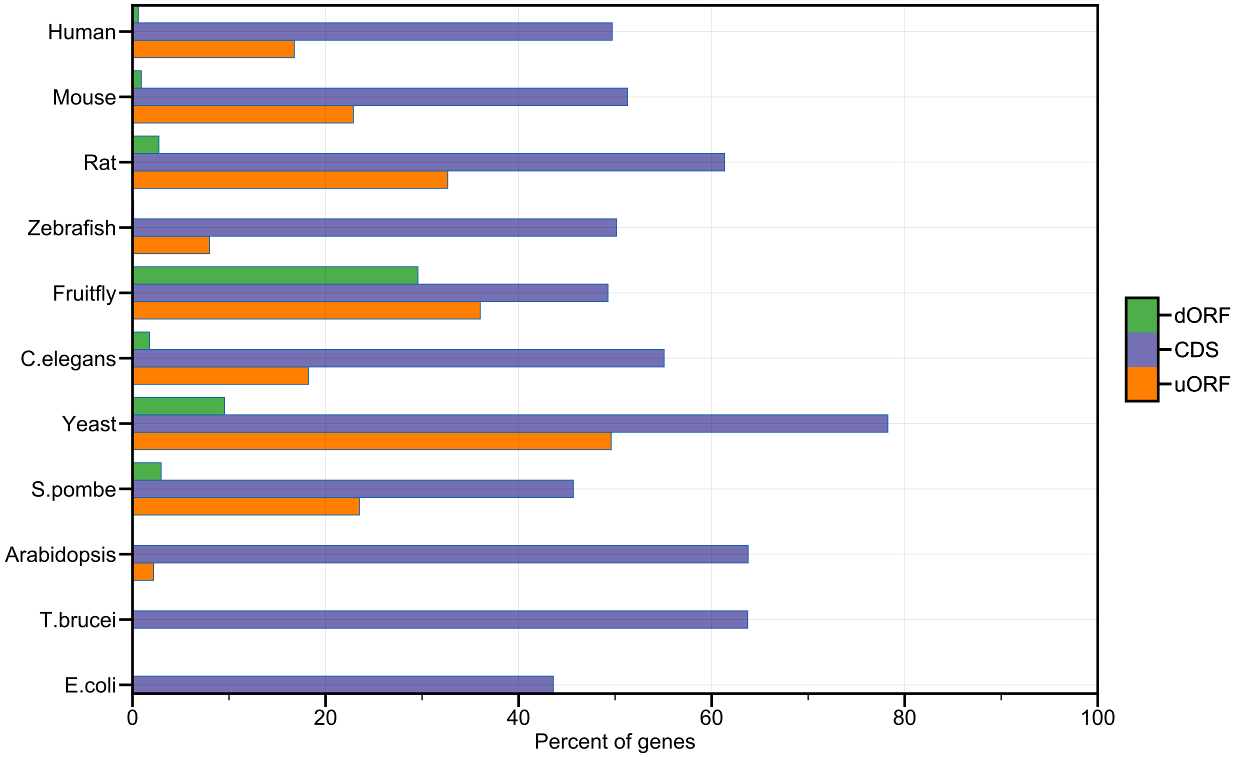


Figure 7.9

Distribution of uORFs and dORFs that are always translating with respect to coding genes. Percentages of uORFs and dORFs that are always translating as percentage of protein coding genes.

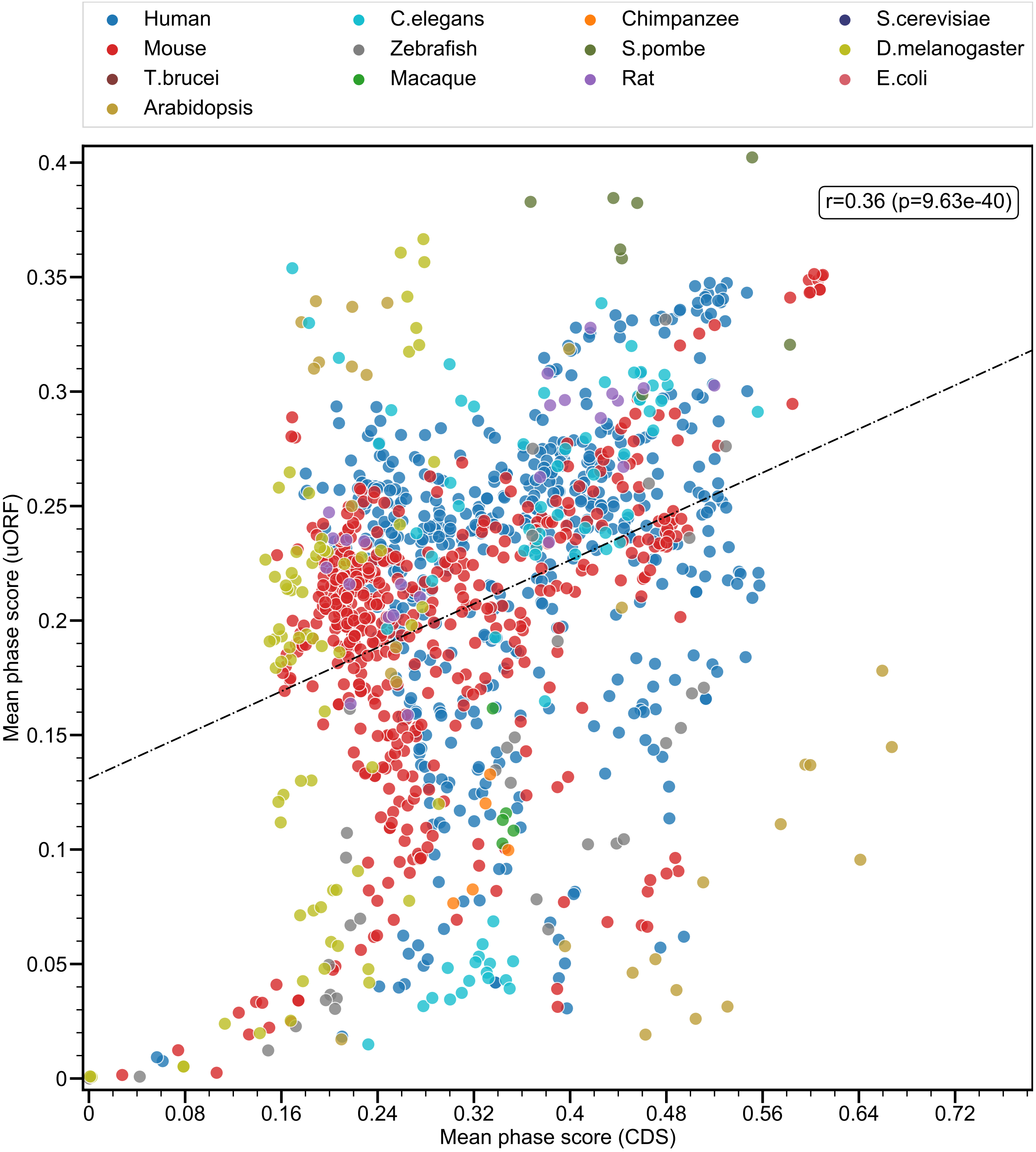


Figure 7.10
Correlation between uORF and CDS phase scores.

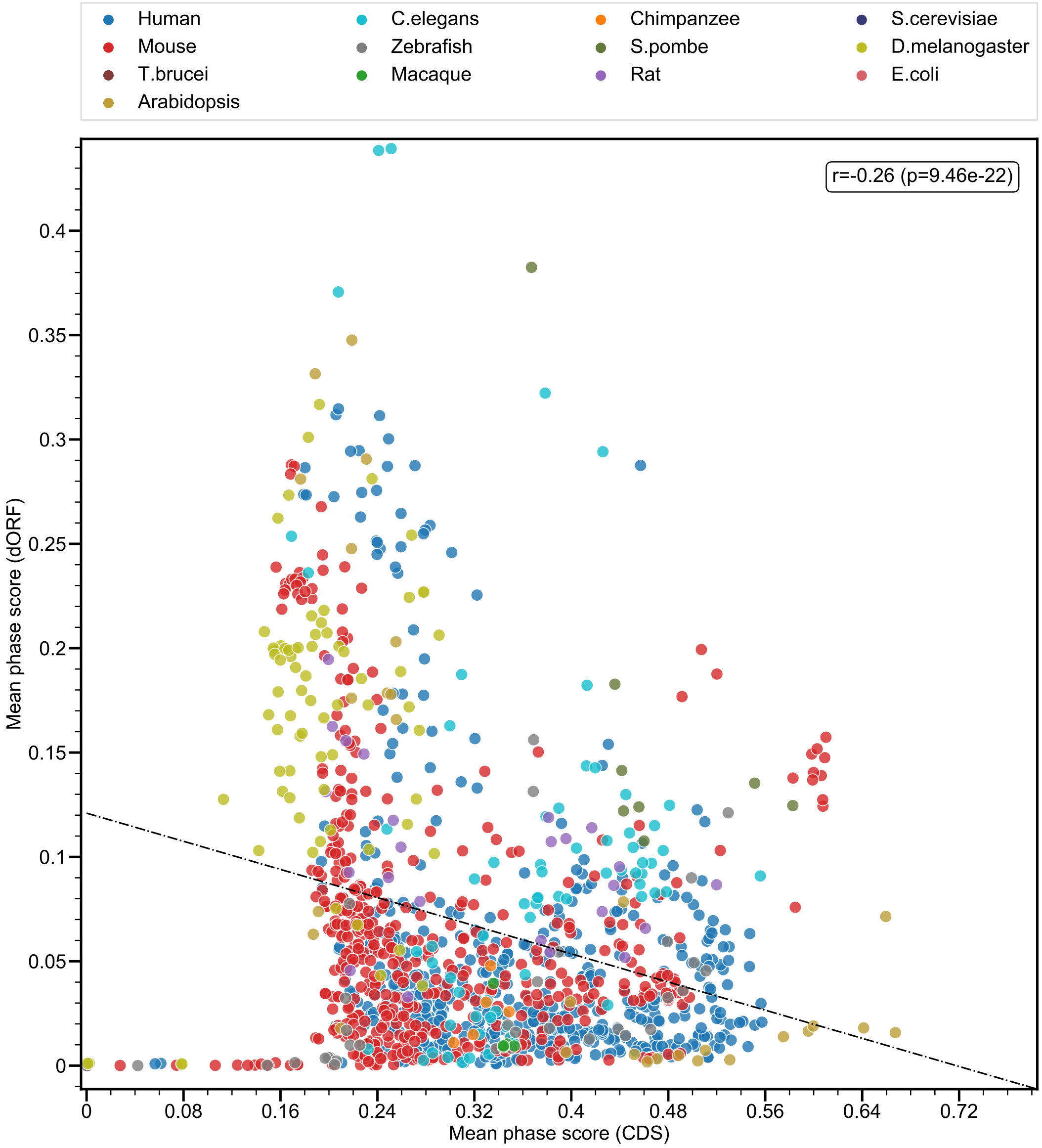


Figure 7.11
Correlation between dORF and CDS phase scores.

7.2.4 uuORFs - uORFs that are almost always translating

Having characterized the translational status of uORFs across species in different physiological and pathological contexts we focused on uORFs that are 'almost always' under active translation. The definition of 'almost always' here is relative to the protein-coding regions. Because of the inherent heterogeneity in the Ribo-seq data itself and given that the datasets are chosen across multiple physiological and pathological contexts, even the high-confidence protein coding regions are not translating always. First, following [149], we learned the species specific for all the datasets cutoff for determining if an ORF is actively-translating. We determined ribotricer generated phase score for each annotated CDS and designated the median of this score calculated across all CDS spanning all datasets as the species-specific cutoff (Table 7.2).

Most of the CDS regions are actively translating in do not exhibit translation in around $60 - 70\%$ of the CDS regions (Figure 7.12). The proportion of samples in which all the protein coding regions are actively-translating, i.e. they meet the minimum phase score cutoff for the corresponding species follows a bimodal distribution (Figure 7.12 and 7.13). We use the proportion corresponding to these peaks to classify any ORF as 'always' translating and 'never' translating. In particular, we use a gaussian mixture model to resolve the mean and variance parameters of the two mixtures for each species. The higher mean gaussian ($\mathcal{N}(\mu_1, \sigma_1)$) will be used to define the 'always' translating and the lower the mean gaussian ($\mathcal{N}(\mu_0, \sigma_0)$) will be used to define the 'never' translating regions. We focus on the $100(1-\alpha)\%$, where $0 < \alpha < 1$ quantile of $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and let z_1 be such that $P(X \leq z_1) = 1 - \alpha$ or equivalently, $P(X \geq z_1) = \alpha$. This defines our universal set - this value of the proportion determines the universality of an ORF. We now call an uORF to be universal as long as the proportion

assembly	mean	median	std. dev
Human	0.355	0.348	0.299
Mouse	0.286	0.255	0.295
Rat	0.327	0.379	0.282
Zebrafish	0.323	0.272	0.335
Drosophila	0.190	0.117	0.255
C. elegans	0.364	0.336	0.270
S. cerevisiae	0.411	0.436	0.225
S. pombe	0.467	0.467	0.149

Table 7.2
Species-specific cutoffs for active translation. The mean median and standard deviation of ribotricer generated phase score as calculated using only protein coding regions for all datasets used in this study. Any ORF in a species is labelled as actively-translating only if its phase score exceeds the median value for the corresponding sites.

of samples which meet the threshold is higher than z_1. This is the most conservative way of defining them, given that this represents the extreme tails of the distribution of CDS which are believed to be actively translating with a higher probability than uORFs. Similarly, for defining the 'never' translating regions, we rely on the middle most values of the gaussian with lower mean $X_0 \sim \mathcal{N}(\mu_0, \sigma_0)$ ensuring we sample similar proportion. In particular, we look for a value of d, such that $P(\mu_0 - d \leq X_0 \leq \mu_0 + d) = \alpha$. We then sample this region from the uuORFs distribution to define 'never' translating regions. We use $\alpha = 0.95$ for this study.

7.2.5 Conservation of uuORFs across species

To identify uuORFs that are conserved across any two species, we *translated* the uORFs to their corresponding amino acid sequences. Considering one species as the reference, we used BLAST [284] to search for conserved uuORFs. In particular, the references species' uuORFs are treated as the target database for blast and then the amino acid sequence of each species is used

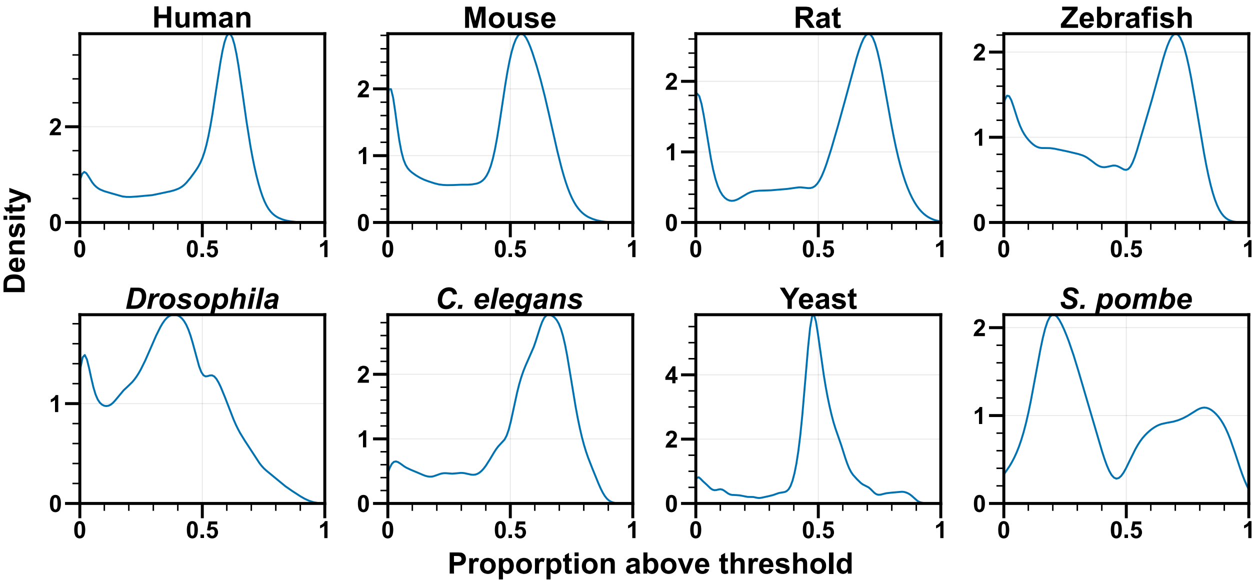


Figure 7.12
Distribution of phase scores of CDS regions that are above the minimum score required for active translation. Kernel density estimates of the phase scores of CDS regions that are above the threshold phase score required for the ORF of the corresponding species to be labelled as active translation.

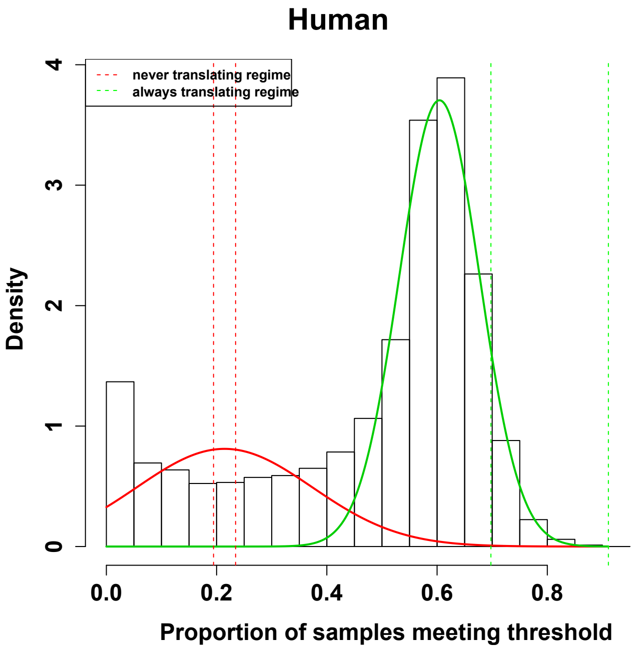


Figure 7.13
Determining thresholds for ‘always active’ translation.

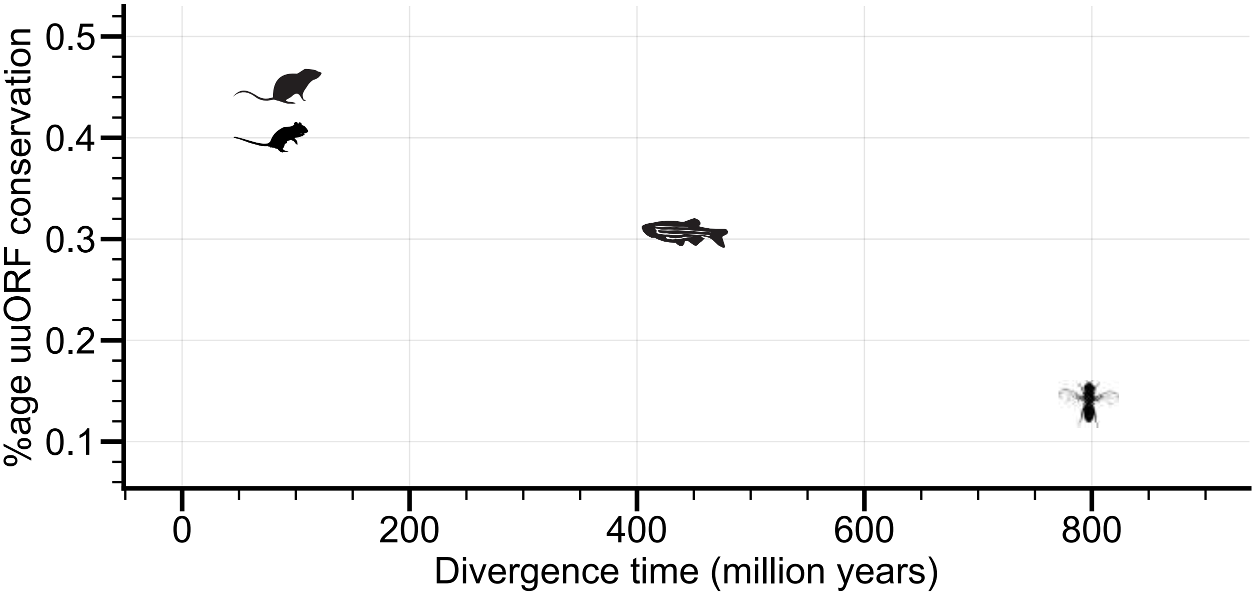


Figure 7.14
Percentage of conserved uuORFs. Human is used as the reference sequence. A uuORF is said to be conserved if there is a unique hit for the corresponding amino acid sequence in the database of human uuORFs.

as the query sequence. We restrict our search to unique hits (`-max_target_seqs 1` , `-evalue 1e-3`) to define 'orthologous' uuORFs. Around 45% of uuORFs are conserved between human and mouse, 40% are conserved between human and rat, 30% is conserved between zebrafish and human and 13% is conserved between human and fruitfly (Figure 7.14).

Next, we hypothesized that the upstream sequence context of these uuORFs might give them a preferential context for translation initiation. The Kozak consensus sequence is a nucleic acid sequence that is present upstream of the start codon [55]. It is considered as the optimum sequence required to be present upstream to initiate translation in vertebrates. In prokaryotes, a similar sequence is present around eight bases upstream referred to as the Shine-Dalgarno sequence [285]. We hypothesized that the uuORFs might have a similar sequence context. Chew *et al.* [286] indeed found evidence for a Kozak like sequence context responsible for increased effect on the translational efficiency of the main CDS.

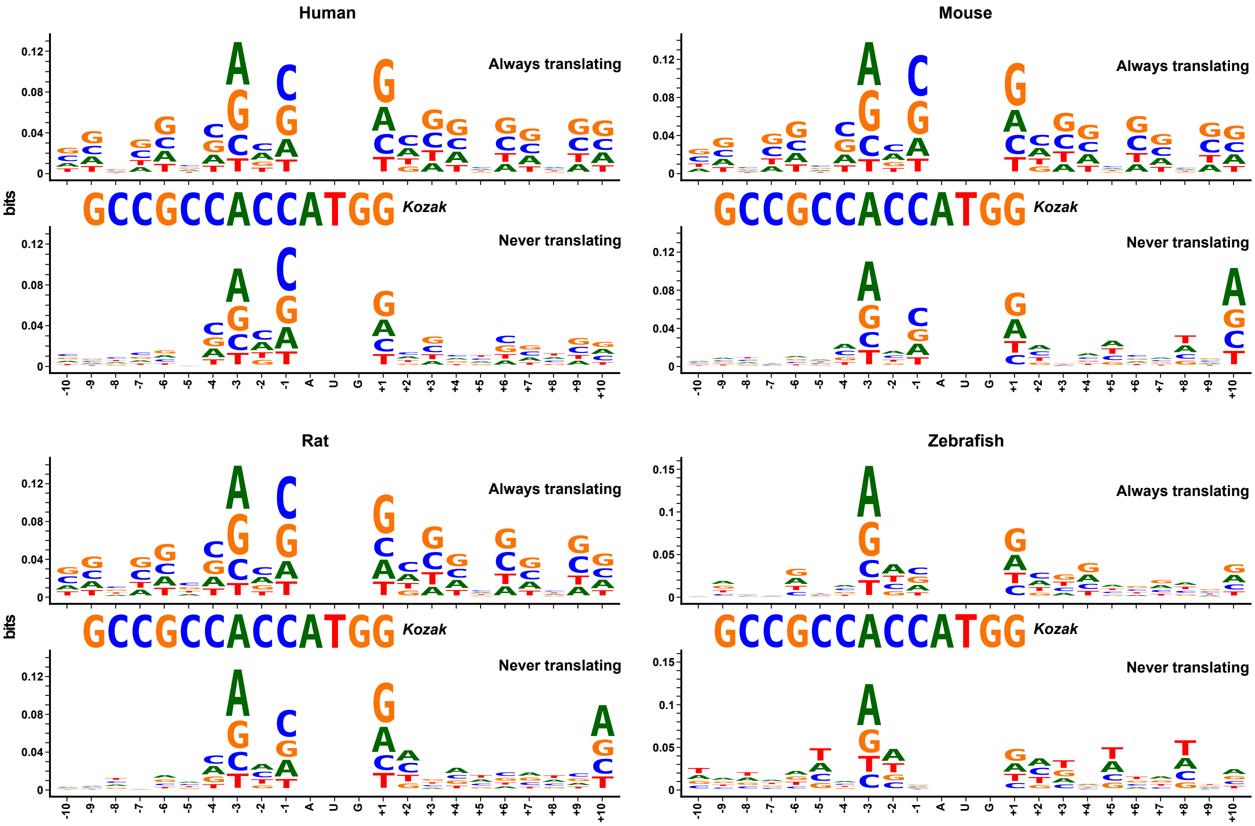


Figure 7.15
Upstream sequence context of CDS in vertebrates.

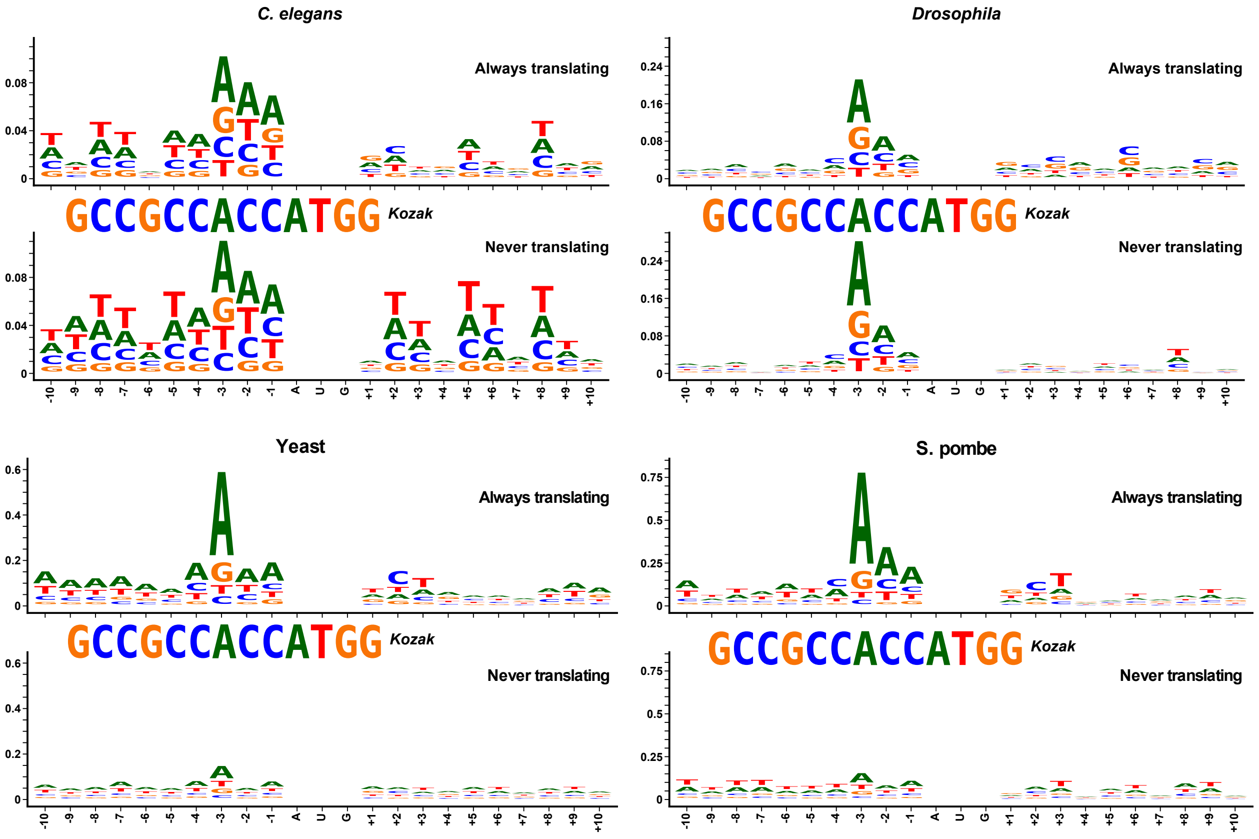


Figure 7.16
Upstream sequence context of CDS in invertebrates.

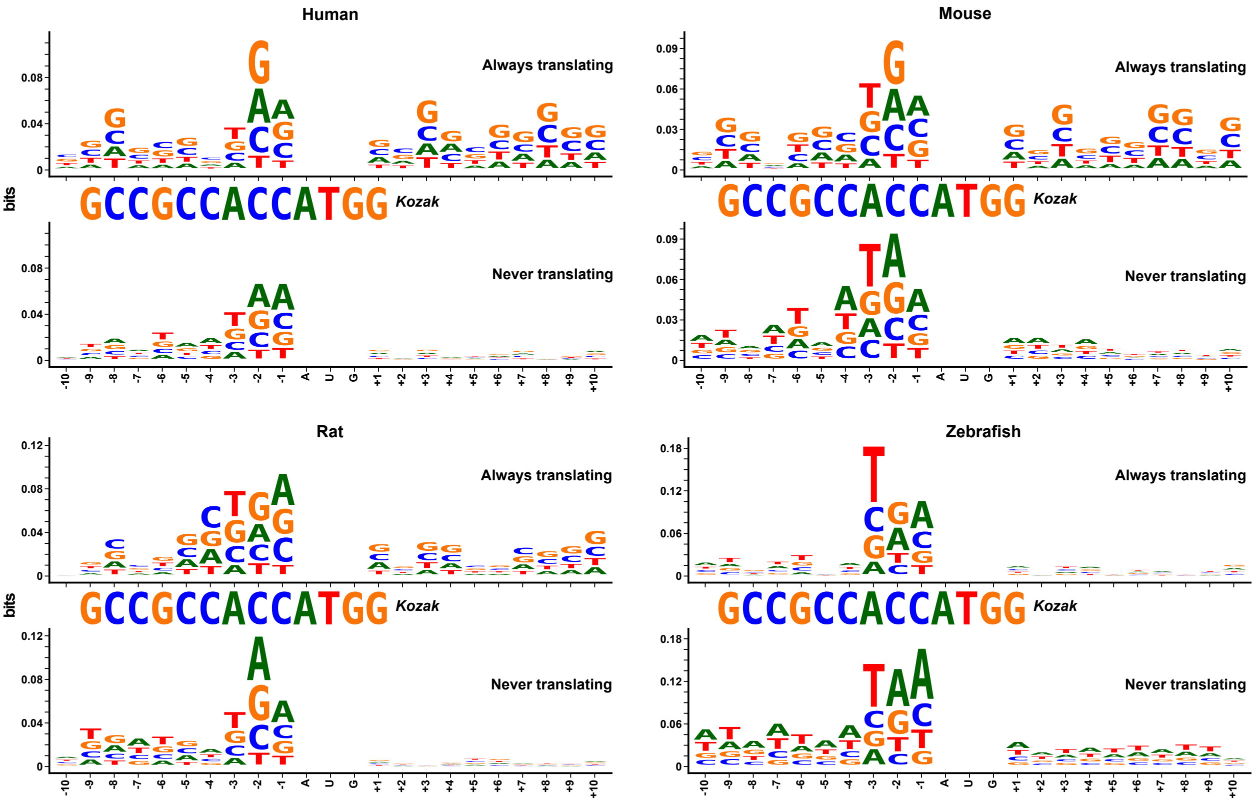


Figure 7.17

Sequence similarity of upstream sequence contexts of always and never translating uORFs in vertebrates. Clustering was performed using pearson correlation coefficient on the position frequency matrix [287].

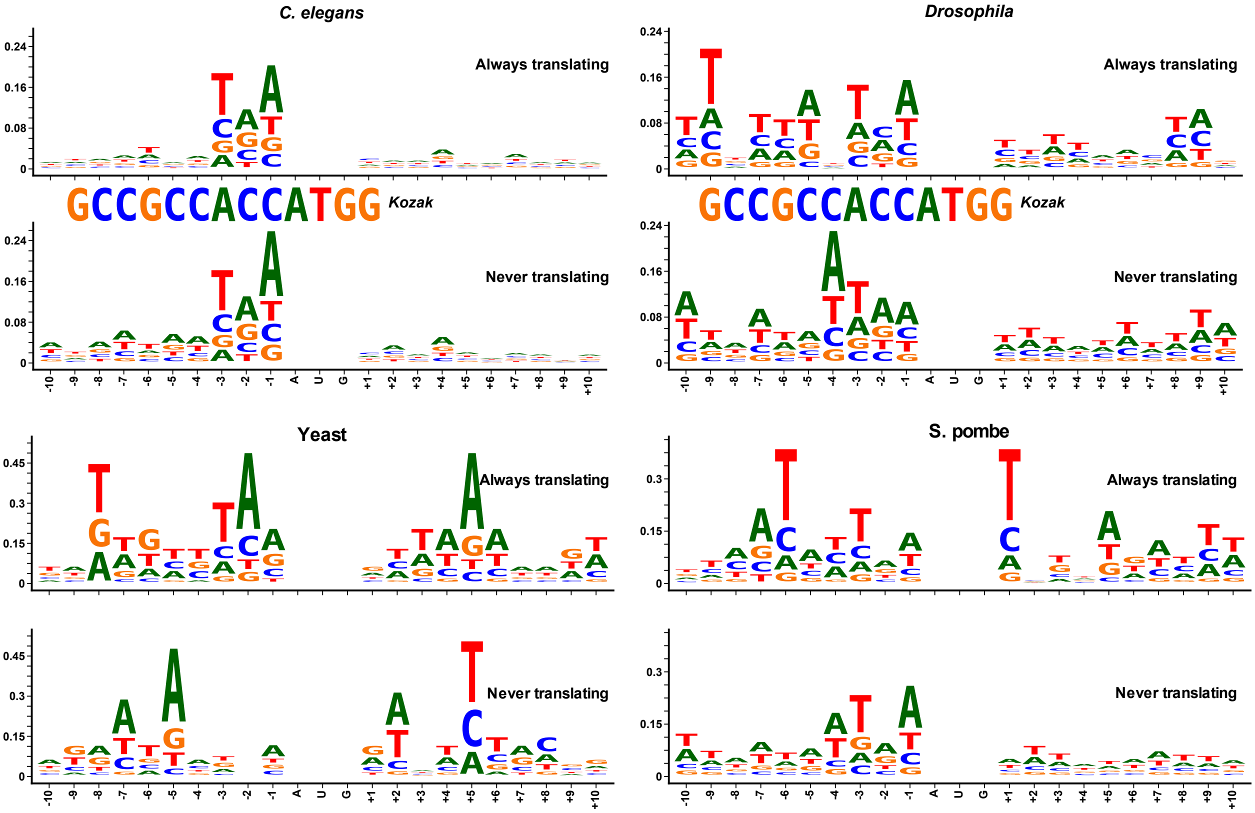


Figure 7.18

Sequence similarity of upstream sequence contexts of always and never translating uORFs in invertebrates. Clustering was performed using pearson correlation coefficient on the position frequency matrix [287].

We determined the 13 nucleotides sequence present upstream of the uuORFs across all species. Our null set is formed by the upstream sequence of those uORFs that are never translating. The sequence context of the never translating uORFs is different from that of the never translating uORFs indicating that the different sequence context might be responsible for initiating translation in the uuORFs (Figures 7.17 and 7.18). Furthermore, the sequence contexts in vertebrates is different from the Kozak sequence context (Figure 7.17) while the sequence context in invertebrates is A and T rich, distinct from both the Kozak and vertebrates' uuORF sequence context.

Next, we investigated if the sequence contexts are similar across species. We make use of the pearson correlation coefficient [287] to quantify the similarity between two sequence motifs. Given two column vectors X and Y representing the frequencies of base frequencies of bases A, C, T and G, Pearson correlation coefficient (PCC) is defined as:

$$PCC(X,Y) = \frac{\sum_{a \in \mathcal{A}} (X_a - \bar{X})(Y_a - \bar{Y})}{\sqrt{\sum_{a \in \mathcal{A}} (X_a - \bar{X})(Y_a - \bar{Y})}},$$

$$\bar{X} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} X_a,$$

$$\bar{Y} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} Y_a.$$

A high PCC value between two sequence motifs reflects their similarity. In order to assess the similarity of the uuORFs motifs we performed hierarchical clusterintg using PCC as the similarity metric. The uuORFs are similar across vertebrates and across invertebrates (Figure 7.19 and

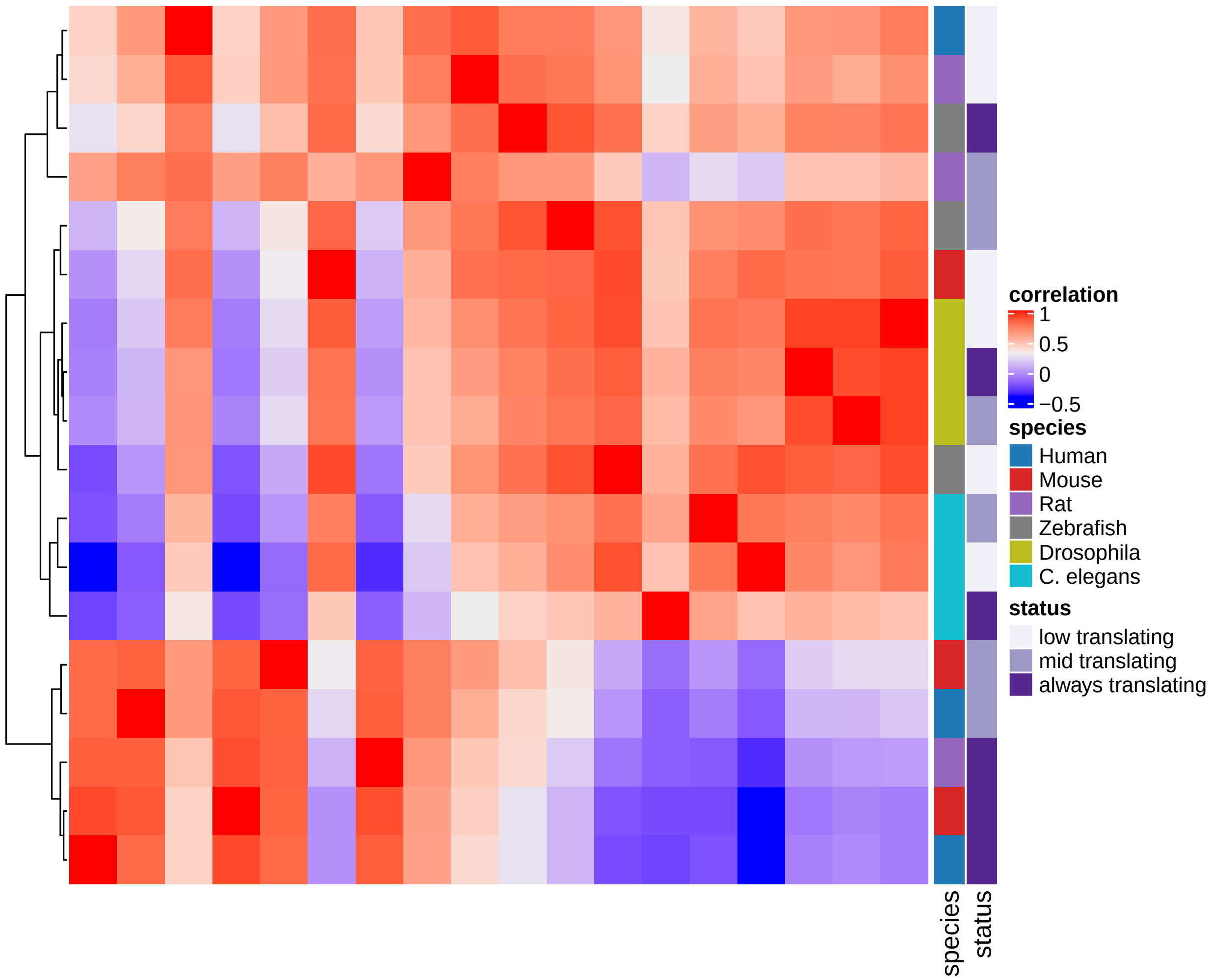


Figure 7.19
Sequence similarity of upstream sequence contexts of always and never translating CDS. Clustering was performed using pearson correlation coefficient on the position frequency matrix [287].

7.20). The uuORFs also show higher similar similarity to the sequence motifs upstream of CDS that are always translating both in vertebrates and in invertebrates (Figure 7.19 and 7.20).

7.3 Discussion

uORFs have been known to play role in regulating translation initiation of the coding domain sequence. Most of the studies thus far have focused on *potential* uORFs, based on their sequence

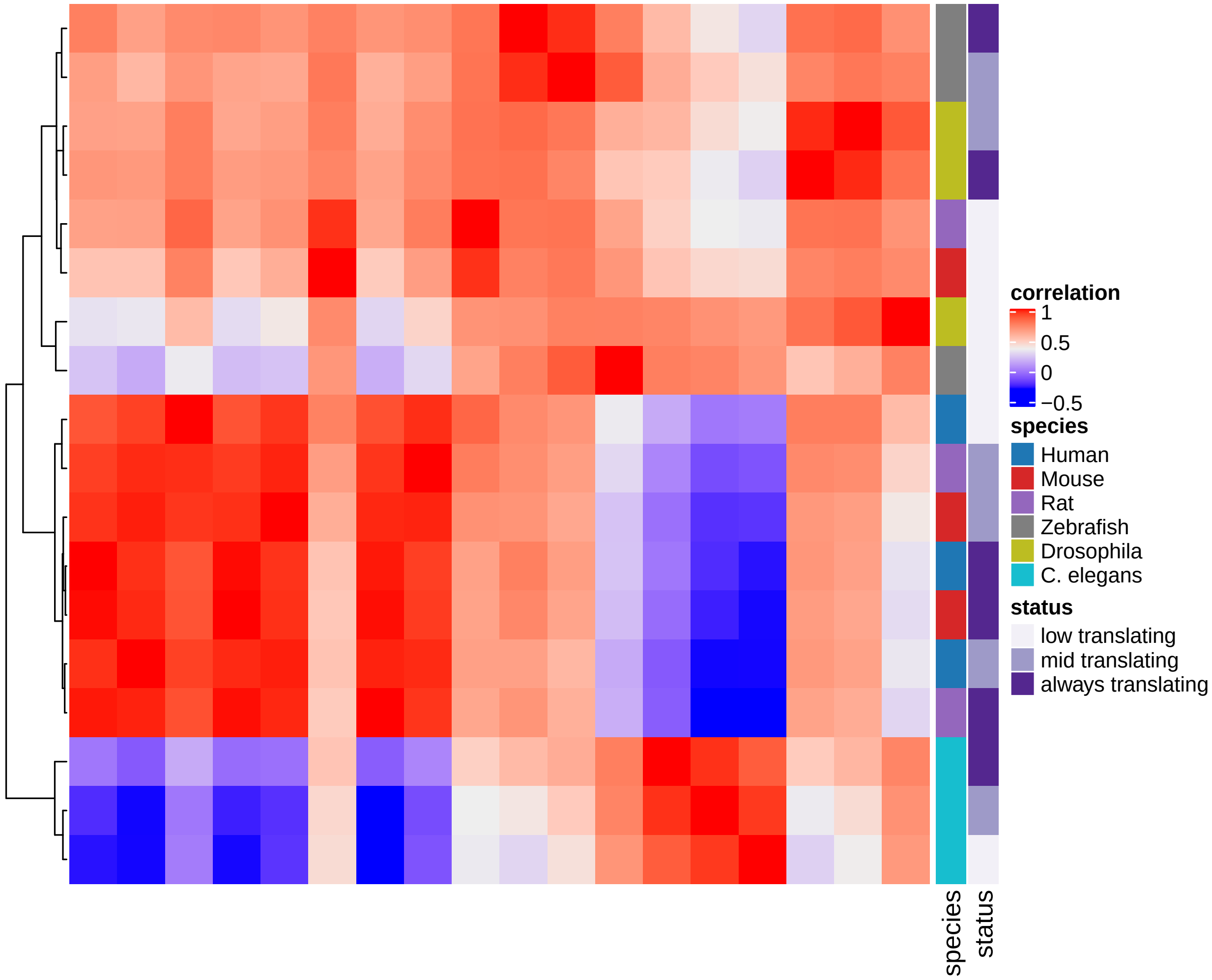


Figure 7.20
Sequence similarity of upstream sequence contexts of always and never translating CDS. Clustering was performed using pearson correlation coefficient on the position frequency matrix [287].

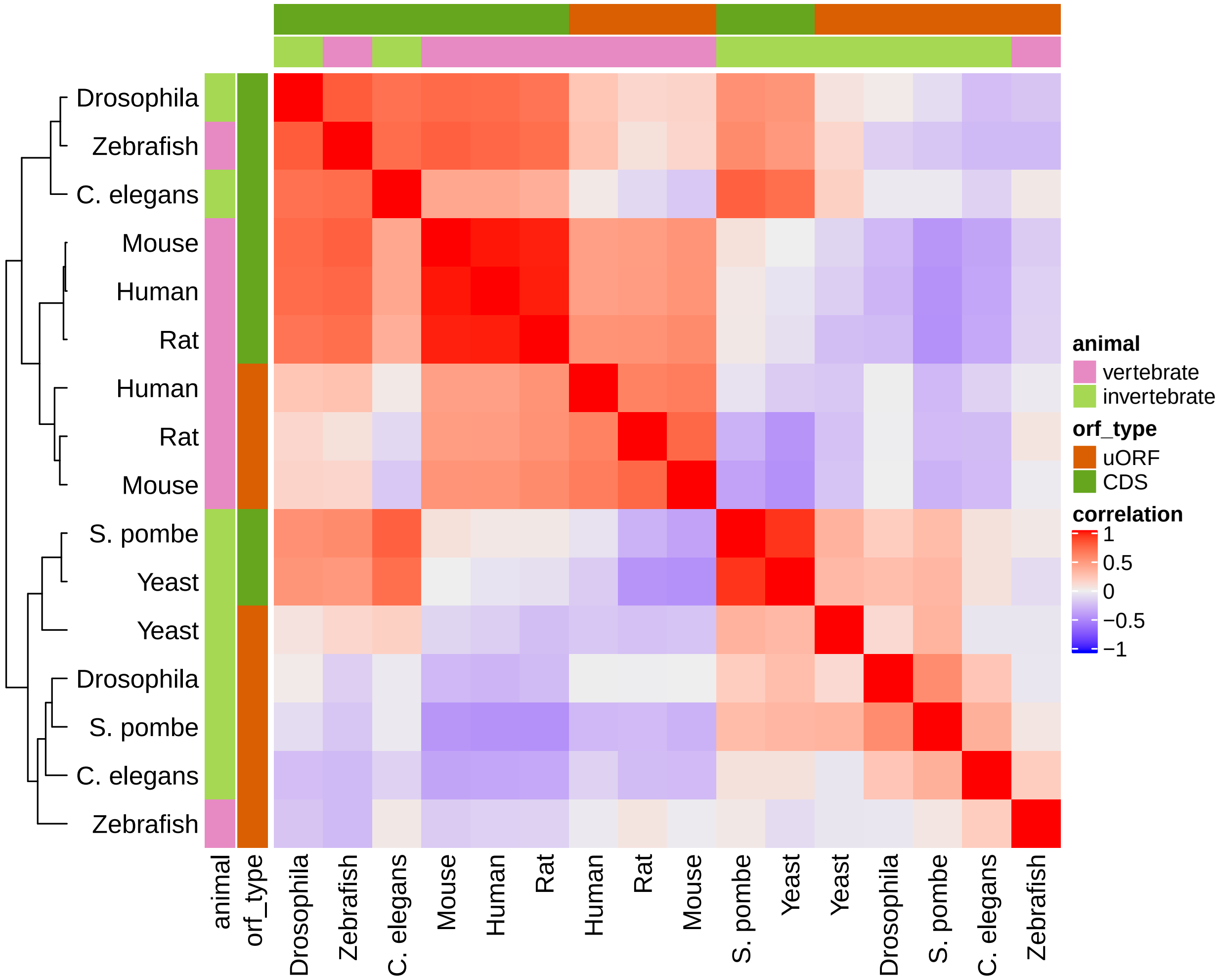


Figure 7.21
Sequence similarity of upstream sequence contexts of always and never translating CDS and uORFs.
Clustering was performed using pearson correlation coefficient on the position frequency matrix [287]. The translating motifs are similar across CDS and uORFs in both vertebrates and invertebrates.

Species	cutoff (never)	cutoff (always)	mean (never)	sigma (never)	mean (always)	sigma (always)
Human	0.194-0.234	0.698-1	0.214	0.159	0.604	0.073
Mouse	0.080-0.104	0.680-1	0.092	0.098	0.550	0.101
Rat	0.111-0.151	0.806-1	0.131	0.158	0.685	0.094
Zebrafish	0.178-0.219	0.790-1	0.199	0.164	0.684	0.0826
Drosophila	0.009-0.012	0.631-1	0.010	0.011	0.383	0.193
C. elegans	0.177-0.212	0.769-1	0.195	0.138	0.637	0.103
Yeast	0.408-0.474	0.564-1	0.441	0.264	0.501	0.05
S. pombe	0.205-0.229	0.901-1	0.217	0.094	0.738	0.127

Table 7.3
Thresholds for defining universal uORFs The cutoffs are learned as the value corresponding to 95 percentile

Species	Status	Motif
Human	Always translating	`CGGGGCCTGA_ATG_GTGGCGGCCG`
Human	Never translating	`ATAATAATAA_ATG_GCGATGGCGG`
Mouse	Always translating	`GGGCGGCGGG_ATG_GCGGGGGCCG`
Mouse	Never translating	`ATAATAATAA_ATG_AATATTTTTA`
Rat	Always translating	`GGCCGGCTGA_ATG_GCGGCCCCCC`
Rat	Never translating	`ATGATGATAA_ATG_GGTGTGATCG`
Zebrafish	Always translating	`GCTCCGGAGA_ATG_CGCTGTATTA`
Zebrafish	Never translating	`ATAATAATAA_ATG_AATATTATTA`
Drosophila	Always translating	`CTCACAATAA_ATG_TGTACAACTA`
Drosophila	Never translating	`ATAATAATAA_ATG_AAAAAAAATA`
C. elegans	Always translating	`TCAACGGACA_ATG_TGCCTGAAAA`
C. elegans	Never translating	`ATAATAATAA_ATG_TTAATTTTTA`
Yeast	Always translating	`TATAGTTTAA_ATG_GTTAAAAAGA`
Yeast	Never translating	`AGAATAAGAA_ATG_GAATTTACCG`
S. pombe	Always translating	`TTAATATTAA_ATG_TCGAAGATTT`
S. pombe	Never translating	`ATAATAATAA_ATG_ATTATAATTA`

Table 7.4
Sequence context of uuORFs Start codon is shighlighed in bold.

Species	Status	Motif
Human	Always translating	CGCCGCCACC_ATG_GCGGAGGTGG
Human	Never translating	CGCCGCCACC_ATG_GCGCTCGTGG
Mouse	Always translating	GGCGGCCAAC_ATG_GCGGCGGTGG
Mouse	Never translating	AATTAAAAAC_ATG_GAGAACATGA
Rat	Always translating	CGCGGCCACC_ATG_GCGGCGGCGG
Rat	Never translating	CATCGACAAC_ATG_GAGATCATCA
Zebrafish	Always translating	AGAAGTCAAC_ATG_GCGGAGGAGG
Zebrafish	Never translating	TATAGTAAAG_ATG_GATATTATGA
Drosophila	Always translating	AAACAACAAA_ATG_GCCAACAACG
Drosophila	Never translating	AAAAATAAAA_ATG_ACTATCATTA
C. elegans	Always translating	TATTGTAAAA_ATG_GCTGTCGTCG
C. elegans	Never translating	TATTTTAAAA_ATG_ATTATTATTA
Yeast	Always translating	AAAAAAAAAA_ATG_TCTAAAAAAA
Yeast	Never translating	AAAAAAAAAA_ATG_AATAATAATA
S. pombe	Always translating	ATTAATCAAA_ATG_GCTGATAATA
S. pombe	Never translating	TTTTTTTAAA_ATG_GATAATAATA

Table 7.5

Sequence context of CDS. Start codon is surrounded by underscores.

context. We bridge this gap by providing a catalogue of uORFs with their translation status across different contexts using public Ribo-seq datasets across multiple species. We use this learned information to investigate the conservation of regulatory roles of uORFs across species which further led to us characterizing uuORFs - uORFs that are almost always translating across different physiological and pathological conditions.

The upstream sequence context of uuORFs in vertebrates is different from the well characterized Kozak sequence context [55] required for translation initiation in vertebrates. The Kozak sequence is given by GCCGCC(A/G)CCAUGG and was initially assumed to be a universal signal required for initiating translation. Later, however, it was found that the sequence context is species dependent. Amongst the characterized species, the sequence context is ACAACCAAAAUGGC for Drosophila, AAAAAAAAAAUGTC for Saccharomyces cerevisiae, and UAAAT(A/C)AACAUG(A/G)C for other invertebrates [54]. Our observations with respect to the sequence context of CDS are consistent with previous findings (Figures 7.15 and 7.16 and Table 7.5). We find that the sequence contexts of uuORFs have a similar context as the CDS of the corresponding species and are conserved within vertebrates and invertebrates (Tables 7.4 and 7.5). Though the consensus sequence is different from the respective expected sequence, we believe the translation initiation at uORFs might have a different mechanism than the corresponding CDS as has been observed previously [272]. The characterization of uuORFs and their associated upstream sequence context enables us to characterize the functional role of the short polypeptides. Our study provides a rich catalogue of these uuORFs which can be further used to investigate their functional implications by proteomics-based or other related approaches.

Chapter 8

Conclusions

Translational control is an integral part of the chain of processes that are employed in the gene to regulate the expression level of protein. The development of assays such as Ribo-seq have provided us with a window to peek into the transcriptome of the cell which is actively engaged by the ribosome thus enabling us to decipher the various modes by which a gene is regulated at the translational level. Ribo-seq has been used to understand translational regulation in a myriad of physiological and pathological contexts across species. The availability of these datasets with different contexts has presented us with avenues of discovering new biological phenomena that are conserved across different species.

The aim of my research was to understand the regulatory role of short open reading frames located in the 5' UTR called upstream open reading frames (uORFs). uORFs have been known to play a repressive role in translation. The previous research in this field is based on the assumption that a mere presence of an ORF characterized by an in-frame start and stop-codon, can be called an uORF. Ribo-seq provides actual evidence if such a potential ORF is actually an uORF, since it allows to evaluate its translation status. The length distribution of uORFs is similar to that of coding exons and being a region of low signal to noise ratio, detecting active translation in uORFs using Ribo-seq data has remained challenging.

In Chapter 3, I developed a novel method, ribotricer, for identifying active translating in both short and long open reading frames using Ribo-seq data. Our method takes an approach of evaluating the inherent periodicity in the Ribo-seq data through the 'high-low-low' pattern exhibited by an actively-translating ribosome as it traverses the mRNA. This approach overcomes the two key problems inherently present in almost all Ribo-seq datasets: uneven coverage and sparsity. While uneven coverage arises because the ribosome moves non-uniformly over the codons and hence is more of biological nature in its origin, sparsity is more of a technical limitation in a lot of experiments. Our approach of transforming the counts information to asses the consistency of qualitative 'high-low-low' pattern gives the highest accuracy on multiple datasets across different species.

In Chapter 4, I use our method, ribotricer, to learn the changes in the translational landscape of a fungal pathogen *Candida albicans* during its morphological transition from yeast-like to filamentous growth. Using deep sequenced Ribo-seq and RNA-seq samples, I first re-learned the transcriptome of *C. albicans* since the available annotation is incomplete. Through this process, I discovered unannotated exons that are also engaged actively by the ribosome implying that these could be genes that were unannotated. I also discover hundreds of genes whose translational efficiency gets upregulated during the morphological transition. These genes could serve as potentially good targets anti-fungal drugs.

In Chapter 5, by an integrated analysis in two glioblastoma cell lines, I mapped the changes that occur in gene expression at three levels of transcription, alternative splicing , and translation in response to ionizing radiation by utilizing Ribo-seq and RNA-seq data. Glioblastoma is the most common intracranial malignant brain tumor with an aggressive clinical course. High-dose radiation is the main component of glioblastoma therapy. By characterizing the alterations at all three levels, I identified new biological processes that lead to altered expression of various oncogenic factors. I also suggested new target options that can increase radiation sensitivity and prevent relapse.

I reanalyzed public Ribo-seq datasets from multiple species to distinguish ORFs that are under active translation from those that are not, using our tool ribotricer. In Chapter 6, I introduce a database of the reanalyzed public Ribo-seq projects that provides users direct access to de-noised Ribo-seq datasets.

Finally, in Chapter 7, I utilized the resource developed in Chapter 6 to study upstream open reading frame (uORF) mediated regulation across eight species. I first establish that uORFs play a repressive role on the translational efficiency of the downstream protein coding regions. I characterized uuORFs - universal uORFs that are 'almost always' translating within a species given the diverse physiological and pathological contexts of the public datasets that were used to characterize them. Further more, I characterize that the upstream sequence context of these uORFs is similar to the upstream sequence context of the coding domain sequences within the species. The uuORFs are conserved across species with total conservation inversely proportional to the divergence time. The sequence context is conserved within the vertebrates and invertebrates.

The characterization of uuORFs is the first step towards deciphering their functional role. The short polypetides synthesized from these uuORFs could also be potential regulators of house-keeping operations in the regulatory system. Thus, one focus of future studies should be on characterizing their functional role, given that they are almost always present.

Bibliography

[1] Y. Zarai, M. Margaliot, and T. Tuller, "Maximizing protein translation rate in the ribosome flow model: the homogeneous case," *arxiv:1407.0207v1*, 2014.

[2] F. Supek, M. Bošnjak, N. Škunca, and T. Šmuc, "REVIGO summarizes and visualizes long lists of gene ontology terms," *PLoS ONE*, vol. 6, no. 7, p. e21800, 2011.

[3] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, *et al.*, "STRING v10: protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. D1, pp. D447–D452, 2014.

[4] R. L. Bowman, Q. Wang, A. Carro, R. G. Verhaak, and M. Squatrito, "GlioVis data portal for visualization and analysis of brain tumor expression datasets," *Neuro-Oncology*, vol. 19, no. 1, pp. 139–141, 2016.

[5] R. Dahm, "Discovering DNA: Friedrich Miescher and the early years of nucleic acid research," *Human genetics*, vol. 122, no. 6, pp. 565–581, 2008.

[6] H. B. Vickery, "The origin of the word Protein," *The Yale journal of biology and medicine*, vol. 22, no. 5, p. 387, 1950.

[7] C. Tanford and J. Reynolds, *Nature's Robots: A History of Proteins* . OUP Oxford, 2003.

[8] J. Brock and S. Davidson, *Human Nutrition and Dietetics*. 1973.

[9] F. H. Crick, "On protein synthesis," in *Symp Soc Exp Biol*, vol. 12, p. 8, 1958.

[10] M. B. Hoagland, M. L. Stephenson, J. F. Scott, L. I. Hecht, and P. C. Zamecnik, "A soluble ribonucleic acid intermediate in protein synthesis," *Journal of Biological Chemistry*, vol. 231, no. 1, pp. 241–257, 1958.

[11] R. Kohler, "The Eighth Day of Creation: Makers of the Revolution in Biology by Horace Freeland Judson," 1997.

[12] T. Nakatani, T. Chen, J. Johnson, J. J. Westendorf, and N. C. Partridge, "The deletion of hdac4 in mouse osteoblasts influences both catabolic and anabolic effects in bone," *Journal of Bone and Mineral Research*, vol. 33, pp. 1362–1375, apr 2018.

[13] G. Meisenberg and W. H. Simmons, *Principles of Medical Biochemistry*. Elsevier Health Sciences, 2016.

[14] J. R. Warner, "The economics of ribosome biosynthesis in yeast," *Trends in Biochemical Sciences*, vol. 24, pp. 437–440, Nov. 1999.

[15] J. W. Hershey, N. Sonenberg, and M. B. Mathews, "Principles of translational control: an overview," *Cold Spring Harbor Perspectives In Biology*, vol. 4, no. 12, p. a011528, 2012.

[16] G. C. Scheper, M. S. van der Knaap, and C. G. Proud, "Translation matters: protein synthesis defects in inherited disease," *Nature Reviews Genetics*, vol. 8, pp. 711–723, jul 2007.

[17] F. Jacob and J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins," p. 39, 1961.

[18] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature methods*, vol. 5, no. 7, p. 621, 2008.

[19] C. Vogel and E. M. Marcotte, "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses," *Nature Reviews Genetics*, vol. 13, pp. 227–232, Mar. 2012.

[20] Y. Liu, A. Beyer, and R. Aebersold, "On the Dependency of Cellular Protein Levels on mRNA Abundance," *Cell*, vol. 165, no. 3, pp. 535–550, 2016.

[21] F. Edfors, F. Danielsson, B. M. Hallström, L. Käll, E. Lundberg, F. Pontén, B. Forsström, and M. Uhlén, "Gene-specific correlation of RNA and protein levels in human cells and tissues," *Molecular systems biology*, vol. 12, no. 10, 2016.

[22] T. Maier, M. Güell, and L. Serrano, "Correlation of mRNA and protein in complex biological samples," *FEBS Letters*, vol. 583, pp. 3966–3973, Oct. 2009.

[23] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature Methods*, vol. 5, no. 1, pp. 16–18, 2008.

[24] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman, "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling," *science*, vol. 324, no. 5924, pp. 218–223, 2009.

[25] C. Gérard and A. Goldbeter, "Dynamics of the mammalian cell cycle in physiological and pathological conditions," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 8, pp. 140–156, nov 2015.

[26] S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, and S.-B. Qian, "Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution," *Proceedings of the National Academy of Sciences*, vol. 109, pp. E2424–E2432, aug 2012.

[27] C. Fritsch, A. Herrmann, M. Nothnagel, K. Szafranski, K. Huse, F. Schumann, S. Schreiber, M. Platzer, M. Krawczak, J. Hampe, and M. Brosch, "Genome-wide search for novel human uORFs and n-terminal protein extensions using ribosomal footprinting," *Genome Research*, vol. 22, pp. 2208–2218, aug 2012.

[28] R. Shalgi, J. A. Hurt, I. Krykbaeva, M. Taipale, S. Lindquist, and C. B. Burge, "Widespread regulation of translation by elongation pausing in heat shock," *Molecular Cell*, vol. 49, pp. 439–452, feb 2013.

[29] G.-L. Chew, A. Pauli, J. L. Rinn, A. Regev, A. F. Schier, and E. Valen, "Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs," *Development*, vol. 140, pp. 2828–2834, may 2013.

[30] M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, and E. S. Lander, "Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins," *Cell*, vol. 154, pp. 240–251, jul 2013.

[31] N. T. Ingolia, G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. Talhouarne, S. E. Jackson, M. R. Wills, and J. S. Weissman, "Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes," *Cell Reports*, vol. 8, pp. 1365–1379, sep 2014.

[32] A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, and A. J. Giraldez, "Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation.," *EMBO J.*, vol. 33, pp. 981–93, May 2014.

[33] A. P. Fields, E. H. Rodriguez, M. Jovanovic, N. Stern-Ginossar, B. J. Haas, P. Mertins, R. Raychowdhury, N. Hacohen, S. A. Carr, N. T. Ingolia, A. Regev, and J. S. Weissman, "A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation," *Molecular Cell*, vol. 60, pp. 816–827, dec 2015.

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

[35] L. Calviello, N. Mukherjee, E. Wyler, H. Zauber, A. Hirsekorn, M. Selbach, M. Landthaler, B. Obermayer, and U. Ohler, "Detecting actively translated open reading frames in ribosome profiling data," *Nature Methods*, vol. 13, pp. 165–170, dec 2015.

[36] B. Malone, I. Atanassov, F. Aeschimann, X. Li, H. Großhans, and C. Dieterich, "Bayesian prediction of RNA translation from ribosome profiling," *Nucleic Acids Research*, vol. 45, no. 6, pp. 2960–2972, 2017.

[37] C. Barbosa, I. Peixeiro, and L. Romão, "Gene expression regulation by upstream open reading frames and human disease," *PLoS Genetics*, vol. 9, p. e1003529, Aug. 2013.

[38] T. Dobzhansky, “Nothing in Biology Makes Sense except in the Light of Evolution,” *The American Biology Teacher*, vol. 35, pp. 125–129, Mar. 1973.

[39] G.-W. Li, D. Burkhardt, C. Gross, and J. S. Weissman, “Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources,” *Cell*, vol. 157, no. 3, pp. 624–635, 2014.

[40] H. Tani, R. Mizutani, K. A. Salam, K. Tano, K. Ijiri, A. Wakamatsu, T. Isogai, Y. Suzuki, and N. Akimitsu, “Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals,” *Genome Research*, vol. 22, no. 5, pp. 947–956, 2012.

[41] L. Y. Chan, C. F. Mugler, S. Heinrich, P. Vallotton, and K. Weis, “Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability,” *eLife*, vol. 7, p. e32536, 2018.

[42] G. A. Brar, “Beyond the triplet code: context cues transform translation,” *Cell*, vol. 167, no. 7, pp. 1681–1692, 2016.

[43] A. S. Zhokhin and V. P. Gachok, “Transition to chaos in the kinetic model of cellulose hydrolysis under enzyme biosynthesis control,” *arxiv:1707.08914v1*, 2017.

[44] L. Ribas de Pouplana, M. A. Santos, J.-H. Zhu, P. J. Farabaugh, and B. Javid, “Protein mistranslation: Friend or foe?,” *Trends in Biochemical Sciences*, vol. 39, pp. 355–362, Aug 2014.

[45] C. Zhang, S. Zhou, E. Groppelli, P. Pellegrino, I. Williams, P. Borrow, B. M. Chain, and C. Jolly, "Hybrid spreading mechanisms and t cell activation shape the dynamics of HIV-1 infection," *PLOS Computational Biology*, vol. 11, p. e1004179, apr 2015.

[46] R. J. Jackson, C. U. Hellen, and T. V. Pestova, "The mechanism of eukaryotic translation initiation and principles of its regulation," *Nature Reviews Molecular Cell Biology*, vol. 11, no. 2, pp. 113–127, 2010.

[47] M. Kozak, "An analysis of vertebrate mRNA sequences: intimations of translational control," *The Journal of Cell Biology*, vol. 115, pp. 887–903, nov 1991.

[48] J. KIEFT, A. GRECH, P. ADAMS, and J. DOUDNA, "Mechanisms of internal ribosome entry in translation initiation," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 66, pp. 277–284, jan 2001.

[49] P. Shah, Y. Ding, M. Niemczyk, G. Kudla, and J. B. Plotkin, "Rate-limiting steps in yeast protein translation," *Cell*, vol. 153, no. 7, pp. 1589–1601, 2013.

[50] M. V. Rodnina, "Translation in prokaryotes," *Cold Spring Harbor Perspectives in Biology*, vol. 10, no. 9, p. a032664, 2018.

[51] P. M. Gentz, G. L. Blatch, and R. A. Dorrington, "Dimerization of the yeast eukaryotic translation initiation factor 5a requires hypusine and is RNA dependent," *FEBS Journal*, vol. 276, pp. 695–706, dec 2008.

[52] Y. Furuichi, "Discovery of m^{7}G-cap in eukaryotic mRNAs," *Proceedings of the Japan Academy, Series B*, vol. 91, no. 8, pp. 394–409, 2015. 00000.

[53] M. Kozak, "Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes," *Cell*, vol. 44, pp. 283–292, jan 1986.

[54] D. R. Cavener and S. C. Ray, "Eukaryotic start and stop translation sites," *Nucleic Acids Research*, vol. 19, no. 12, pp. 3185–3192, 1991.

[55] M. Kozak, "The scanning model for translation: an update.," *The Journal of Cell Biology*, vol. 108, pp. 229–241, Feb. 1989.

[56] G. Hernández, V. G. Osnaya, and X. Pérez-Martínez, "Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes," *Trends in Biochemical Sciences*, vol. 44, pp. 1009–1021, Dec. 2019.

[57] M. Jalsenius and K. Pedersen, "A systematic scan for 7-colourings of the grid," *arxiv:0704.1625v3*, 2007.

[58] J. McManus, G. May, P. Spealman, and A. Shteyman, "Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast," *arxiv:1312.1765v1*, 2013.

[59] S. L. Wolin and P. Walter, "Ribosome pausing and stacking during translation of a eukaryotic mRNA.," *The EMBO Journal*, vol. 7, pp. 3559–3569, nov 1988.

[60] C. J. Woolstenhulme, N. R. Guydosh, R. Green, and A. R. Buskirk, "High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP," *Cell Reports*, vol. 11, pp. 13–21, apr 2015.

[61] A. V. Pisarev, V. G. Kolupaeva, M. M. Yusupov, C. U. Hellen, and T. V. Pestova, "Ribosomal position and contacts of mRNA in eukaryotic translation initiation complexes," *The EMBO Journal*, vol. 27, pp. 1609–1621, Jun 2008. 00171.

[62] Y. Arava, Y. Wang, J. D. Storey, C. L. Liu, P. O. Brown, and D. Herschlag, "Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 3889–3894, Apr. 2003. 00657.

[63] P. Sampath, D. K. Pritchard, L. Pabon, H. Reinecke, S. M. Schwartz, D. R. Morris, and C. E. Murry, "A Hierarchical Network Controls Protein Translation during Murine Embryonic Stem Cell Self-Renewal and Differentiation," *Cell Stem Cell*, vol. 2, pp. 448–460, May 2008. 00216.

[64] C. G. Artieri and H. B. Fraser, "Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation," *Genome Research*, vol. 24, pp. 2011–2021, oct 2014.

[65] J. Li and Y. Zhang, "Translation with frameshifting of ribosome along mrna transcript," *arxiv:1502.02109v1*, 2015.

[66] J. A. Hussmann, S. Patchett, A. Johnson, S. Sawyer, and W. H. Press, "Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast," *PLOS Genetics*, vol. 11, p. e1005732, dec 2015.

[67] D. A. Santos, L. Shi, B. P. Tu, and J. S. Weissman, "Cycloheximide can distort measurements of mRNA levels and translation efficiency," *Nucleic Acids Research*, vol. 47, pp. 4974–4985, mar 2019.

[68] J. Li and Y. Zhang, “Translation with frameshifting of ribosome along mrna transcript,” *arxiv:1502.02109v1*, 2015.

[69] F. Mohammad, R. Green, and A. R. Buskirk, “A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution,” *eLife*, vol. 8, p. e42591, 2019.

[70] M. Chevallet, P. Lescuyer, H. Diemer, A. van Dorsselaer, E. Leize-Wagner, and T. Rabilloud, “Alterations of the mitochondrial proteome caused by the absence of mitochondrial DNA: A proteomic view,” *ELECTROPHORESIS*, vol. 27, pp. 1574–1583, apr 2006.

[71] L. V. Bock, M. H. Kolář, and H. Grubmüller, “Molecular simulations of the ribosome and associated translation factors,” *arxiv:1711.06067v1*, 2017.

[72] N. T. Ingolia, L. F. Lareau, and J. S. Weissman, “Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes,” *Cell*, vol. 147, no. 4, pp. 789–802, 2011.

[73] P. Zhang, D. He, Y. Xu, J. Hou, B.-F. Pan, Y. Wang, T. Liu, C. M. Davis, E. A. Ehli, L. Tan, *et al.*, “Genome-wide identification and differential analysis of translational initiation,” *Nature communications*, vol. 8, no. 1, pp. 1–14, 2017.

[74] N. T. Ingolia, L. F. Lareau, and J. S. Weissman, “Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes,” *Cell*, vol. 147, pp. 789–802, nov 2011.

[75] V. Kasari, T. Margus, G. C. Atkinson, M. J. Johansson, and V. Hauryliuk, "Ribosome profiling analysis of eef3-depleted saccharomyces cerevisiae," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[76] Y. Han, X. Gao, B. Liu, J. Wan, X. Zhang, and S.-B. Qian, "Ribosome profiling reveals sequence-independent post-initiation pausing as a signature of translation," *Cell research*, vol. 24, no. 7, pp. 842–851, 2014.

[77] R. Jackson, L. Kroehling, A. Khitun, W. Bailis, A. Jarret, A. G. York, O. M. Khan, J. R. Brewer, M. H. Skadow, C. Duizer, C. C. D. Harman, L. Chang, P. Bielecki, A. G. Solis, H. R. Steach, S. Slavoff, and R. A. Flavell, "The translation of non-canonical open reading frames controls mucosal immunity," *Nature*, vol. 564, pp. 434–438, dec 2018.

[78] V. Olexiouk, W. V. Criekinge, and G. Menschaert, "An update on sORFs.org: a repository of small ORFs identified by ribosome profiling," *Nucleic Acids Research*, vol. 46, pp. D497–D502, nov 2017.

[79] S. Bianchini, A. Lage-castellanos, and E. Altshuler, "Upstream contamination in water pouring," *arxiv:1105.2585v1*, 2011.

[80] Q. Li, J. K. Eng, and M. Stephens, "A likelihood-based scoring method for peptide identification using mass spectrometry," *The Annals of Applied Statistics*, vol. 6, pp. 1775–1794, dec 2012.

[81] J. Radianti, J. Dugdale, J. J. Gonzalez, and O.-C. Granmo, "Smartphone sensing platform for emergency management," *arxiv:1406.3848v1*, 2014.

[82] C. Barbosa, I. Peixeiro, and L. Romão, "Gene expression regulation by upstream open reading frames and human disease," *PLoS Genetics*, vol. 9, p. e1003529, aug 2013.

[83] J. Li and Y. Zhang, "Translation with frameshifting of ribosome along mrna transcript," *arxiv:1502.02109v1*, 2015.

[84] D. E. Andreev, P. B. O'Connor, C. Fahey, E. M. Kenny, I. M. Terenin, S. E. Dmitriev, P. Cormican, D. W. Morris, I. N. Shatsky, and P. V. Baranov, "Translation of 5' leaders is pervasive in genes resistant to eIF2 repression," *Elife*, vol. 4, p. e03971, 2015.

[85] A. A. Bazzini, T. G. Johnstone, R. Christiano, S. D. Mackowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, and A. J. Giraldez, "Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation," *The EMBO Journal*, vol. 33, pp. 981–993, apr 2014.

[86] A. Raj, S. H. Wang, H. Shim, A. Harpak, Y. I. Li, B. Engelmann, M. Stephens, Y. Gilad, and J. K. Pritchard, "Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling," *eLife*, vol. 5, may 2016.

[87] Z. Xiao, R. Huang, X. Xing, Y. Chen, H. Deng, and X. Yang, "De novo annotation and characterization of the translatome with ribosome profiling data," *Nucleic Acids Research*, vol. 46, pp. e61–e61, mar 2018.

[88] C. Cai, R. Chen, and M. ge Xie, "Individualized group learning," *arxiv:1906.05533v1*, 2019.

[89] S. Y. Chun, C. M. Rodriguez, P. K. Todd, and R. E. Mills, "SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data," *BMC Bioinformatics*, vol. 17, nov 2016.

[90] C. A. Brackley, D. Broomhead, M. C. Romano, and M. Thiel, "A max-plus model of ribosome dynamics during mrna translation," *arxiv:1105.3580v1*, 2011.

[91] J. A. Hussmann, S. Patchett, A. Johnson, S. Sawyer, and W. H. Press, "Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast," *PLoS Genetics*, vol. 11, p. e1005732, Dec. 2015. 00050.

[92] P. B. O'Connor, D. E. Andreev, and P. V. Baranov, "Comparative survey of the relative impact of mRNA features on local ribosome profiling read density," *Nature Communications*, vol. 7, p. 12915, 2016.

[93] S. Choudhary, "pysradb: A python package to query next-generation sequencing metadata and data from NCBI sequence read archive," mar 2019.

[94] S. Ganguly, E. Mossel, and M. Z. Racz, "Sequence assembly from corrupted shotgun reads," *arxiv:1601.07086v1*, 2016.

[95] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, pp. 15–21, oct 2012.

[96] H. H. von Grünberg, M. Peifer, J. Timmer, and M. Kollmann, "Variations in substitution rate in human and mouse genomes," *Physical Review Letters*, vol. 93, nov 2004.

[97] M. Ruffier, A. Kähäri, M. Komorowska, S. Keenan, M. Laird, I. Longden, G. Proctor, S. Searle, D. Staines, K. Taylor, *et al.*, "Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation," *Database*, vol. 2017, 2017.

[98] M. S. Skrzypek, J. Binkley, G. Binkley, S. R. Miyasato, M. Simison, and G. Sherlock, "The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data," *Nucleic Acids Research*, p. gkw924, 2016.

[99] J. R. V. Zandt, "Efficient cubature rules," *arxiv:1712.07309v3*, 2017.

[100] S. A. Eales, M. Baes, N. Bourne, M. Bremer, M. J. I. Brown, C. Clark, D. Clements, P. de Vis, S. Driver, L. Dunne, S. Dye, C. Furlanetto, B. Holwerda, R. J. Ivison, L. S. Kelvin, M. Lara-Lopez, L. Leeuw, J. Loveday, S. Maddox, M. J. Michałowski, S. Phillipps, A. Robotham, D. Smith, M. Smith, E. Valiante, P. van der Werf, and A. Wright, "The causes of the red sequence, the blue cloud, the green valley, and the green mountain," *Monthly Notices of the Royal Astronomical Society*, vol. 481, pp. 1183–1194, aug 2018.

[101] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M.-M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. DiCuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman, "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes," *Genome Research*, vol. 19, pp. 1316–1323, jun 2009.

[102] D. Hinkley, "Bootstrap methods: Another look at the jackknife," in *Springer Series in Statistics*, pp. 178–206, Springer New York.

[103] T. T. Ivancevic, M. J. Bottema, and L. C. Jain, "A theoretical model of chaotic attractor in tumor growth and metastasis," *arxiv:0807.4272v1*, 2008.

[104] J. Noorbakhsh, A. H. Lang, and P. Mehta, "Intrinsic noise of microRNA-regulated genes and the ceRNA hypothesis," *PLoS ONE*, vol. 8, p. e72676, aug 2013.

[105] M. Mariotti, S. Shetty, L. Baird, S. Wu, G. Loughran, P. R. Copeland, J. F. Atkins, and M. T. Howard, "Multiple RNA structures affect translation initiation and UGA redefinition efficiency during synthesis of selenoprotein p," *Nucleic Acids Research*, vol. 45, pp. 13004–13015, oct 2017.

[106] M. Castellana, S. H.-J. Li, and N. S. Wingreen, "Spatial organization of bacterial transcription and translation," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 9286–9291, aug 2016.

[107] Y. Chua and C. Tan, "Neurogenesis and multiple plasticity mechanisms enhance associative memory retrieval in a spiking network model of the hippocampus," *arxiv:1704.07526v1*, 2017.

[108] S. Sengupta, X. Yang, and P. G. Higgs, "The mechanisms of codon reassignments in mitochondrial genetic codes," *arxiv:q-bio/0703066v1*, 2007.

[109] Y. T. Lin, P. G. Hufton, E. J. Lee, and D. A. Potoyan, "A stochastic and dynamical view of pluripotency in mouse embryonic stem cells," *PLOS Computational Biology*, vol. 14, p. e1006000, feb 2018.

[110] G. Xu, G. H. Greene, H. Yoo, L. Liu, J. Marqués, J. Motley, and X. Dong, "Global translational reprogramming is a fundamental layer of immune regulation in plants.," *Nature*, vol. 545, pp. 487–490, 05 2017.

[111] P. Juntawong, T. Girke, J. Bazin, and J. Bailey-Serres, "Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 111, pp. E203–12, Jan 2014.

[112] R. Lukoszek, P. Feist, and Z. Ignatova, "Insights into the adaptive response of Arabidopsis thaliana to prolonged thermal stress by ribosomal profiling and RNA-Seq.," *BMC Plant Biol.*, vol. 16, p. 221, 10 2016.

[113] M.-J. Liu, S.-H. Wu, J.-F. Wu, W.-D. Lin, Y.-C. Wu, T.-Y. Tsai, H.-L. Tsai, and S.-H. Wu, "Translational landscape of photomorphogenic Arabidopsis.," *Plant Cell*, vol. 25, pp. 3699–710, Oct 2013.

[114] H. M. Blank, R. Perez, C. He, N. Maitra, R. Metz, J. Hill, Y. Lin, C. D. Johnson, V. A. Bankaitis, B. K. Kennedy, R. Aramayo, and M. Polymenis, "Translational control of lipogenic enzymes in the cell cycle of synchronous, growing yeast cells.," *EMBO J.*, vol. 36, pp. 487–502, 02 2017.

[115] N. R. Guydosh and R. Green, "Dom34 rescues ribosomes in 3' untranslated regions.," *Cell*, vol. 156, pp. 950–62, Feb 2014.

[116] C. G. Artieri and H. B. Fraser, "Evolution at two levels of gene expression in yeast.," *Genome Res.*, vol. 24, pp. 411–21, Mar 2014.

[117] C. J. McManus, G. E. May, P. Spealman, and A. Shteyman, "Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast.," *Genome Res.*, vol. 24, pp. 422–30, Mar 2014.

[118] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling.," *Science*, vol. 324, pp. 218–23, Apr 2009.

[119] D. D. Nedialkova and S. A. Leidel, "Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity.," *Cell*, vol. 161, pp. 1606–18, Jun 2015.

[120] M. Stadler and A. Fire, "Conserved translatome remodeling in nematode species executing a shared developmental transition.," *PLoS Genetics*, vol. 9, no. 10, p. e1003739, 2013.

[121] M. Stadler, K. Artiles, J. Pak, and A. Fire, "Contributions of mRNA abundance, ribosome loading, and post- or peri-translational effects to temporal repression of C. elegans heterochronic miRNA targets.," *Genome Res.*, vol. 22, pp. 2418–26, Dec 2012.

[122] M. Stadler and A. Fire, "Wobble base-pairing slows in vivo translation elongation in metazoans.," *RNA*, vol. 17, pp. 2063–73, Dec 2011.

[123] X. Chen and D. Dickman, "Development of a tissue-specific ribosome profiling approach in Drosophila enables genome-wide evaluation of translational adaptations.," *PLoS Genetics*, vol. 13, p. e1007117, Dec 2017.

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

[125] S. W. Eichhorn, A. O. Subtelny, I. Kronja, J. C. Kwasnieski, T. L. Orr-Weaver, and D. P. Bartel, “mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in Drosophila oocytes and early embryos.,” *Elife*, vol. 5, 07 2016.

[126] J. L. Aspden, Y. C. Eyre-Walker, R. J. Phillips, U. Amin, M. A. S. Mumtaz, M. Brocard, and J.-P. Couso, “Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq.,” *Elife*, vol. 3, p. e03528, Aug 2014.

[127] D. E. Andreev, P. B. F. O’Connor, A. V. Zhdanov, R. I. Dmitriev, I. N. Shatsky, D. B. Papkovsky, and P. V. Baranov, “Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes.,” *Genome Biol.*, vol. 16, p. 90, May 2015.

[128] A. Ori, B. H. Toyama, M. S. Harris, T. Bock, M. Iskar, P. Bork, N. T. Ingolia, M. W. Hetzer, and M. Beck, “Integrated Transcriptome and Proteome Analyses Reveal Organ-Specific Proteome Deterioration in Old Rats.,” *Cell Syst*, vol. 1, pp. 224–37, Sep 2015.

[129] S. Schafer, E. Adami, M. Heinig, K. E. C. Rodrigues, F. Kreuchwig, J. Silhavy, S. van Heesch, D. Simaite, N. Rajewsky, E. Cuppen, M. Pravenec, M. Vingron, S. A. Cook, and N. Hubner, “Translational regulation shapes the molecular landscape of complex disease phenotypes.,” *Nat Commun*, vol. 6, p. 7200, May 2015.

[130] A. A. Bazzini, M. T. Lee, and A. J. Giraldez, “Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish.,” *Science*, vol. 336, pp. 233–7, Apr 2012.

[131] M. T. Lee, A. R. Bonneau, C. M. Takacs, A. A. Bazzini, K. R. DiVito, E. S. Fleming, and A. J. Giraldez, “Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition.,” *Nature*, vol. 503, pp. 360–4, Nov 2013.

[132] D. Muzzey, G. Sherlock, and J. S. Weissman, “Extensive and coordinated control of allele-specific expression by both transcription and translation in Candida albicans.,” *Genome Res.*, vol. 24, pp. 963–73, Jun 2014.

[133] N. R. Guydosh, P. Kimmig, P. Walter, and R. Green, “Regulated Ire1-dependent mRNA decay requires no-go mRNA degradation to maintain endoplasmic reticulum homeostasis in S. pombe.,” *Elife*, vol. 6, 09 2017.

[134] S. H. Wang, C. J. Hsiao, Z. Khan, and J. K. Pritchard, “Post-translational buffering leads to convergent protein expression levels between primates.,” *Genome Biol.*, vol. 19, p. 83, 06 2018.

[135] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC bioinformatics*, vol. 12, no. 1, p. 77, 2011.

[136] A. M. Michel, S. J. Kiniry, P. B. F. O’Connor, J. P. Mullan, and P. V. Baranov, “GWIPS-viz: 2018 update,” *Nucleic acids research*, vol. 46, no. D1, pp. D823–D830, 2017.

[137] H. Wang, L. Yang, Y. Wang, L. Chen, H. Li, and Z. Xie, “Rpfdb v2. 0: an updated database for genome-wide information of translated mrna generated from ribosome profiling,” *Nucleic acids research*, vol. 47, no. D1, pp. D230–D234, 2018.

[138] M. A. Pfaller and D. J. Diekema, "Epidemiology of Invasive Candidiasis: a Persistent Public Health Problem," *Clinical Microbiology Reviews*, vol. 20, pp. 133–163, Jan. 2007.

[139] M. A. Pfaller, D. R. Andes, D. J. Diekema, D. L. Horn, A. C. Reboli, C. Rotstein, B. Franks, and N. E. Azie, "Epidemiology and Outcomes of Invasive Candidiasis Due to Non-albicans Species of Candida in 2,496 Patients: Data from the Prospective Antifungal Therap y (PATH) Registry 2004–2008," *PLoS ONE*, vol. 9, p. e101510, July 2014.

[140] M. E. Falagas, N. Roussos, and K. Z. Vardakas, "Relative frequency of albicans and the various non-albicans Candida spp among candidemia isolates from inpatients in various parts of the world: a systematic review," *International Journal of Infectious Diseases*, vol. 14, no. 11, pp. e954–e966, 2010.

[141] T. H. Koornwinder, "q-special functions, an overview," *arxiv:math/0511148v1*, 2005.

[142] A. Čufar, A. Mrhar, and M. Robnik-Šikonja, "Identifying roles of clinical pharmacy with survey evaluation," *arxiv:1406.4287v1*, 2014.

[143] R. Hirano, Y. Sakamoto, K. Kudo, and M. Ohnishi, "Retrospective analysis of mortality and Candida isolates of 75 patients with candidemia: a single hospital experience," *Infection and drug resistance*, vol. 8, p. 199, 2015.

[144] Z. Xiao, Q. Wang, F. Zhu, and Y. An, "Epidemiology, species distribution, antifungal susceptibility and mortality risk factors of candidemia among critically ill patients: a retrospective study from 2011 to 2017 in a teaching hospital in China," *Antimicrobial Resistance & Infection Control*, vol. 8, no. 1, p. 89, 2019.

[145] A. Boullis, F. Francis, and F. Verheggen, "Aphid-hoverfly interactions under elevated CO2 concentrations: oviposition and larval development," *Physiological Entomology*, vol. 43, pp. 245–250, jun 2018.

[146] D. Kadosh, "Control of Candida albicans morphology and pathogenicity by post-transcriptional mechanisms," *Cellular and Molecular Life Sciences*, vol. 73, pp. 4265–4278, Nov. 2016.

[147] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biology*, vol. 15, no. 12, p. 550, 2014.

[148] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, "Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling," *Science*, vol. 324, pp. 218–223, apr 2009.

[149] S. Choudhary, W. Li, and A. D. Smith, "Accurate detection of short and long active ORFs using Ribo-seq data," *Bioinformatics*, 2019.

[150] W. Li, W. Wang, P. J. Uren, L. O. Penalva, and A. D. Smith, "Riborex: fast and flexible identification of differential translation from Ribo-seq data," *Bioinformatics*, vol. 33, no. 11, pp. 1735–1737, 2017.

[151] V. M. Bruno, Z. Wang, S. L. Marjani, G. M. Euskirchen, J. Martin, G. Sherlock, and M. Snyder, "Comprehensive annotation of the transcriptome of the human fungal pathogen candida albicans using RNA-seq," *Genome Research*, vol. 20, pp. 1451–1458, sep 2010.

[152] A. S. Hinrichs, D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson, M. Diekhans, T. S. Furey, R. A. Harte, F. Hsu, *et al.*, "The UCSC genome browser

database: update 2006," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D590–D598, 2006.

[153] M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, and S. L. Salzberg, "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads," *Nature Biotechnology*, vol. 33, no. 3, p. 290, 2015.

[154] I. H. Madshus, T. I. Tønnessen, S. Olsnes, and K. Sandvig, "Effect of potassium depletion of hep 2 cells on intracellular pH and on chloride uptake by anion antiport," *Journal of Cellular Physiology*, vol. 131, pp. 6–13, apr 1987.

[155] D. K. Okamoto, "Competition among eggs shifts to cooperation along a sperm supply gradient in an external fertilizer," *arxiv:1510.08813v3*, 2015.

[156] H. Teppola, S. Okujeni, M. L. Linne, and U. Egert, "Ampa, nmda and gabaa receptor mediated network burst dynamics in cortical cultures in vitro," *arxiv:1802.00217v1*, 2018.

[157] J. C. Darnell, S. J. V. Driesche, C. Zhang, K. Y. S. Hung, A. Mele, C. E. Fraser, E. F. Stone, C. Chen, J. J. Fak, S. W. Chi, D. D. Licatalosi, J. D. Richter, and R. B. Darnell, "FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism," *Cell*, vol. 146, pp. 247–261, jul 2011.

[158] W. J. Faller, T. J. Jackson, J. R. P. Knight, R. A. Ridgway, T. Jamieson, S. A. Karim, C. Jones, S. Radulescu, D. J. Huels, K. B. Myant, K. M. Dudek, H. A. Casey, A. Scopelliti, J. B. Cordero, M. Vidal, M. Pende, A. G. Ryazanov, N. Sonenberg, O. Meyuhas, M. N. Hall, M. Bushell, A. E. Willis, and O. J. Sansom, "mTORC1-mediated translational elongation limits intestinal tumour initiation and growth," *Nature*, vol. 517, pp. 497–500, nov 2014.

[159] S. Andrews *et al.*, "Fastqc: a quality control tool for high throughput sequence data," 2010.

[160] F. Krueger, "TrimGalore! A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files," 2012.

[161] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He, "clusterprofiler: an r package for comparing biological themes among gene clusters," *Omics: a journal of integrative biology*, vol. 16, no. 5, pp. 284–287, 2012.

[162] S. Zhang, H. Hu, J. Zhou, X. He, T. Jiang, and J. Zeng, "Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning," *Cell Systems*, vol. 5, pp. 212–220.e6, Sep 2017. 00006.

[163] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures.," *Analytical Chemistry*, vol. 36, pp. 1627–1639, jul 1964.

[164] E. García-Portugués, R. M. Crujeiras, and W. González-Manteiga, "Smoothing-based tests with directional random variables," *arxiv:1804.00230v1*, 2018.

[165] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[166] R. Stupp, W. P. Mason, M. J. Van Den Bent, M. Weller, B. Fisher, M. J. Taphoorn, K. Belanger, A. A. Brandes, C. Marosi, U. Bogdahn, *et al.*, "Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma," *New England Journal of Medicine*, vol. 352, no. 10, pp. 987–996, 2005.

[167] M. R. Gilbert, J. J. Dignam, T. S. Armstrong, J. S. Wefel, D. T. Blumenthal, M. A. Vogelbaum, H. Colman, A. Chakravarti, S. Pugh, M. Won, *et al.*, “A Randomized Trial of Bevacizumab for Newly Diagnosed Glioblastoma,” *New England Journal of Medicine*, vol. 370, no. 8, pp. 699–708, 2014.

[168] M. E. Hegi, A.-C. Diserens, T. Gorlia, M.-F. Hamou, N. De Tribolet, M. Weller, J. M. Kros, J. A. Hainfellner, W. Mason, L. Mariani, *et al.*, “MGMT gene silencing and benefit from temozolomide in glioblastoma,” *New England Journal of Medicine*, vol. 352, no. 10, pp. 997–1003, 2005.

[169] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, J. M. Stuart, C. G. A. R. Network, *et al.*, “The Cancer Genome Atlas Pan-Cancer analysis project,” *Nature Genetics*, vol. 45, no. 10, p. 1113, 2013.

[170] Y. Lee, K. H. Kim, D. G. Kim, H. J. Cho, Y. Kim, J. Rheey, K. Shin, Y. J. Seo, Y.-S. Choi, J.-I. Lee, *et al.*, “FoxM1 Promotes Stemness and Radio-Resistance of Glioblastoma by Regulating the Master Stem Cell Regulator Sox2,” *PLoS ONE*, vol. 10, no. 10, p. e0137703, 2015.

[171] U. B. Maachani, U. Shankavaram, T. Kramp, P. J. Tofilon, K. Camphausen, and A. T. Tandle, “FOXM1 and STAT3 interaction confers radioresistance in glioblastoma cells,” *Oncotarget*, vol. 7, no. 47, p. 77365, 2016.

[172] L. Cheng, Q. Wu, Z. Huang, O. A. Guryanova, Q. Huang, W. Shou, J. N. Rich, and S. Bao, “L1CAM regulates DNA damage checkpoint response of glioblastoma stem cells through NBS1,” *The EMBO Journal*, vol. 30, no. 5, pp. 800–813, 2011.

[173] X. Han, P. Ranganathan, C. Tzimas, K. L. Weaver, K. Jin, L. Astudillo, W. Zhou, X. Zhu, B. Li, D. J. Robbins, *et al.*, “Notch Represses Transcription by PRC2 Recruitment to the Ternary Complex,” *Molecular Cancer Research*, vol. 15, no. 9, pp. 1173–1183, 2017.

[174] A. Balbous, U. Cortes, K. Guilloteau, P. Rivet, B. Pinel, M. Duchesne, J. Godet, O. Boissonnade, M. Wager, R. J. Bensadoun, *et al.*, “A radiosensitizing effect of rad51 inhibition in glioblastoma stem-like cells,” *BMC Cancer*, vol. 16, no. 1, p. 604, 2016.

[175] S.-H. Kim, K. Joshi, R. Ezhilarasan, T. R. Myers, J. Siu, C. Gu, M. Nakano-Okuno, D. Taylor, M. Minata, E. P. Sulman, *et al.*, “EZH2 protects glioma stem cells from radiation-induced cell death in a MELK/FOXM1-dependent manner,” *Stem Cell Reports*, vol. 4, no. 2, pp. 226–238, 2015.

[176] S. U. Ahmed, R. Carruthers, L. Gilmour, S. Yildirim, C. Watts, and A. J. Chalmers, “Selective inhibition of parallel dna damage response pathways optimizes radiosensitization of glioblastoma stem-like cells,” *Cancer Research*, vol. 75, no. 20, pp. 4416–4428, 2015.

[177] A. Karim, K. McCarthy, A. Jawahar, D. Smith, B. Willis, and A. Nanda, “Differential cyclooxygenase-2 enzyme expression in radiosensitive versus radioresistant glioblastoma multiforme cell lines,” *Anticancer Research*, vol. 25, no. 1B, pp. 675–679, 2005.

[178] S. H. K. Vellanki, A. Grabrucker, S. Liebau, C. Proepper, A. Eramo, V. Braun, T. Boeckers, K.-M. Debatin, and S. Fulda, “Small-molecule xiap inhibitors enhance γ-irradiation-induced apoptosis in glioblastoma,” *Neoplasia*, vol. 11, no. 8, pp. 743–W9, 2009.

[179] H. Ma, L. Rao, H. Wang, Z. Mao, R. Lei, Z. Yang, H. Qing, and Y. Deng, “Transcriptome analysis of glioma cells for the dynamic response to γ-irradiation and dual regulation of

apoptosis genes: a new insight into radiotherapy for glioblastomas," *Cell Death & Disease*, vol. 4, no. 10, p. e895, 2013.

[180] P. Godoy, S. Mello, D. Magalhães, F. Donaires, P. Nicolucci, E. A. Donadi, G. Passos, and E. T. Sakamoto-Hojo, "Ionizing radiation-induced gene expression changes in TP53 proficient and deficient glioblastoma cell lines," *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, vol. 756, no. 1-2, pp. 46–55, 2013.

[181] K. P. Bhat, V. Balasubramaniyan, B. Vaillant, R. Ezhilarasan, K. Hummelink, F. Hollingsworth, K. Wani, L. Heathcock, J. D. James, L. D. Goodman, *et al.*, "Mesenchymal differentiation mediated by NF-κB promotes radiation resistance in glioblastoma," *Cancer Cell*, vol. 24, no. 3, pp. 331–346, 2013.

[182] K. A. Effenberger, V. K. Urabe, and M. S. Jurica, "Modulating splicing with small molecular inhibitors of the spliceosome," *Wiley Interdisciplinary Reviews: RNA*, vol. 8, no. 2, p. e1381, 2017.

[183] H. Dvinge, E. Kim, O. Abdel-Wahab, and R. K. Bradley, "RNA splicing factors as oncoproteins and tumour suppressors," *Nature Reviews Cancer*, vol. 16, no. 7, p. 413, 2016.

[184] F. M. Meliso, C. G. Hubert, P. A. F. Galante, and L. O. Penalva, "Rna processing as an alternative route to attack glioblastoma," *Human Genetics*, vol. 136, no. 9, pp. 1129–1141, 2017.

[185] C. G. Hubert, R. K. Bradley, Y. Ding, C. M. Toledo, J. Herman, K. Skutt-Kakaria, E. J. Girard, J. Davison, J. Berndt, P. Corrin, *et al.*, "Genome-wide RNAi screens in human brain

tumor isolates reveal a novel viability requirement for PHF5A," *Genes & Development*, vol. 27, no. 9, pp. 1032–1045, 2013.

[186] M. Grzmil and B. A. Hemmings, "Translation regulation as a therapeutic target in cancer," *Cancer research*, vol. 72, no. 16, pp. 3891–3900, 2012.

[187] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, *et al.*, "GENCODE: the reference human genome annotation for The ENCODE Project," *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.

[188] S. Anders, P. T. Pyl, and W. Huber, "HTSeq - a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[189] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[190] P. Langfelder, R. Luo, M. C. Oldham, and S. Horvath, "Is my network module preserved and reproducible?," *PLoS Computational Biology*, vol. 7, no. 1, p. e1001057, 2011.

[191] K. Strimmer, "fdrtool: a versatile R package for estimating local and tail area-based false discovery rates," *Bioinformatics*, vol. 24, no. 12, pp. 1461–1462, 2008.

[192] S. Shen, J. W. Park, Z.-x. Lu, L. Lin, M. D. Henry, Y. N. Wu, Q. Zhou, and Y. Xing, "rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data," *Proceedings of the National Academy of Sciences*, vol. 111, no. 51, pp. E5593–E5601, 2014.

[193] J. A. Martin and Z. Wang, “Next-generation transcriptome assembly,” *Nature Reviews Genetics*, vol. 12, no. 10, p. 671, 2011.

[194] A. Fabregat, K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, *et al.*, “The reactome pathway knowledgebase,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D481–D487, 2015.

[195] H. Mi, B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, *et al.*, “The PANTHER database of protein families, subfamilies, functions and pathways,” *Nucleic Acids Research*, vol. 33, no. suppl_1, pp. D284–D288, 2005.

[196] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, “GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses,” *Nucleic Acids Research*, vol. 45, no. W1, pp. W98–W102, 2017.

[197] S. Ning, J. Zhang, P. Wang, H. Zhi, J. Wang, Y. Liu, Y. Gao, M. Guo, M. Yue, L. Wang, *et al.*, “Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D980–D985, 2015.

[198] K. C. Cotto, A. H. Wagner, Y.-Y. Feng, S. Kiwala, A. C. Coffman, G. Spies, A. Wollam, N. C. Spies, O. L. Griffith, and M. Griffith, “DGIdb 3.0: a redesign and expansion of the drug–gene interaction database,” *Nucleic Acids Research*, vol. 46, no. D1, pp. D1068–D1073, 2017.

[199] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, *et al.*, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Research*, vol. 44, no. W1, pp. W90–W97, 2016.

[200] D. Nandi, P. S. Cheema, N. Jaiswal, and A. Nag, "FoxM1: repurposing an oncogene as a biomarker," in *Seminars in Cancer Biology*, vol. 52, pp. 74–84, Elsevier, 2018.

[201] H.-Z. Chen, S.-Y. Tsai, and G. Leone, "Emerging roles of E2Fs in cancer: an exit from cell cycle control," *Nature Reviews Cancer*, vol. 9, no. 11, p. 785, 2009.

[202] L. S. Payne and P. H. Huang, "The pathobiology of collagens in glioma," *Molecular Cancer Research*, vol. 11, no. 10, pp. 1129–1140, 2013.

[203] S. Monferran, N. Skuli, C. Delmas, G. Favre, J. Bonnet, E. Cohen-Jonathan-Moyal, and C. Toulas, "αvβ3 and αvβ5 integrins control glioma cell response to ionising radiation through ilk and rhob," *International Journal of Cancer*, vol. 123, no. 2, pp. 357–364, 2008.

[204] Z. Du, C. Cai, M. Sims, F. A. Boop, A. M. Davidoff, and L. M. Pfeffer, "The effects of type I interferon on glioblastoma cancer stem cells," *Biochemical and Biophysical Research Communications*, vol. 491, no. 2, pp. 343–348, 2017.

[205] M. Budhwani, R. Mazzieri, and R. Dolcetti, "Plasticity of type I interferon-mediated responses in cancer therapy: from anti-tumor immunity to resistance," *Frontiers in Oncology*, vol. 8, p. 322, 2018.

[206] Z. Peng, C. Liu, and M. Wu, "New insights into long noncoding rnas and their roles in glioma," *Molecular Cancer*, vol. 17, no. 1, p. 61, 2018.

[207] X. Wu, Y. Wang, T. Yu, E. Nie, Q. Hu, W. Wu, T. Zhi, K. Jiang, X. Wang, X. Lu, *et al.*, "Blocking mir155hg/mir-155 axis inhibits mesenchymal transition in glioma," *Neuro-oncology*, vol. 19, no. 9, pp. 1195–1205, 2017.

[208] C. Miao, K. Zhao, J. Zhu, C. Liang, A. Xu, Y. Hua, J. Zhang, S. Liu, Y. Tian, C. Zhang, *et al.*, "Clinicopathological and prognostic role of long noncoding rna linc00152 in various human neoplasms: Evidence from meta-analysis," *BioMed Research International*, vol. 2017, pp. 1–11, 2017.

[209] H. Zhang, Y. Cai, L. Zheng, Z. Zhang, X. Lin, and N. Jiang, "Long noncoding rna neat1 regulate papillary thyroid cancer progression by modulating mir-129-5p/klk7 expression," *Journal of Cellular Physiology*, vol. 233, no. 10, pp. 6638–6648, 2018.

[210] W. Zhang, Y. Bi, J. Li, F. Peng, H. Li, C. Li, L. Wang, F. Ren, C. Xie, P. Wang, *et al.*, "Long noncoding rna ftx is upregulated in gliomas and promotes proliferation and invasion of glioma cells by negatively regulating mir-342-3p," *Laboratory Investigation*, vol. 97, no. 4, p. 447, 2017.

[211] H. Climente-González, E. Porta-Pardo, A. Godzik, and E. Eyras, "The functional impact of alternative splicing in cancer," *Cell Reports*, vol. 20, no. 9, pp. 2215–2226, 2017.

[212] E. Macaeva, Y. Saeys, K. Tabury, A. Janssen, A. Michaux, M. A. Benotmane, W. H. De Vos, S. Baatout, and R. Quintens, "Radiation-induced alternative transcription and splicing events and their applicability to practical biodosimetry," *Scientific Reports*, vol. 6, p. 19251, 2016.

[213] N. H. Binh, K. Satoh, K. Kobayashi, M. Takamatsu, Y. Hatano, A. Hirata, H. Tomita, T. Kuno, and A. Hara, “Galectin-3 in preneoplastic lesions of glioma,” *Journal of Neuro-Oncology*, vol. 111, no. 2, pp. 123–132, 2013.

[214] K. Orino and K. Watanabe, “Molecular, physiological and clinical aspects of the iron storage protein ferritin,” *The Veterinary Journal*, vol. 178, no. 2, pp. 191–201, 2008.

[215] M. Pang, X. Liu, B. Slagle-Webb, A. Madhankumar, and J. Connor, “Role of h-ferritin in radiosensitivity of human glioma cells,” *J Cancer Biol Treat*, vol. 3, no. 006, pp. 1–10, 2016.

[216] D.-H. Cho, Y.-M. Hong, H.-J. Lee, H.-N. Woo, J.-O. Pyo, T. W. Mak, and Y.-K. Jung, “Induced inhibition of ischemic/hypoxic injury by apip, a novel apaf-1-interacting protein,” *Journal of Biological Chemistry*, vol. 279, no. 38, pp. 39942–39950, 2004.

[217] C. Simion, M. E. Cedano-Prieto, and C. Sweeney, “The lrig family: enigmatic regulators of growth factor receptor signaling,” *Endocrine-related cancer*, vol. 21, no. 6, pp. R431–R443, 2014.

[218] Q. Xiao, Y. Tan, Y. Guo, H. Yang, F. Mao, R. Xie, B. Wang, T. Lei, and D. Guo, “Soluble lrig2 ectodomain is released from glioblastoma cells and promotes the proliferation and inhibits the apoptosis of glioblastoma cells in vitro and in vivo in a similar manner to the full-length lrig2,” *PLoS ONE*, vol. 9, no. 10, p. e111419, 2014.

[219] N. Dalla Venezia, A. Vincent, V. Marcel, F. Catez, and J.-J. Diaz, “Emerging Role of Eukaryote Ribosomes in Translational Control,” *International Journal of Molecular Sciences*, vol. 20, no. 5, p. 1226, 2019.

[220] L. N. Kent and G. Leone, “The broken cycle: E2f dysfunction in cancer,” *Nature Reviews Cancer*, vol. 19, no. 6, pp. 326–338, 2019.

[221] I. Thurlings and A. de Bruin, “E2f transcription factors control the roller coaster ride of cell cycle gene expression,” in *Methods in Molecular Biology*, pp. 71–88, Springer New York, 2016.

[222] M. M. Alonso, J. Fueyo, J. W. Shay, K. D. Aldape, H. Jiang, O.-H. Lee, D. G. Johnson, J. Xu, Y. Kondo, T. Kanzawa, *et al.*, “Expression of Transcription Factor E2F1 and Telomerase in Glioblastomas: Mechanistic Linkage and Prognostic Significance,” *Journal of the National Cancer Institute*, vol. 97, no. 21, pp. 1589–1600, 2005.

[223] X. Li, H. Zhang, and X. Wu, “Long noncoding RNA DLX6-AS1 accelerates the glioma carcinogenesis by competing endogenous sponging miR-197-5p to relieve e2f1,” *Gene*, vol. 686, pp. 1–7, 2019.

[224] L. Xia, D. Nie, G. Wang, C. Sun, and G. Chen, “FER1l4/miR-372/e2f1 works as a ceRNA system to regulate the proliferation and cell cycle of glioma cells,” *Journal of Cellular and Molecular Medicine*, vol. 23, no. 5, pp. 3224–3233, 2019.

[225] B. Yang, Q. Meng, Y. Sun, L. Gao, and J. Yang, “Long non-coding RNA SNHG16 contributes to glioma malignancy by competitively binding miR-20a-5p with E2F1.,” *Journal of Biological Regulators & Homeostatic Agents*, vol. 32, no. 2, pp. 251–261, 2018.

[226] N. Wu, L. Xiao, X. Zhao, J. Zhao, J. Wang, F. Wang, S. Cao, and X. Lin, “miR-125b regulates the proliferation of glioblastoma stem cells by targeting e2f2,” *FEBS Letters*, vol. 586, no. 21, pp. 3831–3839, 2012.

[227] O. K. Okamoto, S. M. Oba-Shinjo, L. Lopes, and S. K. N. Marie, "Expression of HOXC9 and E2F2 are up-regulated in CD133+ cells isolated from human astrocytomas and associate with transformation of human astrocytes," *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, vol. 1769, no. 7-8, pp. 437–442, 2007.

[228] H. Song, Y. Zhang, N. Liu, D. Zhang, C. Wan, S. Zhao, Y. Kong, and L. Yuan, "Let-7b inhibits the malignant behavior of glioma cells and glioma stem-like cells via downregulation of E2F2," *Journal of Physiology and Biochemistry*, vol. 72, no. 4, pp. 733–744, 2016.

[229] S. Qiu, D. Huang, D. Yin, F. Li, X. Li, H. fu Kung, and Y. Peng, "Suppression of tumorigenicity by MicroRNA-138 through inhibition of EZH2-CDK4/6-pRb-e2f1 signal loop in glioblastoma multiforme," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1832, no. 10, pp. 1697–1707, 2013.

[230] Y. Zhang, D. Han, W. Wei, W. Cao, R. Zhang, Q. Dong, J. Zhang, Y. Wang, and N. Liu, "MiR-218 inhibited growth and metabolism of human glioblastoma cells by directly targeting e2f2," *Cellular and Molecular Neurobiology*, vol. 35, no. 8, pp. 1165–1173, 2015.

[231] Y. Chen, F. Zhao, D. Cui, R. Jiang, J. Chen, Q. Huang, and J. Shi, "HOXD-AS1/miR-130a sponge regulates glioma development by targeting e2f8," *International Journal of Cancer*, vol. 142, no. 11, pp. 2313–2322, 2018.

[232] V. Gouazé-Andersson, M.-J. Ghérardi, A. Lemarié, J. Gilhodes, V. Lubrano, F. Arnauduc, E. C.-J. Moyal, and C. Toulas, "FGFR1/FOXM1 pathway: a key regulator of glioblastoma stem cells radioresistance and a prognosis biomarker," *Oncotarget*, vol. 9, no. 60, p. 31637, 2018.

[233] Q. Ma, Y. Liu, L. Shang, J. Yu, and Q. Qu, “The FOXM1/BUB1b signaling pathway is essential for the tumorigenicity and radioresistance of glioblastoma,” *Oncology Reports*, vol. 38, no. 6, pp. 3367–3375, 2017.

[234] S. Zhang, B. S. Zhao, A. Zhou, K. Lin, S. Zheng, Z. Lu, Y. Chen, E. P. Sulman, K. Xie, O. Bögler, S. Majumder, C. He, and S. Huang, “m 6 a demethylase ALKBH5 maintains tumorigenicity of glioblastoma stem-like cells by sustaining FOXM1 expression and cell proliferation program,” *Cancer Cell*, vol. 31, no. 4, pp. 591–606.e6, 2017.

[235] J.-J. Quan, J.-N. Song, and J.-Q. Qu, “PARP3 interacts with FoxM1 to confer glioblastoma cell radioresistance,” *Tumor Biology*, vol. 36, no. 11, pp. 8617–8624, 2015.

[236] A. hua Gong, P. Wei, S. Zhang, J. Yao, Y. Yuan, A. dong Zhou, F. F. Lang, A. B. Heimberger, G. Rao, and S. Huang, “FoxM1 drives a feed-forward STAT3-activation signaling loop that promotes the self-renewal and tumorigenicity of glioblastoma stem-like cells,” *Cancer Research*, vol. 75, no. 11, pp. 2337–2348, 2015.

[237] N. Zhang, X. Wu, L. Yang, F. Xiao, H. Zhang, A. Zhou, Z. Huang, and S. Huang, “FoxM1 inhibition sensitizes resistant glioblastoma cells to temozolomide by downregulating the expression of DNA-repair gene rad51,” *Clinical Cancer Research*, vol. 18, no. 21, pp. 5961–5971, 2012.

[238] M. Fischer, P. Grossmann, M. Padi, and J. A. DeCaprio, “Integration of TP53, DREAM, MMB-FOXM1 and RB-e2f target gene analyses identifies cell cycle gene regulatory networks,” *Nucleic Acids Research*, vol. 44, no. 13, pp. 6070–6086, 2016.

[239] H. Bai, A. S. Harmancı, E. Z. Erson-Omay, J. Li, S. Coșkun, M. Simon, B. Krischek, K. Özduman, S. B. Omay, E. A. Sorensen, Ș. Turcan, M. Bakırcığlu, G. Carrión-Grant, P. B. Murray, V. E. Clark, A. G. Ercan-Sencicek, J. Knight, L. Sencar, S. Altınok, L. D. Kaulen, B. Gülez, M. Timmer, J. Schramm, K. Mishra-Gorur, O. Henegariu, J. Moliterno, A. Louvi, T. A. Chan, S. L. Tannheimer, M. N. Pamir, A. O. Vortmeyer, K. Bilguvar, K. Yasuno, and M. Günel, "Integrated genomic characterization of IDH1-mutant glioma malignant progression," *Nature Genetics*, vol. 48, no. 1, pp. 59–66, 2015.

[240] A. C. Belkina and G. V. Denis, "BET domain co-regulators in obesity, inflammation and cancer," *Nature Reviews Cancer*, vol. 12, no. 7, pp. 465–477, 2012.

[241] L. Xu, Y. Chen, A. Mayakonda, L. Koh, Y. K. Chong, D. L. Buckley, E. Sandanaraj, S. W. Lim, R. Y.-T. Lin, X.-Y. Ke, M.-L. Huang, J. Chen, W. Sun, L.-Z. Wang, B. C. Goh, H. Q. Dinh, D. Kappei, G. E. Winter, L.-W. Ding, B. T. Ang, B. P. Berman, J. E. Bradner, C. Tang, and H. P. Koeffler, "Targetable BET proteins- and e2f1-dependent transcriptional program maintains the malignancy of glioblastoma," *Proceedings of the National Academy of Sciences*, vol. 115, no. 22, pp. E5086–E5095, 2018.

[242] F. M. Meliso, C. G. Hubert, P. A. F. Galante, and L. O. Penalva, "RNA processing as an alternative route to attack glioblastoma," *Human Genetics*, vol. 136, no. 9, pp. 1129–1141, 2017.

[243] B. R. Correa, P. R. de Araujo, M. Qiao, S. C. Burns, C. Chen, R. Schlegel, S. Agarwal, P. A. F. Galante, and L. O. F. Penalva, "Functional genomics analyses of RNA-binding

proteins reveal the splicing regulator SNRPB as an oncogenic candidate in glioblastoma," *Genome Biology*, vol. 17, no. 1, p. 125, 2016.

[244] Y. Liu, Y. Shen, T. Sun, and W. Yang, "Mechanisms regulating radiosensitivity of glioma stem cells," *Neoplasma*, vol. 64, no. 05, pp. 655–665, 2017.

[245] J. Wang, T. P. Wakeman, J. D. Lathia, A. B. Hjelmeland, X.-F. Wang, R. R. White, J. N. Rich, and B. A. Sullenger, "Notch Promotes Radioresistance of Glioma Stem Cells," *Stem Cells*, vol. 28, no. 1, pp. 17–28, 2010.

[246] R. Hannen, M. Hauswald, and J. W. Bartsch, "A rationale for targeting extracellular regulated kinases ERK1 and ERK2 in glioblastoma," *Journal of Neuropathology & Experimental Neurology*, vol. 76, no. 10, pp. 838–847, 2017.

[247] L. Xu, Y. Chen, M. Dutra-Clarke, A. Mayakonda, M. Hazawa, S. E. Savinoff, N. Doan, J. W. Said, W. H. Yong, A. Watkins, *et al.*, "Bcl6 promotes glioma and serves as a therapeutic target," *Proceedings of the National Academy of Sciences*, vol. 114, no. 15, pp. 3981–3986, 2017.

[248] E. Cho and Y. Yen, "Novel regulators and molecular mechanisms of p53R2 and its disease relevance," *Biochimie*, vol. 123, pp. 81–84, 2016.

[249] L. Zhai, E. Ladomersky, A. Lenzen, B. Nguyen, R. Patel, K. L. Lauing, M. Wu, and D. A. Wainwright, "Ido1 in cancer: a gemini of immune checkpoints," *Cellular & Molecular Immunology*, vol. 15, no. 5, p. 447, 2018.

[250] P. Kesarwani, A. Prabhu, S. Kant, P. Kumar, S. F. Graham, K. L. Buelow, G. D. Wilson, C. R. Miller, and P. Chinnaiyan, "Tryptophan metabolism contributes to radiation-induced

immune checkpoint reactivation in glioblastoma," *Clinical Cancer Research*, vol. 24, no. 15, pp. 3632–3643, 2018.

[251] R. O. Hynes and A. Naba, "Overview of the matrisome–an inventory of extracellular matrix constituents and functions," *Cold Spring Harbor Perspectives in Biology*, vol. 4, no. 1, pp. a004903–a004903, 2011.

[252] C. Heit, B. C. Jackson, M. McAndrews, M. W. Wright, D. C. Thompson, G. A. Silverman, D. W. Nebert, and V. Vasiliou, "Update of the human and mouse SERPIN gene superfamily," *Human Genomics*, vol. 7, no. 1, p. 22, 2013.

[253] S. Bonnal, L. Vigevani, and J. Valcárcel, "The spliceosome as a target of novel antitumour drugs," *Nature Reviews Drug Discovery*, vol. 11, no. 11, pp. 847–859, 2012.

[254] I. Pal, M. Safari, M. Jovanovic, S. E. Bates, and C. Deng, "Targeting translation of mRNA as a therapeutic strategy in cancer," *Current Hematologic Malignancy Reports*, pp. 1–9, 2019.

[255] S. L. Salzberg, "Open questions: How many genes do we have?," vol. 16, no. 1.

[256] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. F. Banet, K. Billis, C. G. Girón, T. Hourlier, K. Howe, A. Kähäri, F. Kokocinski, F. J. Martin, D. N. Murphy, R. Nag, M. Ruffier, M. Schuster, Y. A. Tang, J.-H. Vogel, S. White, A. Zadissa, P. Flicek, and S. M. J. Searle, "The ensembl gene annotation system," *Database*, vol. 2016, p. baw093, 2016.

[257] A. Bairoch, "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000," *Nucleic Acids Research*, vol. 28, pp. 45–48, jan 2000.

[258] S. J. Kiniry, A. M. Michel, and P. V. Baranov, "The GWIPS-viz browser," *Current Protocols in Bioinformatics*, vol. 62, p. e50, may 2018.

[259] S.-Q. Xie, P. Nie, Y. Wang, H. Wang, H. Li, Z. Yang, Y. Liu, J. Ren, and Z. Xie, "RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling," *Nucleic Acids Research*, vol. 44, pp. D254–D258, oct 2015.

[260] W.-S. Wu, Y.-X. Jiang, J.-W. Chang, Y.-H. Chu, Y.-H. Chiu, Y.-H. Tsao, T. E. M. Nordling, Y.-Y. Tseng, and J. T. Tseng, "HRPDviewer: Human ribosome profiling data viewer," *Database*, vol. 2018, Jan 2018. 00000.

[261] M. Matsui, N. Yachie, Y. Okada, R. Saito, and M. Tomita, "Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse," *FEBS Letters*, vol. 581, pp. 4184–4188, aug 2007.

[262] S. E. Calvo, D. J. Pagliarini, and V. K. Mootha, "Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 7507–7512, apr 2009.

[263] G. Menschaert, W. V. Criekinge, T. Notelaers, A. Koch, J. Crappé, K. Gevaert, and P. V. Damme, "Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events," *Molecular & Cellular Proteomics*, vol. 12, pp. 1780–1790, feb 2013.

[264] M. Kozak, "Pushing the limits of the scanning mechanism for initiation of translation," *Gene*, vol. 299, pp. 1–34, oct 2002.

[265] H. M. Hood, D. E. Neafsey, J. Galagan, and M. S. Sachs, "Evolutionary roles of upstream open reading frames in mediating gene regulation in fungi," *Annual Review of Microbiology*, vol. 63, pp. 385–409, oct 2009.

[266] N. Sonenberg and A. G. Hinnebusch, "Regulation of translation initiation in eukaryotes: Mechanisms and biological targets," *Cell*, vol. 136, pp. 731–745, feb 2009.

[267] M. Iacono, F. Mignone, and G. Pesole, "uAUG and uORFs in human and rodent 5′untranslated mRNAs," *Gene*, vol. 349, pp. 97–105, apr 2005.

[268] A. Franks, E. Airoldi, and N. Slavov, "Post-transcriptional regulation across human tissues," *PLOS Computational Biology*, vol. 13, p. e1005535, may 2017.

[269] P. D. Lu, H. P. Harding, and D. Ron, "Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response," *The Journal of Cell Biology*, vol. 167, pp. 27–33, oct 2004.

[270] M. J. Clemens, "Initiation Factor eIF2α Phosphorylation in Stress Responses and Apoptosis," in *Signaling Pathways for Translation*, pp. 57–89, Springer Berlin Heidelberg, 2001.

[271] J. Medenbach, M. Seiler, and M. W. Hentze, "Translational control via protein-regulated upstream open reading frames," *Cell*, vol. 145, pp. 902–913, jun 2011.

[272] H. A. Meijer and A. A. M. Thomas, "Control of eukaryotic protein synthesis by upstream open reading frames in the 5h-untranslated region of an mRNA," p. 11, 2002.

[273] P. D. Lu, H. P. Harding, and D. Ron, “Translation reinitiation at alternative open reading frames regulates gene expression in an integrated stress response,” *The Journal of Cell Biology*, vol. 167, pp. 27–33, oct 2004.

[274] J. Medenbach, M. Seiler, and M. W. Hentze, “Translational control via protein-regulated upstream open reading frames,” *Cell*, vol. 145, pp. 902–913, jun 2011.

[275] S. STONE, “PATHOLOGICAL SCIENCE,” in *Flavor Physics for the Millennium*, WORLD SCIENTIFIC, sep 2001.

[276] J. Schulz, N. Mah, M. Neuenschwander, T. Kischka, R. Ratei, P. M. Schlag, E. Castaños-Vélez, I. Fichtner, P.-U. Tunn, C. Denkert, *et al.*, “Loss-of-function uorf mutations in human malignancies,” *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[277] T. Ouspenskaia, T. Law, K. R. Clauser, S. Klaeger, S. Sarkizova, F. Aguet, B. Li, E. Christian, B. A. Knisbacher, P. M. Le, *et al.*, “Thousands of novel unannotated proteins expand the MHC I immunopeptidome in cancer,” *bioRxiv*, 2020.

[278] F.-B. Gao, J. D. Richter, and D. W. Cleveland, “Rethinking unconventional translation in neurodegeneration,” *Cell*, vol. 171, pp. 994–1000, Nov. 2017.

[279] C. W. Dunn, X. Luo, and Z. Wu, “Phylogenetic analysis of gene expression,” *Integrative and Comparative Biology*, vol. 53, pp. 847–856, jun 2013.

[280] S. H. Wang, C. J. Hsiao, Z. Khan, and J. K. Pritchard, “Post-translational buffering leads to convergent protein expression levels between primates,” *Genome Biology*, vol. 19, p. 83, Dec 2018. 00000.

[281] Z. Khan, M. J. Ford, D. A. Cusanovich, A. Mitrano, J. K. Pritchard, and Y. Gilad, "Primate Transcript and Protein Expression Levels Evolve Under Compensatory Selection Pressures," *Science*, vol. 342, pp. 1100–1104, Nov 2013. 00133.

[282] P. Spealman, A. W. Naik, G. E. May, S. Kuersten, L. Freeberg, R. F. Murphy, and J. McManus, "Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data," *Genome Research*, vol. 28, pp. 214–222, Feb. 2018. 00005.

[283] T. F. Martinez, Q. Chu, C. Donaldson, D. Tan, M. N. Shokhirev, and A. Saghatelian, "Accurate annotation of human protein-coding small open reading frames," *Nature Chemical Biology*, Dec. 2019. 00000.

[284] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, Oct. 1990.

[285] J. Shine and L. Dalgarno, "Occurrence of heat-dissociable ribosomal RNA in insects: The presence of three polynucleotide chains in 26 s RNA from cultured aedes aegypti cells," *Journal of Molecular Biology*, vol. 75, pp. 57–72, Mar. 1973.

[286] G.-L. Chew, A. Pauli, and A. F. Schier, "Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish," *Nature Communications*, vol. 7, May 2016.

[287] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. Noble, "Quantifying similarity between motifs," *Genome Biology*, vol. 8, no. 2, p. R24, 2007.

[288] A. Cassese, M. Guindani, M. G. Tadesse, F. Falciani, and M. Vannucci, "A hierarchical bayesian model for inference of copy number variants and their association to gene expression," *The Annals of Applied Statistics*, vol. 8, pp. 148–175, mar 2014.

[289] C. Briceno, “Massive variability surveys from venezuela,” *arxiv:astro-ph/0304081v1*, 2003.

[290] D. Koop, J. Freire, and C. T. Silva, “Enabling reproducible science with vistrails,” *arxiv:1309.1784v2*, 2013.

[291] N. J. Kalton, “The basic sequence problem,” *arxiv:math/9408205v1*, 1994.

[292] R. Luo, J. Xu, Y. Zhang, X. Ren, and X. Sun, “Pkuseg: A toolkit for multi-domain chinese word segmentation,” *arxiv:1906.11455v2*, 2019.

[293] R. Cogranne, “Determining jpeg image standard quality factor from the quantization tables,” *arxiv:1802.00992v1*, 2018.

[294] Y. Zhu, R. M. Stephens, P. S. Meltzer, and S. R. Davis, “SRAdb: query and use public next-generation sequencing data from within r,” *BMC Bioinformatics*, vol. 14, no. 1, p. 19, 2013.

[295] J. Zhu and S. Davis, “Bioconductor:sradb,” Dec 2018.

[296] J. Wilson, “Statistical computing with r: selecting the right tool for the job-r commander or something else?,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 4, pp. 518–526, aug 2012.

[297] S. Choudhary, “saketkc/pysradb v0.9.0,” Feb 2019.

[298] D. Faes, “Use of python programming language in astronomy and science,” *Journal of Computational Interdisciplinary Sciences*, vol. 3, no. 3, 2012.

[299] B. Grüning, , R. Dale, A. Sjödin, B. A. Chapman, J. Rowe, C. H. Tomkins-Tinch, R. Valieris, and J. Köster, "Bioconda: sustainable and comprehensive software distribution for the life sciences," *Nature Methods*, vol. 15, pp. 475–476, jul 2018.

[300] M. A. Breddels and J. Veljanoski, "Vaex: big data exploration in the era of gaia," *Astronomy & Astrophysics*, vol. 618, p. A13, oct 2018.

[301] C. da Costa-Luis, S. L., H. Mary, noamraph, M. Korobov, I. Ivanov, M. Bargull, James, G. Chen, M. D. Pagel, S. Malmgren, Socialery, J. McCracken, F. Dill, D. Panteleit, A. Rothberg, Y. Halchenko, T. Ostasevicius, S. Pokharel, ReadmeCritic, P. VandeHaar, K. che Wu, jcea, Hugo, F. Hurley, E. Betts, D. Bau, A. Persaud, Alexander, and A. Umer, "tqdm/tqdm: tqdm v4.20.0 stable," Apr 2018.

[302] R. Teymourzadeh, S. A. Ahmed, K. W. Chan, and M. V. Hoong, "Smart GSM based home automation system," in *2013 IEEE Conference on Systems, Process & Control (ICSPC)*, IEEE, dec 2013.

[303] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, and C. Willing, "Jupyter notebooks – a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (F. Loizides and B. Schmidt, eds.), pp. 87 – 90, IOS Press, 2016.

[304] D. R. Schrider and A. D. Kern, "Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain," *Genome Biology and Evolution*, vol. 7, pp. 3511–3528, nov 2015.

[305] J. D. Blair, D. Hockemeyer, J. A. Doudna, H. S. Bateup, and S. N. Floor, “Widespread translational remodeling during human neuronal differentiation,” *Cell reports*, vol. 21, no. 7, pp. 2005–2016, 2017.

[306] J. Merkin, C. Russell, P. Chen, and C. B. Burge, “Evolutionary dynamics of gene and isoform regulation in mammalian tissues,” *Science*, vol. 338, no. 6114, pp. 1593–1599, 2012.

[307] M. D. Schultz, Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, *et al.*, “Human body epigenome maps reveal noncanonical dna methylation variation,” *Nature*, vol. 523, no. 7559, p. 212, 2015.

Appendix A

pysradb: A Python package to query next-generation sequencing metadata and data from NCBI Sequence Read Archive

A.1 Introduction

Several projects have made efforts to analyze and publish summaries of DNA- [288] and RNA-seq [289, 290] datasets. Obtaining metadata and raw data from the NCBI Sequence Read Archive (SRA) [291] is often the first step towards re-analyzing public next-generation sequencing datasets in order to compare them to private data or test a novel hypothesis. The NCBI SRA toolkit [292] provides utility methods to download raw sequencing data, while the metadata can be obtained by querying the website or through the Entrez `efetch` command line utility [293]. Most workflows analyzing public data rely on first searching for relevant keywords in the metadata either through the command line utility or the website, gathering relevant sample(s) of interest and then downloading these. A more streamlined workflow can enable the performance of all these steps at once.

In order to make querying both metadata and data more precise and robust, the SRAdb [294] project provides a frequently updated SQLite database containing all the metadata parsed from SRA. SRAdb tracks the five main data objects in SRA's metadata: submission, study, sample, experiment and run. These are mapped to five different relational database tables that are made available in the SQLite file. The metadata semantics in the file remain as they are on SRA. The accompanying package, SRAdb [295], made available in the R programming language [296], provides a convenient framework to handle metadata queries and raw data downloads by utilizing the SQLite database. Though powerful, SRAdb requires the end user to be familiar with the R programming language and does not provide a command-line interface for querying or downloading operations.

The pysradb package [297] builds upon the principles of SRAdb, providing a simple and user-friendly command-line interface for querying metadata and downloading datasets from SRA. It obviates the need for the user to be familiar with any programming language for querying and downloading datasets from SRA. Additionally, it provides utility functions that will further help a user perform more granular queries, which are often required when dealing with multiple datasets on a large scale. By enabling both metadata search and download operations at the command-line, pysradb aims to bridge the gap in seamlessly retrieving public sequencing datasets and the associated metadata.

pysradb [297] is written in Python [298] and is currently developed on GitHub under the open-source BSD 3-Clause License. To simplify the installation procedure for the end-user, it is also available for download through PyPI and bioconda [299].

A.2 Methods

A.2.1 Implementation

pysradb [297] is implemented in Python and uses pandas [300] for data frame based operations. Since downloading datasets can often take a long time, pysradb displays progress for long haul tasks using tqdm [301]. The metadata information is read in the form of an SQLite [302] database, made available by SRAdb [294].

Each sub-command of pysradb contains a self-contained help string that describes its purpose and usage example. The help text can be accessed by passing the '–help' flag. There is also additional documentation available for the sub-commands on the project's website. We also provide example Jupyter [303] notebooks that demonstrate the functionality of the Python API.

pysradb's development primarily occurred on GitHub and the code is tested continuously using Travis CI webhook. This monitors all incoming pull requests and commits to the master branch. The testing happens on Python version 3.5, 3.6, and 3.7 on an Ubuntu 16.04 LTS virtual machine, while testing webhooks on the bioconda channel provide additional testing on

Mac-based systems. Nevertheless, pysradb should run on most Unix derivatives.

A.2.2 Operation

pysradb [297] can be run on either Linux- or Mac-based operating systems. It supports Python 3.5, 3.6 and 3.7. Requiring just two additional dependencies, pysradb can be easily installed using either a pip- or conda- based package manager via the bioconda [299] channel.

An earlier version of this article can be found on bioRxiv https://doi.org/10.1101/578500

A.3 Use cases

pysradb [297] provides a chain of sub-commands for retrieving metadata, converting one accession to other and downloading. Each sub-command is designed to perform a single operation by default, while additional operations can be performed by passing additional flags. In the following section we demonstrate some of the use cases of these sub-commands.

pysradb uses SRAmetadb.sqlite, a SQLite file produced and made available by SRAdb [294] project. The file itself can be downloaded using pysradb as:

```
$ pysradb srametadb
```

The SRAmetadb.sqlite file is required for all other operations supported by pysradb. This file is required for all the sub-commands to function. By default, pysradb assumes that the file is located in the current working directory. Alternatively, it can supplied using the '–db path/to/SRAmetadb.sqlite' argument. The SRAmetadb.sqlite that is required for all underlying operations of pysradb is available at: https://s3.amazonaws.com/starbuck1/sradb/SRAmetadb.sqlite.gz or alternatively at https://gbnci-abcc.ncifcrf.gov/backup/SRAmetadb.sqlite.gz. The examples here were run using SRAmetadb.sqlite with schema version 1.0 and creation timestamp 2019-01-25 00:38:19.

A.3.1 Search

Consider a case where a user is looking for Ribo-seq [148] public datasets on SRA. These datasets will often have 'ribosome profiling' appearing in the abstract or sample description. We can search for such projects using the 'search' sub-command:

```
$ pysradb search `"ribosome profiling"' | head
```

study_accession	experiment_accession	sample_accession	run_accession
DRP003075	DRX019536	DRS026974	DRR021383
DRP003075	DRX019537	DRS026982	DRR021384
DRP003075	DRX019538	DRS026979	DRR021385
DRP003075	DRX019540	DRS026984	DRR021387
DRP003075	DRX019541	DRS026978	DRR021388
DRP003075	DRX019543	DRS026980	DRR021390
DRP003075	DRX019544	DRS026981	DRR021391
ERP013565	ERX1264364	ERS1016056	ERR1190989

The results here list all relevant ‘ribosome profiling’ projects.

A.3.2 Getting metadata for a SRA project

Each SRA project (accession prefix ‘SRP’) on SRA consists of single or multiple experiments (accession prefix ‘SRX’) which are sequenced as single or multiple runs (accession prefix ‘SRR’). Each experiment is carried out on an individual biological sample (accession prefix ‘SRS’). `pysradb metadata` can be used to obtain all the experiment, sample, and run accessions associated with a SRA project as:

```
$ pysradb metadata SRP010679 | head
```

study_accession	experiment_accession	sample_accession	run_accession
SRP010679	SRX118285	SRS290854	SRR403882
SRP010679	SRX118286	SRS290855	SRR403883
SRP010679	SRX118287	SRS290856	SRR403884
SRP010679	SRX118288	SRS290857	SRR403885
SRP010679	SRX118289	SRS290858	SRR403886
SRP010679	SRX118290	SRS290859	SRR403887
SRP010679	SRX118291	SRS290860	SRR403888
SRP010679	SRX118292	SRS290861	SRR403889
SRP010679	SRX118293	SRS290862	SRR403890
SRP010679	SRX118294	SRS290863	SRR403891
SRP010679	SRX118295	SRS290864	SRR403892
SRP010679	SRX118296	SRS290865	SRR403893

However, this information by itself is often incomplete. We require detailed metadata associated with each sample to perform any downstream analysis. For example, the assays used for different samples and the corresponding treatment conditions. This can be done by supplying the ‘–desc’ flag:

```
$ pysradb metadata SRP010679 --desc | head -5
```

study_accession	experiment_accession	sample_accession	run_accession	sample_attribute
SRP010679	SRX118285	SRS290854	SRR403882	source_name: PC3 human prostate cancer cells || cell line: PC3 || sample type: polyA RNA || treatment: vehicle
SRP010679	SRX118286	SRS290855	SRR403883	source_name: PC3 human prostate cancer cells || cell line: PC3 || sample type: ribosome protected RNA || treatment: vehicle
SRP010679	SRX118287	SRS290856	SRR403884	source_name: PC3 human prostate cancer cells || cell line: PC3 || sample type: polyA RNA || treatment: rapamycin
SRP010679	SRX118288	SRS290857	SRR403885	source_name: PC3 human prostate cancer cells || cell line: PC3 || sample type: ribosome protected RNA || treatment: rapamycin

This can be further expanded to reveal the data in ‘sample_attribute’ column into separate columns via ‘–expand’ flag. This is most useful for samples that have associated treatment or cell type metadata available.

```
$ pysradb metadata SRP010679 --desc --expand
```

```
... [truncated]
run_accession  cell_line  sample_type              source_name                      treatment
SRR403882      pc3        polya rna                pc3 human prostate cancer cells  vehicle
SRR403883      pc3        ribosome protected rna   pc3 human prostate cancer cells  vehicle
SRR403884      pc3        polya rna                pc3 human prostate cancer cells  rapamycin
SRR403885      pc3        ribosome protected rna   pc3 human prostate cancer cells  rapamycin
SRR403886      pc3        polya rna                pc3 human prostate cancer cells  pp242
SRR403887      pc3        ribosome protected rna   pc3 human prostate cancer cells  pp242
SRR403888      pc3        polya rna                pc3 human prostate cancer cells  vehicle
SRR403889      pc3        ribosome protected rna   pc3 human prostate cancer cells  vehicle
SRR403890      pc3        polya rna                pc3 human prostate cancer cells  rapamycin
SRR403891      pc3        ribosome protected rna   pc3 human prostate cancer cells  rapamycin
SRR403892      pc3        polya rna                pc3 human prostate cancer cells  pp242
SRR403893      pc3        ribosome protected rna   pc3 human prostate cancer cells  pp242
```

Any SRA project might consist of experiments involving multiple assay types. The assay associated with any project can be obtained by providing `--assay` flag:

```
$ pysradb metadata SRP000941 --assay | tr -s ' ' | cut -f5 -d ' ' | tail -n +2 | sort | uniq
-c
```

```
999   Bisulfite-Seq
768   ChIP-Seq
121   OTHER
353   RNA-Seq
28    WGS
```

A.3.3 Getting SRPs from GSE

The Gene Expression Omnibus database (GEO) [304] is the NCBI data repository for functional genomics data. It accepts array and sequence-based data from gene profiling experiments. For sequence-based data, the corresponding raw files are deposited to the SRA. GEO assigns a dataset accession (accession prefix ‘GSE’) that is linked to the corresponding accession on the SRA (accession prefix ‘SRP’). It is often necessary to interpolate between the two accessions. `gse-to-srp` sub-command allows converting GSE to SRP:

```
$ pysradb gse-to-srp GSE24355 GSE25842
```

```
study_alias   study_accession
GSE24355      SRP003870
GSE25842      SRP005378
```

It can be further expanded to obtain the corresponding experiment and run accessions:

```
$ pysradb gse-to-srp --detailed --expand GSE100007 | head
```

study_alias	study_accession	experiment_accession	sample_accession	experiment_alias	sample_alias
GSE100007	SRP109126	SRX2916198	SRS2282390	GSM2667747	GSM2667747
GSE100007	SRP109126	SRX2916199	SRS2282391	GSM2667748	GSM2667748
GSE100007	SRP109126	SRX2916200	SRS2282392	GSM2667749	GSM2667749
GSE100007	SRP109126	SRX2916201	SRS2282393	GSM2667750	GSM2667750
GSE100007	SRP109126	SRX2916202	SRS2282394	GSM2667751	GSM2667751
GSE100007	SRP109126	SRX2916203	SRS2282395	GSM2667752	GSM2667752
GSE100007	SRP109126	SRX2916204	SRS2282396	GSM2667753	GSM2667753
GSE100007	SRP109126	SRX2916205	SRS2282397	GSM2667754	GSM2667754
GSE100007	SRP109126	SRX2916206	SRS2282400	GSM2667755	GSM2667755

A.3.4 Getting a list of GEO experiments for a GEO study

Any GEO study (accession prefix ‘GSE’) will involve a collection of experiments (accession prefix ‘GSM’). We can obtain an entire list of experiments corresponding to the study using the gse-to-gsm sub-command from pysradb:

```
$ pysradb gse-to-gsm GSE41637 | head
```

study_alias	experiment_alias
GSE41637	GSM1020640_1
GSE41637	GSM1020641_1
GSE41637	GSM1020642_1
GSE41637	GSM1020643_1
GSE41637	GSM1020644_1
GSE41637	GSM1020645_1
GSE41637	GSM1020646_1
GSE41637	GSM1020647_1
GSE41637	GSM1020648_1

However, a list of GSM accessions is not useful if one is performing any downstream analysis, which essentially requires more detailed information about the metadata associated with each experiment. This relevant metadata associated with each sample can be obtained by providing `gse-to-gsm` additional flags:

```
$ pysradb gse-to-gsm --desc GSE41637 | head
```

study_alias	experiment_alias	sample_attribute
GSE41637	GSM1020640_1	source_name: mouse_brain || strain: DBA/2J || tissue: brain
GSE41637	GSM1020641_1	source_name: mouse_colon || strain: DBA/2J || tissue: colon
GSE41637	GSM1020642_1	source_name: mouse_heart || strain: DBA/2J || tissue: heart
GSE41637	GSM1020643_1	source_name: mouse_kidney || strain: DBA/2J || tissue: kidney
GSE41637	GSM1020644_1	source_name: mouse_liver || strain: DBA/2J || tissue: liver
GSE41637	GSM1020645_1	source_name: mouse_lung || strain: DBA/2J || tissue: lung
GSE41637	GSM1020646_1	source_name: mouse_skm || strain: DBA/2J || tissue: skeletal muscle
GSE41637	GSM1020647_1	source_name: mouse_spleen || strain: DBA/2J || tissue: spleen
GSE41637	GSM1020648_1	source_name: mouse_testes || strain: DBA/2J || tissue: testes

The metadata information can then be parsed from the `sample_attribute` column. To obtain more structured metadata, we can use an additional flag '`--expand`':

```
$ pysradb gse-to-gsm --desc --expand GSE41637 | head
```

study_alias	experiment_alias	source_name	strain	tissue
GSE41637	GSM1020640_1	mouse_brain	dba/2j	brain
GSE41637	GSM1020641_1	mouse_colon	dba/2j	colon
GSE41637	GSM1020642_1	mouse_heart	dba/2j	heart
GSE41637	GSM1020643_1	mouse_kidney	dba/2j	kidney
GSE41637	GSM1020644_1	mouse_liver	dba/2j	liver
GSE41637	GSM1020645_1	mouse_lung	dba/2j	lung
GSE41637	GSM1020646_1	mouse_skm	dba/2j	skeletal muscle

A.3.5 Getting SRR from GSM

`gsm-to-srr` allows conversion from GEO experiments (accession prefix ‘GSM’) to SRA runs (accession prefix ‘SRR’):

```
$ pysradb gsm-to-srr GSM1020640 GSM1020646
```

experiment_alias	run_accession
GSM1020640_1	SRR594393
GSM1020646_1	SRR594399

A.3.6 Downloading SRA datasets

`pysradb` enables seemless downloads from SRA. It organizes the downloaded data following the NCBI hiereracby: ‘SRP => SRX => SRR’ of storing data. Each ‘SRP’ (project) has multiple ‘SRX’ (experiments) and each ‘SRX’ in turn has multiple ‘SRR’ (runs). Multiple projects can be downloaded at once using the `download` sub-command:

```
$ pysradb download -p SRP000941 -p SRP010679
```

download also allows Unix pipes-based inputs. Consider our previous example of the project SRP000941 with different assays. However, we want to be able to download only ‘RNA-seq’ samples. We can do this by subsetting the metadata output for only ‘RNA-seq’ samples:

```
$ pysradb metadata SRP000941 --assay | grep `study|RNA-Seq' | pysradb download
```

This will only download the ‘RNA-seq’ samples from the project.

A.4 Summary

pysradb [297] provides a command-line interface to query metadata and download sequencing datasets from the SRA. It enables seamless retrieval of metadata and conversion between different accessions. pysradb is written in Python 3 and is available on Linux and Mac OS. The source code is hosted on GitHub and licensed under BSD 3-clause license. It is available for installation through PyPI and bioconda.

A.5 Data availability

Underlying data

Dataset from DDBJ Sequence Read Archive, Accession number DRP003075: https://identifiers.org/insdc.sra/DRP003075

Dataset from EMBL-EBI Sequence Read Archive, Accession number ERP013565: https://identifiers.org/insdc.sra/ERP013565

Dataset from Gene Expression Omnibus, Accession number GSE24355: https://identifiers.org/geo/GSE24355

Dataset from Gene Expression Omnibus, Accession number GSE25842: https://identifiers.org/geo/GSE25842

Dataset from Gene Expression Omnibus, Accession number GSE100007: https://identifiers.org/geo/GSE100007 [305]

Dataset from Gene Expression Omnibus, Accession number GSE41637: https://identifiers.org/geo/GSE41637 [306]

Dataset from NCBI Sequence Read Archive, Accession number SRP010679: https://identifiers.org/insdc.sra/SRP010679 [103]

Dataset from NCBI Sequence Read Archive, Accession number SRP000941: https://identifiers.org/insdc.sra/SRP000941 [307]

A.6 Software availability

Software available from: https://pypi.org/project/pysradb/.

Source code available from: https://github.com/saketkc/pysradb.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.2579446 [297].

License: BSD 3-Clause