CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Representing translation: concepts, methods, and resources for protein coding loci representation and ribosome profiling data analysis
Authors	Tierney, Jack
Publication date	2024
Original Citation	Tierney, J. 2024. Representing translation: concepts, methods, and resources for protein coding loci representation and ribosome profiling data analysis. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2024, Jack Tierney. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-16 14:55:47
Item downloaded from	https://hdl.handle.net/10468/16960

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork



Representing Translation: Concepts, Methods, and Resources for Protein Coding Loci Representation and Ribosome Profiling Data Analysis

Thesis presented by

Jack Tierney MSc. BSc.

for the degree of

Doctor of Philosophy

University College Cork

School of Biochemistry and Cell Biology

Head of School/Department: Professor Justin McCarthy

Supervisor(s): Professor Pavel Baranov

2024

Table of Contents

Table of Contents 1
Acknowledgements 3
Declaration 6
Abstract 7
Lay Abstract 9
1. Introduction 11
1.1 Representations of -Omic Information 13
1.2 Translatomic Complexity 16
1.3 Unveiling the Landscape of Translatomic Complexity 20
1.4 Thesis Aims and Significance 23
1.5 References 25
2. Ribosome decision graphs for the representation of eukaryotic RNA translation complexity. 37
2.1 Foreword 38
2.2 Abstract 39
2.3 Nascent need for abstract representation of mRNA decoding complexity 40
2.4 Representation of the complexity of mRNA translation using Open Reading Frames leads to ambiguity 42
2.5 RNA translation is segmented 46
2.6 Ribosome Paths 49
2.7 Ribosome Decision Graphs 52
2.8 Implementations of RDGs 54
2.9 Conclusion 61
2.10 References 62
2.11 Supplemental Material 70
2.11.1 Glossary of Concepts 70
2.11.2 RDG-Viewer 71
2.11.3 Complex Examples Visualised 72
2.11.4 Annotation Schema 73
2.11.5 Methods 74
2.11.6 Supplemental References 75
3. RDG: A Python package for the visualisation and analysis of translation complexity with Ribosome Decision Graphs 76
3.1 Foreword 77
3.2 Abstract 78
3.3 Introduction 79
3.4 Methods 81
3.4.1 RDG Class 82
3.4.2 RDG Construct Functionality 83
3.4.3 RDG Visualise Functionality 85

3.4.4 RDG Read and Write Functionality....85
3.4.5 RDG Prune Functionality....86
3.4.6 Installation and Accessibility....87
3.5 Discussion....88
3.6 References....89
4. Ribo-Seq Data Portal: Empowering Large-Scale Translation Reanalysis through Access to Pre-processed Samples and Standardised Metadata....91
4.1 Foreword....92
4.2 Abstract....93
4.3 Background....94
4.4 Results....98
4.4.1 Ribo-Seq Data Portal - A Platform for the Exploration of Ribosome Profiling Datasets 98
4.4.2 Programmatic Access Facilitates Large-Scale Analysis Projects....101
4.4.3 Ribo-Seq Data Portal enables Large-Scale Interrogation of Translation with RiboSeq.Org 101
4.5 Discussion....104
4.6 Conclusions....107
4.7 Methods....108
4.7.1 Ribo-Seq Study and Sample Identification....108
4.7.2 Metadata Curation....108
4.7.3 File Processing....109
4.7.4 Data Portal Development....110
4.8 Author contributions....111
4.9 References....112
5. Metric-Based Analysis of Ribosome Profiling Data Quality with RiboMetric....118
5.1 Foreword....119
5.2 Abstract....120
5.3 Introduction....121
5.4 Materials and Methods....127
5.4.1 Datasets....127
5.4.2 RiboMetric Implementation....129
5.4.3 Quality Control Metrics....129
Periodicity....129
Codon Coverage Uniformity of Metagenes....132
Terminal Nucleotide Bias....134
Read Length Distribution....135
5.4.4 Data Processing....137
5.4.5 Implementation and Reproducibility....137
5.4.6 Metric Decile Calculation....138
5.5 Results....139
5.5.1 RiboMetric....139

5.5.2 Assessment of Metrics for Data Properties....140
Periodicity....140
Uniformity....144
Terminal Nucleotide Bias....151
Read Length Distribution....152
5.5.3 Relationships Between Metrics....155
5.6 Discussion....158
5.6.1 Overview of RiboMetric's Contributions....158
5.6.2 Interpretation of Metric Performance....159
5.6.3 Potential for Facilitating large-scale Analysis....162
5.7 Conclusion....163
5.8 References....164
6. Discussion....169
6.1 Advances in Translatomics....169
6.2 References....176
7. Supplementary Manuscript....177
7.1 Translation complex profile sequencing allows discrimination of leaky scanning and reinitiation in upstream open reading frame-controlled translation....178
7.1.1 Abstract....179
7.1.2 Introduction....179
7.1.3 Materials and Methods....183
7.1.4 Results....184
7.1.5 Conclusion....193
7.1.6 References....194

Acknowledgements

I would like to express my deepest gratitude to all those who have supported me throughout my PhD. This work would not have been possible without their guidance, encouragement, and support.

First and foremost, I extend my sincere thanks to Professor Pasha Baranov for his patience, invaluable guidance, challenging discussions, and inspiring mentorship throughout my time as his PhD student.

I am deeply grateful to the LAPTI lab, the RiboSeq.Org team and the School of Biochemistry and Cell Biology for creating a stimulating and collaborative research environment. Particularly, Dr. Audrey Michel for her support and encouragement in the early days of my academic career. I may not have even begun this PhD without her advice and guidance. Special thanks to Dr. Anmol Kiran and Dr. Alla Fedorova for their bioinformatic support, insightful discussions, and camaraderie. As the senior bioinformaticians in the group throughout my PhD they both provided significant inspiration as to how to conduct thorough projects in an efficient manner.

My heartfelt appreciation goes to my two closest collaborators, Michal Świrski and Dr. Håkon Tjeldnes, whose contributions and expertise have greatly enriched this work.

I would like to thank Dr. Jon Mudge and the team at the European Bioinformatics Institute for hosting me during my external internship in Cambridge, UK. The experience and knowledge gained during this placement and our ongoing collaborations have been invaluable to my development as a researcher.

To my partner, Aisling O'Neill, thank you for your unwavering support, patience, and understanding throughout this challenging journey. Your ability to bring balance to my life, even during the most intense periods of work, has been invaluable. The achievement of completing this PhD is as much yours as it is mine.

Finally, I wish to express my profound gratitude to my parents, Martha and John, as well as my family and friends for their endless support and belief in me. Your encouragement and genuine interest in my research, despite its complexity, has been a great source of inspiration.

To all those who have contributed to my growth and success during this PhD, whether named here or not, I offer my sincere thanks.

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Digital signature of the candidate:



Abstract

The advent of the ribosome profiling (Ribo-Seq) sequencing technique has led to the genome-wide characterisation of translatomic complexity across the tree of life. The understanding of the extent to which translation occurs outside of annotated protein coding regions has been greatly expanded, as has the understanding of the role of translation dynamics in gene expression. As more and more Ribo-Seq data is made available and consequently more translated regions identified, the need for resources and methods for annotating, representing and analysing this translation grows. Current representation schemas all but ignore translational complexity and where it is accounted for, the reality of the gene's expression is often further obscured. The Ribo-Seq data used to uncover this translational complexity is often generated for independent investigations prior to its deposition in public depositories. The re-use of this data is essential for the annotation of translational complexity yet the need for elaborate data processing and lack of standardised metadata hinder this reuse at scale.

In this thesis, we introduce a set of interoperable frameworks and resources for the representation of eukaryotic translational complexity, and the increased accessibility of high-quality Ribo-Seq data in order to facilitate further investigations into translational complexity.

First, introduce the concept of a Ribosome Decision Graph (RDG) for the accurate representation of translatomic complexity. RDGs address the shortcomings of current annotation frameworks by representing translation events as branching points in a directed acyclic graph enabling the representation of the set of potential paths a ribosome may take as it traverses the mRNA. In order to facilitate the adoption of this concept, we proceed to introduce a comprehensive Python package that enables researchers to construct, visualise and analyse these graphs.

To address the barriers to large-scale reanalysis of public Ribo-Seq data, we introduce the RiboSeq Data Portal. This resource makes 14,840 Ribo-Seq samples from 969 studies across 96 different species/strains. This resource serves as a database of standardised metadata for each of these samples and also offers pre-processed data in various formats. This portal’s

integration with the existing resources maintained by RiboSeq.Org enhances the exploration of Ribo-Seq data within the browser.

We also introduce RiboMetric a comprehensive command line application for the assessment of Ribo-Seq dataset properties. RiboMetric generates detailed sample reports as well as a set of metrics that quantify important Ribo-Seq data properties including, periodicity, uniformity and read length distribution. In combination with the data available in the RiboSeq Data Portal, these metrics enable large-scale comparisons of datasets enhancing the reproducibility of translatomic research.

Collectively, these advances provide robust foundations for future research into translatomic complexity in particular through the reanalysis of Ribo-Seq data. RDGs provide new perspectives on how translatomic can be represented and analysed while the RiboSeq Data Portal and RiboMetric make large-scale reanalysis of Ribo-Seq datasets with desired properties more accessible to the entire translatomic research community. This work collectively lays the groundwork for more comprehensive and reproducible investigations into translatomic complexity, potentially revolutionising our understanding of translation dynamics in gene regulation.

Lay Abstract

Our cells are like bustling cities, constantly building and maintaining structures (proteins) essential for maintaining life. The set of flexible building guidelines for the entire city are stored in the ‘genome’ of the cell that is written in DNA sequence. These guidelines aren't strict blueprints, but rather adaptable plans that can be interpreted differently based on the city's needs.

Genes are specific sections of these guidelines, each describing how to build particular structures. When a structure is needed, copies of the relevant guidelines are transcribed into ‘messenger RNA’ sequence and distributed to the builders. The builders in our cellular cities are ribosomes, which read these copied guidelines and translate them into the required structure. Multiple ribosomes can translate one set of instructions at a time and when a cell needs a particular protein they can make hundreds of copies of these instructions.

In 2009 researchers from University of California, Berkeley developed a powerful technique called ‘ribosome profiling’. This is the equivalent of taking a snapshot of every construction site in the city simultaneously, showing exactly what stage each project is at, how many ‘builders’ are working on it and what stage of the instructions each ‘builder’ is at. This citywide survey revealed surprising complexities in how our cellular cities operate.

Over the following years researchers went on to discover that some areas of the guidelines originally thought to be just notes or unused sections actually contain instructions for important structures. Even more surprisingly, we learned that it is not uncommon for ‘builders’ to interpret the guidelines differently than we expected, leading to a variety of structures from the same set of instructions.

Ribosome profiling became popular because it gives us an important perspective into how our cellular cities function. Scientists share their citywide surveys, allowing others to verify their findings and use the data in new ways. This sharing helps us understand how our cellular cities work on a massive scale.

However, we lacked good ways to map out these complex building processes, making it hard for scientists to share and understand this information. In order to address these shortcomings we:

1. Developed Ribosome Decision Graphs, a new way to map how building guidelines are commonly interpreted by these ‘builders’.
2. Created a computer tool to help scientists easily make and use these new maps.
3. Built a website that collects and organises data from thousands of citywide surveys, showing how building guidelines are being used across many cellular cities under different conditions.
4. Developed a tool to check the quality of this data, helping scientists choose the most reliable information.

These tools will help the research community better understand how our cellular cities build and maintain themselves as well as respond to changes in their environment. This will lead to new insights into diseases as it helps explain these diseases from the ribosomes perspective - which originate from problems in the city's operations - and potentially help develop new treatments. By making it easier to study our cellular cities, we hope to speed up research that could improve human health.

1. Introduction

Nobel laureate Renato Dulbecco first proposed the concept of sequencing an entire genome in a 1986 perspective in Science, where he suggested it as a potential turning point in cancer research (1). In this piece, he illustrates how complexity reduction by using various model systems was commonly employed to make investigations into the mechanisms of cancer more tractable to great effect. However, these reduced complexity systems limit potential insights into how oncogene expression is linked to the progression of a cancer towards malignancy. To address this gap, Dulbecco proposed the sequencing of an entire human genome in order to aid the identification and classification of the genes involved in cancer progression. Concluding this perspective with the following:

"The next generation can look forward to exciting new tasks that may lead to a completion of our knowledge about cancer, closing one of the most challenging chapters in biological research." (1)

This hopeful forecast was echoed, and further amplified, by scientists and politicians alike as the Human Genome Project (HGP) progressed through its various tumultuations, overlooking many known complexities of biology. The hype induced at the time in order to justify the budget assigned to the project was not unfounded, and the majority of the potential outcomes of the work that have not yet been achieved are still in the works. Frustrations around the relatively slow progress towards the revolutionised "treatment of most, if not all, human diseases" — as advertised in U.S President Bill Clinton's announcement of the completion of the project — can be attributed to lack of weight attributed to known biological complexity when generating this hype.

Complexity and an unwillingness of observations to adhere to generalisations are defining features of biological sciences. Consequently, there is considerable variability in the appropriate degrees of complexity that should be accounted for in investigations of biological processes. This leads to the generation of results with unclear context-specificity and scaling any generalisations made outside of these contexts may produce multitudes of exceptions. Nonetheless, it is through the study of reduced complexity systems such as model organisms that much of what we know about gene expression and regulation is known.

Post-Human Genome Project initiatives, including the 1000 Genomes Project and the Cancer Genome Atlas, have demonstrated the potential of transparent data generation and sharing procedures. These pioneering projects have shown how knowledge representation, data sharing, and comparative analysis can illuminate the complexities in the field of genomic variability. The application of the principles deduced from these projects to other -omics fields has the potential to greatly aid our ability to grapple with further complexities, produce more accurate generalisations and even leverage exceptions to fulfil the outcomes promised of the age of genomics.

1.1 Representations of -Omic Information

Hans Winkler, a German botanist, coined the term ‘genome’ in 1920 to refer to the haploid chromosome set of a given species (2). The Oxford English Dictionary provides an etymology suggesting that genome is a portmanteau of ‘gene’ and ‘chromosome’, however, this is disputed. Prior to the coining of ‘genome’ the ‘-ome’ suffix had been in extensive use in terms such as ‘rhizome’ and ‘biome’ to refer to the complete set of whatever is in the stem of the word. The term ‘genome’ refers to the complete set of the genetic material. Throughout the late 20th and early 21st centuries, the ‘-ome’ suffix quickly became ingrained in the scientific vernacular to represent the comprehensive set of any given biological entity. The transcriptome is the comprehensive set of RNA transcripts, proteome of proteins, translatome of translated regions and so on.

Revolutions in sequencing, spectrometry and microscopy have resulted in an inundation of the scientific community with large amounts of data from a wide variety of -omics. This has enabled extensive characterisation of the ‘-omes’ of numerous species under even more numerous conditions (3–5). Even in a simplified system, the mode in which the comprehensive set of anything is represented has implications for how that set is utilised. Simplifications at the point of representation are often introduced to aid the storage and distribution of the collected data. However, as with the use of simplified systems for the study of cancer progression biological complexity is often lost in these simplifications.

The production of a haploid linear human reference genome transformed the field of genomics enabling major advances in our understanding of genetic variation and variant effects on phenotypic presentations. However, the inclusion of just one allele at each locus leads to the global adoption of a reference genome that poorly represents many of the populations for which it is used. The 1000 Genomes Project consortium revealed that a typical human genome differs from the reference genome at between 4.1 and 5 million loci yet the same linear reference genome was, until recently, used for all humans. This led to a significant bias towards discoveries pertaining to the variants represented in the reference (6,7).

Since the early 2000s, graph data structures have been ubiquitously used to represent DNA sequences in the process of genome assembly and multiple sequence alignment (8–10), and for genome comparison since the early 2010s (11). This laid the foundation for the representation of population-scale genomes with genome graphs to incorporate genomic complexity and address reference bias (12–14). The application of genome graphs gained considerable technical development and investment prior to the release of an official human reference genome pangenome (15). The human reference pangenome from the NHGRI provides a more comprehensive reference of genomic variation as it contains a diverse set of 47 phased diploid assemblies and has already been shown to improve variant calling (16).

Genome graphs, as novel representations of genomic complexity, are a significant step towards enabling the achievement of the prophesied outcomes of the Human Genome Project. The analysis, and therefore interpretation, of genomic data is no longer limited by constraints imposed by the simplified linear representation of genomic complexity.

Francis Crick's sequence hypothesis (1958), widely popularised by James Watson under the misnomer of "The Central Dogma", describes the standard information flow from DNA sequence to amino acid sequence (17,18). François Jacob and Jacques Monod proposed that this standard information flow was via intermediate messenger RNA transcripts (19) whose existence was jointly proven in experiments by both Jacob and Sydney Brenner, and François Gros in 1961 (20,21). The study of the comprehensive set of these transcripts in a given sample later became the focus of the field of transcriptomics.

However, it was not until the late 1970s and early 1980s that the true complexity of transcriptomics began to be uncovered with the discovery of alternative splicing, first in adenovirus and later, mammalian cells (22,23). Through alternative splicing, each eukaryotic gene has the potential to encode multiple different mature mRNA transcripts significantly increasing the diversity of the proteome and increasing transcriptomic complexity.

The advent of high throughput RNA Sequencing (RNA-Seq) has led to extensive characterisation of the unique splicing patterns of exonic regions exhibited by different cell lines and tissues under various conditions. Subsequent to the Human Genome Project discrete RNA isoforms composed of exons were annotated across the entire genome based on a

combination of mapped protein homology and Hidden Markov Models (24), followed by expressed sequence tags (25) and RNA-Seq data (26). Despite over two decades of both automated and manual annotation of the human genome, there are still many RNA transcripts that are detected in expression data that are not represented in annotation sets, however, long-read RNA-Seq methods have the potential to accurately characterise any transcriptome (27).

Characterising the entire transcriptome for each experimental condition is however intangible. Known RNA isoforms need to be represented and made available for use in downstream analysis. The primary mode by which this is achieved is either an annotation file that documents individual transcript models mapped to genome coordinates from which the transcript sequence can be obtained or via an individual sequence file containing RNA transcript sequences. However, delineating which RNA isoform is transcribed, especially from short-read data, is not trivial.

Splice graphs or splicing graphs, are commonly used in the identification of RNA isoforms and quantification of splice events from sequencing data regardless of read length (28–32). In the mid-2000s, splice graphs were also distributed as references of transcriptomic complexity through the Alternative Splicing Graph Server and TromER (still accessible at https://epd.expasy.org/tromer) (33,34). However, their popularity for this purpose decreased as transcriptomic complexity increased and the community gravitated towards simpler models as they are easier to interact with.

These data structures are similar to genome graphs in the sense that nodes representing tracts of sequences and walks along edges are potential combinations of these sequences. The combination of these two data structures in spliced pangenome graphs has recently been shown to bring the advances in population-level genomic references to transcript quantification in a haplotype-aware manner (35).

Genome graphs are conceptually more difficult to comprehend than linear reference genomes and in this case, comprehensive representations won out. Perhaps the same is destined for splice graphs.

1.2 Translatomic Complexity

The ultimate step in the aforementioned standard information flow from Francis Crick's sequence hypothesis is the translation of messenger RNA sequence into amino acid sequence. Direct evidence of this translation of RNA to protein, as well as the discovery of the first codon, was produced by Nirenberg and Matthaei (also in 1961) further validating the discoveries of the experiments of Brenner, Jacob and Gros (36,37). These findings provided proof that the ribosome, which had already been shown to be responsible for protein synthesis (38,39), decodes the sequence of RNA templates to produce peptides. The use of synthetic RNA by Nirenberg and Matthaei, combined with the determination of the triplet nature of the genetic code (40–42), proved instrumental in the deciphering of the genetic code in experiments by Har Gobind Khorana (43,44).

Following these foundational insights into principles of translation, a large number of experiments were published exploring the bounds of this process through which the significant differences between eukaryotic and prokaryotic translation initiation were observed by the research community and subsequently summarised by Marilyn Kozak (45). In this Review, Kozak proceeds to summarise the known characteristics of eukaryotic translation initiation and subsequently proposes a scanning mechanism of translation that accounts for the noted characteristics of initiation. In this initial scanning model, Kozak proposed that the 40S ribosomal subunit scans until the first AUG triplet on the mRNA sequence where the 60S subunit is recruited and translation is initiated. Although she later noted that this may be (46), and in fact was (47), too dogmatic it did have a lasting effect on the popular perception of the rules of translation initiation which persists today.

The scanning model for translation was subsequently updated to account for a number of exceptions (47). Should the AUG be either too proximal to the 5’ cap or be in an unfavourable context for initiation it may not be recognised by the scanning complex, which she termed ‘leaky scanning’. Finally, following the translation of a short open reading frame (ORF), scanning towards the 3` end is resumed permitting reinitiation at the next AUG. Remarkably, this updated scanning model still accounts for the vast majority of translatomic complexity

known to date if we permit the use of non-AUG initiation codons and a broader definition of favourable context to what was initially described.

Translation initiated at non-AUG codons has been widely documented since before the publication of the updated scanning model (48–52) and was even already linked to specific pathologies (53). This phenomenon has continued to be explored and the discoveries have been periodically reviewed (54–56). In brief, start codon selection by the scanning complex is not deterministic and is influenced by a number of factors. The canonical start codon, AUG, is preferentially initiated at in less favourable context than non-AUG codons due to stronger codon/anticodon with $Met - tRNA_i^{Met}$. The favourability of a given codon's context is dynamic and is documented to vary depending on the composition of the preinitiation complex (57). Kozak initially referred to the identity of the nucleotides in the immediate flanking sequences as the primary determinants of favorability (45); however, she later exhaustively described the contribution of secondary structure to this context (58).

In light of the scanning model for translation and diversity of the factors influencing start codon selection, it is not surprising that the degree of translatomic complexity rivals, if not exceeds, that of the transcriptomic and genomic complexity previously described above. This, however, covers just one of three stages of the process of translation, albeit the one that contributes most to translatomic complexity.

Once translation is initiated, the fidelity at which the elongating ribosome decodes the RNA sequence through to the termination site is remarkable. Mechanisms for which were recently summarised by Marina Rodnina (59). Thus making the rare instances where the ribosome's strict rules for RNA decoding are programmatically subverted to produce alternative proteoforms, termed 'Recoding', all the more extraordinary.

Numerous different classes of recoding event have been described to date. Examples include; programmed ribosomal frameshifting, stop codon readthrough, and translational bypassing. Programmed ribosomal frameshifting was first discovered in viruses (60), where it is commonly utilised to coordinate the expression of essential viral proteins, but has since been identified in a diverse range of organisms (61–64). Here, the rule of triplet decoding is broken as the ribosome slips either forward or backwards to change the reading frame and continue

decoding until the termination codon is reached. In the case of stop codon readthrough, it is the termination at stop codon rule that is broken. Rather than altering the reading frame, stop codons are decoded as sense codons extending the produced proteoform at its C-terminus. These stop codons are either decoded by near-cognate aminoacyl-tRNAs or specialised cognate tRNAs that often encode non-canonical amino acids selenocysteine or pyrrolysine (65). Finally, whole tracts of mRNA can be skipped by an elongating ribosome prior to resumed translation downstream leading to the production of a protein from discontinuous segments of RNA in a process called translational bypassing (66). It is postulated that at least one example of these recoding events will occur in every organism (67,68).

Although the principles by which translatomic complexity comes about were described in the early days of molecular biology, the ubiquity of recoding events and pervasiveness of alternative initiation were not fully understood until the availability of genomic sequences and, ultimately, high throughput sequencing data increased. As a result, the development of comprehensive schemas for the representation of translatomic complexity lags considerably in comparison with representations of genomic and transcriptomic complexity. To facilitate the representation of translatomic complexity, canonical representation schemas often resort to the corruption of the representation of transcriptional complexity. In order to facilitate the representation of alternative initiation or termination sites on one transcript, the transcript model is duplicated with different coding sequence coordinates. Even more egregious, the representation of programmed ribosomal frameshifting involves the creation of a faux-intron enabling the encoding of the trans-frame product.

The lack of accurate representations of the knowledge obtained from decades of experimentation significantly hinders the progression of scientific discovery. This lack particularly restricts large-scale bioinformatic investigations as the potential for computational interface with well known phenomena is limited. The development of accurate and interoperable representations of all -omic information types, including translatomic, in concert with diverse, high quality and freely available data are essential in order to maximise scientific progress.

In Chapter 2 of this thesis, we introduce a novel concept for the accurate representation of translatomic complexity and subsequently, in Chapter 3, a Python implementation of this concept.

1.3 Unveiling the Landscape of Translatomic Complexity

Perspectives of global translation have been obtainable since 1963 when brothers Hans and Markus Noll introduced polysome profiling via sucrose gradient (69). The advent of this technique, along with many further fascinating insights from the previously alluded-to discovery of the mechanisms of protein synthesis, are detailed in an essay by Hans Noll from 2008 (70). This method posits that mRNAs with more ribosomes bound will be heavier than those with fewer ribosomes. The quantities of mRNA in complex with varying numbers of translating ribosomes (termed polysomes) can be approximated from the height of peaks in the absorbance across the gradient. Northern blot-type approaches were used to identify the presence or absence of a specific mRNA isoform in a given fraction (71). Site-specific RNA cleavage followed by Northern blotting enabled the identification of specific RNA fragments in polysomes (72). However, the combination of polysome fractionation with DNA microarray technology provided the first insights into translation at the genome wide level providing insights into the scale at which translational regulation occurs (73–76). However, even though high throughput sequencing-by-synthesis was being applied to RNA since 2006 (77), polysomes were not sequenced at a genome-wide level until the mid-2010s (78,79).

The observation that the ribosome protects the segment of mRNA to which it is bound from nuclease digestion is credited to Takanami, Yan and Jukes (80,81). Four years later these ~30nt long ribosome-protected fragments (RPFs) were first sequenced by Joan Steitz providing seminal insights into the translatome of bacteriophage R17 (81). This property of RNA protection was later used to make further seminal discoveries regarding translation dynamics (82,83). However, it wasn't until 2009 when Ingolia et al described the high throughput sequencing of these RPFs, termed ribosome profiling or Ribo-Seq, that the genome wide identification of translating ribosome positions was achieved for large genomes (84).

Specifics of ribosome profiling and its various derivatives are described in further detail in chapters 4 and 5 of this thesis. In brief, Ribo-Seq typically involves halting ribosomes at a specific stage of translation (eg. initiation or elongation), followed by nuclease digestion and sequencing of the resulting RPFs. One could argue that Joan Steitz conducted the first

initiating ribosome profiling experiment 40 years before the term was even coined. She conducted genome wide profiling of RNA fragments protected from nuclease digestion by initiating ribosomes. This differs from the modern approach of conducting Ingolias Ribo-Seq using harringtonine or lactimidomycin followed by Illumina sequencing only in terms of the mode of enrichment for initiating ribosomes and sequencing technology. Admittedly Steitzs Ribo-Seq would not scale nearly as well as Ingolias.

The advent of ribosome profiling revolutionised the study of translatomic complexity as it uncovered pervasive translation outside of annotated CoDing Sequences (CDSs) (84). This revelation in combination with the apparent applicability of this technique led to a spate of studies that looked to further characterise the extent of this translatomic complexity. In 2011, Ingolia et al followed up with further evidence of the translation in the 5` leader in the form of translated upstream open reading frames (ORFs) and N-terminal extensions/truncations in mouse embryonic stem cells (85). Subsequent pioneering work expanded on these findings by applying this technique in humans (86–90), zebrafish (91,92), drosophila (79) and bacteria (93–95). Novel translated regions were uncovered in various geographies relative to the CDS (86,96–99) as well as in supposedly non-coding RNAs (79,100–104). The pervasiveness of this translation was highlighted in numerous cell lines/tissues (105–107) and uncovering a variety of recoding examples (108,109). Various Ribo-Seq datasets were subsequently re-used for calling translated regions (110–114) ultimately leading to the initiation of a community-wide effort to produce a reference set of annotated translated regions referred to as Ribo-Seq ORFs (115).

The full potential of Ribo-Seq, and its derivatives (116–119), comes through the combined reanalysis of existing data. The genome-wide nature of the data lends itself to incidental findings outside of the focus of the experimental design highlighting the importance of data sharing. As ribosome profiling gained popularity and the amount of available data increased, the need for resources that compile, process, aggregate and make datasets available became apparent.

RiboSeq.Org is a collection of resources that have been actively developed since 2012 to aid the study of translatomic complexity by serving processed data for visualisation and analysis in the web browser (120–123). In Chapter 4 we introduce the newest addition to RiboSeq.Org,

the RiboSeq Data Portal (RDP), that aims to further facilitate the standardisation of Ribo-Seq data accessibility. Through the implementation of metadata standardisation steps and provision of pre-processed Ribo-Seq datasets, the RDP aims to lower the barrier to large-scale re-analysis projects and ultimately help comprehensively characterise translatomic complexity.

There is considerable variability in the properties of Ribo-Seq data across the substantial body of available datasets. These properties, including the characteristic 3nt periodicity observed in ribosome profiles, read length distribution and profile uniformity have significant impacts on the suitability of the dataset for use in reanalysis projects. However, the systematic characterisation of these properties at scale served as a limiting factor in large-scale reanalysis efforts. In Chapter 5, we introduce RiboMetric, a command line application for the standardised assessment of these properties. RiboMetric serves as the first metric based method for the quantitative assessment of key features of Ribo-Seq data, facilitating informed dataset selection and integration.

Large strides have been made to date in the field of translatomics yet a number of obstacles still hinder further understanding of translatomic complexity. The introduction to this thesis advocates for two major advances, i) accurate representations of translatomic complexity and ii) readily accessible Ribo-Seq data, towards which the succeeding chapters will strive to address. By addressing these needs we strive to set the foundation for the comprehensive modelling of the process of protein synthesis and its regulation, hopefully contributing to further advances across the fields of bioinformatics, genetics and molecular biology.

1.4 Thesis Aims and Significance

Large strides have been made to date in the field of translatomics however a number of obstacles still hinder further understanding of translatomic complexity. The introduction to this thesis advocates for two major advances:

1. The development of accurate representations of translatomic complexity
2. The improved accessibility and standardisation of translatomic (Ribo-Seq) data

The primary objectives of the research in the following chapters are:

1. To introduce a novel framework for the representation of translatomic complexity that can accurately capture the dynamics of translation events and enable the elucidation of regulatory mechanisms.
2. To provide an implementation of this novel framework in a computational application that enables research via operations on these complex representations.
3. To facilitate community wide access to Ribo-Seq data via a web platform containing standardised metadata and preprocessed read files, lowering the barriers to large-scale reanalysis projects
4. To develop a method for assessing key properties of Ribo-Seq datasets, enabling more informed dataset selection and integration in translatomic studies.

By addressing these objectives we strive to set the foundation for the comprehensive modelling of the process of protein synthesis and its regulation, with the potential to have the following significance:

1. Translatomics will have an accurate representation of known complexity enabling knowledge sharing and reducing the frequency of misinterpretation of translational complexity originating from poor representations.

2. The provision of a flexible implementation of the novel data structure will facilitate the adoption of these concepts throughout the translatomic and wider omics research communities.

3. Increasing accessibility of translatomic data and associated metadata in standardised formats enables large-scale translatomic research reusing the wealth of data available to the community.

The combination of these major advances to produce accurate representations of translatomic complexity informed by accessible Ribo-Seq datasets has the potential to induce a paradigm shift in the study of translational complexity. In the following chapters we detail our work on these goals and discuss specific details and potential applications throughout.

1.5 References

1. Dulbecco R. A Turning Point in Cancer Research: Sequencing the Human Genome. Science. 1986 Mar 7;231(4742):1055–6.

2. Winkler H. Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche. Jena: G. Fischer; 1920. doi Available from: https://doi.org/10.5962/bhl.title.1460

3. Uhlén M, Björling E, Agaton C, Szigyarto CAK, Amini B, Andersen E, et al. A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics. Mol Cell Proteomics. 2005 Dec 1;4(12):1920–32.

4. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science. 2015 May 8;348(6235):660–5.

5. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007 Jan;35(Database issue):D521-526.

6. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinforma Oxf Engl. 2009 Dec 15;25(24):3207–12.

7. Brandt DYC, Aguiar VRC, Bitarello BD, Nunes K, Goudet J, Meyer D. Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. G3 Bethesda Md. 2015 Mar 17;5(5):931–41.

8. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci U S A. 2001 Aug 14;98(17):9748–53.

9. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 May;18(5):821–9.

10. Raphael B, Zhi D, Tang H, Pevzner P. A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res. 2004 Nov;14(11):2336–46.

11. Paten B, Diekhans M, Earl D, John JSt, Ma J, Suh B, et al. Cactus Graphs for Genome Comparisons. J Comput Biol. 2011 Mar 1;18(3):469–81.

12. Dilthey A, Cox C, Iqbal Z, Nelson MR, McVean G. Improved genome inference in the MHC using a population reference graph. Nat Genet. 2015 Jun;47(6):682–8.

13. Novak AM, Hickey G, Garrison E, Blum S, Connelly A, Dilthey A, et al. Genome Graphs. bioRxiv. 2017 Jan 18;101378.

14. Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018 Oct;36(9):875–9.

15. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023 May;617(7960):312–24.

16. Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, et al. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet. 2023 Jul;24(7):464–83.

17. Crick FH. On protein synthesis. Symp Soc Exp Biol. 1958;12:138–63.

18. Crick F. Central Dogma of Molecular Biology. Nature. 1970 Aug;227(5258):561–3.

19. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol. 1961 Jun;3:318–56.

20. Gros F, Hiatt H, Gilbert W, Kurland CG, Risebrough RW, Watson JD. Unstable Ribonucleic Acid Revealed by Pulse Labelling of Escherichia Coli. Nature. 1961 May;190(4776):581–5.

21. Brenner S, Jacob F, Meselson M. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. Nature. 1961 May;190(4776):576–81.

22. Berget SM, Moore C, Sharp PA. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA*. Proc Natl Acad Sci. 1977 Aug;74(8):3171–5.

23. Padgett RA, Grabowski PJ, Konarska MM, Seiler S, Sharp PA. Splicing of messenger RNA precursors. Annu Rev Biochem. 1986;55:1119–50.

24. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, et al. The Ensembl genome database project. Nucleic Acids Res. 2002 Jan 1;30(1):38–41.

25. Eyras E, Caccamo M, Curwen V, Clamp M. ESTGenes: Alternative Splicing From ESTs in Ensembl. Genome Res. 2004 May 1;14(5):976–87.

26. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, et al. Ensembl 2011. Nucleic Acids Res. 2011 Jan;39(Database issue):D800–6.

27. Pardo-Palacios FJ, Wang D, Reese F, Diekhans M, Carbonell-Sala S, Williams B, et al. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. Nat Methods. 2024 Jun 7;

28. Su Y, Yu Z, Jin S, Ai Z, Yuan R, Chen X, et al. Comprehensive assessment of mRNA isoform detection methods for long-read sequencing data. Nat Commun. 2024 May 10;15(1):3972.

29. Goldstein LD, Cao Y, Pau G, Lawrence M, Wu TD, Seshagiri S, et al. Prediction and Quantification of Splice Events from RNA-Seq Data. PloS One. 2016;11(5):e0156132.

30. Baharlou Houreh M, Ghorbani Kalkhajeh P, Niazi A, Ebrahimi F, Ebrahimie E. SpliceDetector: a software for detection of alternative splicing events in human and model organisms directly from transcript IDs. Sci Rep. 2018 Mar 22;8(1):5063.

31. LeGault LH, Dewey CN. Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs. Bioinforma Oxf Engl. 2013 Sep 15;29(18):2300–10.

32. Ma C, Zheng H, Kingsford C. Exact transcript quantification over splice graphs. Algorithms Mol Biol. 2021 May 10;16(1):5.

33. Bollina D, Lee BTK, Tan TW, Ranganathan S. ASGS: an alternative splicing graph web service. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W444-447.

34. Sperisen P, Iseli C, Pagni M, Stevenson BJ, Bucher P, Jongeneel CV. trome, trEST and

trGEN: databases of predicted protein sequences. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D509–11.

35. Sibbesen JA, Eizenga JM, Novak AM, Sirén J, Chang X, Garrison E, et al. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. Nat Methods. 2023 Feb;20(2):239–47.

36. Nirenberg MW, Matthaei JH. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci U S A. 1961 Oct 15;47(10):1588–602.

37. Matthaei H, Nirenberg MW. The dependence of cell-free protein synthesis in E. coli upon RNA prepared from ribosomes. Biochem Biophys Res Commun. 1961 Apr 28;4:404–8.

38. Tissières A, Watson JD, Schlessinger D, Hollingworth BR. Ribonucleoprotein particles from *Escherichia coli*. J Mol Biol. 1959 Sep 1;1(3):221–33.

39. Zamecnik PC, Keller EB. Relation between phosphate energy donors and incorporation of labeled amino acids into proteins. J Biol Chem. 1954 Jul;209(1):337–54.

40. Nirenberg M, Leder P. RNA Codewords and Protein Synthesis. Science. 1964 Sep 25;145(3639):1399–407.

41. Leder P, Nirenberg M. Rna codewords and protein synthesis, ii. nucleotide sequence of a valine rna codeword. Proc Natl Acad Sci. 1964 Aug;52(2):420–7.

42. Leder P, Nirenberg MW. Rna codewords and protein synthesis, iii. on the nucleotide sequence of a cysteine and a leucine rna codeword. Proc Natl Acad Sci. 1964 Dec;52(6):1521–9.

43. Nirenberg M, Caskey T, Marshall R, Brimacombe R, Kellogg D, Doctor B, et al. The RNA Code and Protein Synthesis. Cold Spring Harb Symp Quant Biol. 1966 Jan 1;31(0):11–24.

44. Söll D, Ohtsuka E, Jones DS, Lohrmann R, Hayatsu H, Nishimura S, et al. Studies on

polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. Proc Natl Acad Sci U S A. 1965 Nov;54(5):1378–85.

45. Kozak M. How do eucaryotic ribosomes select initiation regions in messenger RNA? Cell. 1978 Dec 1;15(4):1109–23.

46. Kozak M. Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. Cell. 1980 Nov;22(1 Pt 1):7–8.

47. Kozak M. The scanning model for translation: an update. J Cell Biol. 1989 Feb;108(2):229–41.

48. Peabody DS. Translation initiation at an ACG triplet in mammalian cells. J Biol Chem. 1987 Aug 25;262(24):11847–51.

49. Peabody DS. Translation initiation at non-AUG triplets in mammalian cells. J Biol Chem. 1989 Mar 25;264(9):5031–5.

50. Kozak M. Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. Mol Cell Biol. 1989 Nov;9(11):5073–80.

51. Clements JM, Laz TM, Sherman F. Efficiency of translation initiation by non-AUG codons in Saccharomyces cerevisiae. Mol Cell Biol. 1988 Oct;8(10):4533–6.

52. Zitomer RS, Walthall DA, Rymond BC, Hollenberg CP. Saccharomyces cerevisiae ribosomes recognize non-AUG initiation codons. Mol Cell Biol. 1984 Jul;4(7):1191–7.

53. Hann SR, King MW, Bentley DL, Anderson CW, Eisenman RN. A non-AUG translational initiation in c-myc exon 1 generates an N-terminally distinct protein whose synthesis is disrupted in Burkitt's lymphomas. Cell. 1988 Jan 29;52(2):185–95.

54. Touriol C, Bornes S, Bonnal S, Audigier S, Prats H, Prats AC, et al. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. Biol Cell. 2003;95(3–4):169–78.

55. Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in

eukaryotes. Genes Dev. 2017 Sep 1;31(17):1717–31.

56. Andreev DE, Loughran G, Fedorova AD, Mikhaylova MS, Shatsky IN, Baranov PV. Non-AUG translation initiation in mammals. Genome Biol. 2022 May 9;23(1):111.

57. Hinnebusch AG. Molecular Mechanism of Scanning and Start Codon Selection in Eukaryotes. Microbiol Mol Biol Rev MMBR. 2011 Sep;75(3):434–67.

58. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene. 2005 Nov 21;361:13–37.

59. Rodnina MV. Decoding and Recoding of mRNA Sequences by the Ribosome. Annu Rev Biophys. 2023 May 9;52:161–82.

60. Jacks T, Varmus HE. Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. Science. 1985 Dec 13;230(4731):1237–42.

61. Matsufuji S, Matsufuji T, Miyazaki Y, Murakami Y, Atkins JF, Gesteland RF, et al. Autoregulatory frameshifting in decoding mammalian ornithine decarboxylase antizyme. Cell. 1995 Jan 13;80(1):51–60.

62. Clark MB, Jänicke M, Gottesbühren U, Kleffmann T, Legge M, Poole ES, et al. Mammalian Gene *PEG10* Expresses Two Reading Frames by High Efficiency –1 Frameshifting in Embryonic-associated Tissues*. J Biol Chem. 2007 Dec 28;282(52):37359–69.

63. Lobanov AV, Heaphy SM, Turanov AA, Gerashchenko MV, Pucciarelli S, Devaraj RR, et al. Position-dependent termination and widespread obligatory frameshifting in Euplotes translation. Nat Struct Mol Biol. 2017 Jan;24(1):61–8.

64. Craigen WJ, Caskey CT. Expression of peptide chain release factor 2 requires high-efficiency frameshift. Nature. 1986 Jul 17;322(6076):273–5.

65. Zhang Y, Baranov PV, Atkins JF, Gladyshev VN. Pyrrolysine and Selenocysteine Use Dissimilar Decoding Strategies *. J Biol Chem. 2005 May 27;280(21):20740–51.

66. Huang WM, Ao SZ, Casjens S, Orlandi R, Zeikus R, Weiss R, et al. A persistent

untranslated sequence within bacteriophage T4 DNA topoisomerase gene 60. Science. 1988 Feb 26;239(4843):1005–12.

67. Gesteland RF, Atkins JF. Recoding: dynamic reprogramming of translation. Annu Rev Biochem. 1996;65:741–68.

68. Baranov PV, Gesteland RF, Atkins JF. Recoding: translational bifurcations in gene expression. Gene. 2002 Mar 20;286(2):187–201.

69. Wettstein FO, Staehelin T, Noll H. Ribosomal Aggregate Engaged in Protein Synthesis: Characterization of the Ergosome. Nature. 1963 Feb;197(4866):430–5.

70. Noll H. The discovery of polyribosomes. BioEssays News Rev Mol Cell Dev Biol. 2008 Nov;30(11–12):1220–34.

71. Penn MD, Thireos G, Greer H. Temporal analysis of general control of amino acid biosynthesis in Saccharomyces cerevisiae: role of positive regulatory genes in initiation and maintenance of mRNA derepression. Mol Cell Biol. 1984 Mar;4(3):520–8.

72. Arava Y, Boas FE, Brown PO, Herschlag D. Dissecting eukaryotic translation and its control by ribosome density mapping. Nucleic Acids Res. 2005;33(8):2421–32.

73. Johannes G, Carter MS, Eisen MB, Brown PO, Sarnow P. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. Proc Natl Acad Sci U S A. 1999 Nov 9;96(23):13118–23.

74. Kuhn KM, DeRisi JL, Brown PO, Sarnow P. Global and specific translational regulation in the genomic response of Saccharomyces cerevisiae to a rapid transfer from a fermentable to a nonfermentable carbon source. Mol Cell Biol. 2001 Feb;21(3):916–27.

75. Zong Q, Schummer M, Hood L, Morris DR. Messenger RNA translation state: the second dimension of high-throughput expression screening. Proc Natl Acad Sci U S A. 1999 Sep 14;96(19):10632–6.

76. Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci. 2003

Apr;100(7):3889–94.

77. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC Genomics. 2006 Sep 29;7:246.

78. Floor SN, Doudna JA. Tunable protein synthesis by transcript isoforms in human cells. eLife. 2016 Jan 6;5:e10921.

79. Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, et al. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. eLife. 2014 Aug 21;3:e03528.

80. Takanami M, Yan Y, Jukes TH. Studies on the site of ribosomal binding of f2 bacteriophage RNA. J Mol Biol. 1965 Jul;12(3):761–73.

81. Steitz JA. Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. Nature. 1969 Dec;224(5223):957–64.

82. Hinnebusch AG. A hierarchy of trans-acting factors modulates translation of an activator of amino acid biosynthetic genes in Saccharomyces cerevisiae. Mol Cell Biol. 1985 Sep;5(9):2349–60.

83. Wolin SL, Walter P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. EMBO J. 1988 Nov;7(11):3559–69.

84. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009 Apr 10;324(5924):218–23.

85. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity of Mammalian Proteomes. Cell. 2011 Nov 11;147(4):789–802.

86. Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using

ribosomal footprinting. Genome Res. 2012 Nov 1;22(11):2208–18.

87. Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2012 Sep 1;109(37):E2424-32.

88. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016 Feb;13(2):165–70.

89. Raj A, Wang SH, Shim H, Harpak A, Li YI, Engelmann B, et al. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. Ingolia NT, editor. eLife. 2016 May 27;5:e13328.

90. Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN, Saghatelian A. Accurate annotation of human protein-coding small open reading frames. Nat Chem Biol. 2020 Apr;16(4):458–68.

91. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014 May 2;33(9):981–93.

92. Pauli A, Norris ML, Valen E, Chew GL, Gagnon JA, Zimmerman S, et al. Toddler: an embryonic signal that promotes cell movement via Apelin receptors. Science. 2014 Feb 14;343(6172):1248636.

93. Oh E, Becker AH, Sandikci A, Huber D, Chaba R, Gloge F, et al. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell. 2011 Dec 1;147(6):1295–308.

94. Mohammad F, Woolstenhulme CJ, Green R, Buskirk AR. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. Cell Rep. 2016 Feb 2;14(4):686–94.

95. Woolstenhulme CJ, Guydosh NR, Green R, Buskirk AR. High Precision Analysis of Translational Pausing by Ribosome Profiling in Bacteria Lacking EFP. Cell Rep. 2015

Apr 7;11(1):13–21.

96. Benitez-Cantos MS, Yordanova MM, O'Connor PBF, Zhdanov AV, Kovalchuk SI, Papkovsky DB, et al. Translation initiation downstream from annotated start codons in human mRNAs coevolves with the Kozak context. Genome Res. 2020 Jul;30(7):974–84.

97. Stern-Ginossar N, Weisburd B, Michalski A, Le VTK, Hein MY, Huang SX, et al. Decoding human cytomegalovirus. Science. 2012 Nov 23;338(6110):1088–93.

98. Fields AP, Rodriguez EH, Jovanovic M, Stern-Ginossar N, Haas BJ, Mertins P, et al. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. Mol Cell. 2015 Dec 3;60(5):816–27.

99. Andreev DE, O'Connor PBF, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, et al. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. eLife. 2015 Jan 26;4:e03971.

100. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. eLife. 2014 Sep 16;3:e03523.

101. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, et al. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Rep. 2014 Sep 11;8(5):1365–79.

102. Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Dev Camb Engl. 2013 Jul 1;140(13):2828–34.

103. Guttman M, Russell P, Ingolia NT, Weissman JS, Lander ES. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell. 2013 Jul 3;154(1):240–51.

104. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Sonenberg N, editor. eLife. 2015 Dec 19;4:e08890.

105. van Heesch S, Witte F, Schneider-Lunitz V, Schulz JF, Adami E, Faber AB, et al. The Translational Landscape of the Human Heart. Cell. 2019 Jun 27;178(1):242-260.e29.

106. Gonzalez C, Sims JS, Hornstein N, Mela A, Garcia F, Lei L, et al. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. J Neurosci Off J Soc Neurosci. 2014 Aug 13;34(33):10924–36.

107. Janich P, Arpat AB, Castelo-Szekely V, Lopes M, Gatfield D. Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames. Genome Res. 2015 Dec;25(12):1848–59.

108. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. eLife. 2013 Dec 3;2:e01179.

109. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012 Nov;22(11):2219–29.

110. Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D497–502.

111. Brunet MA, Lucier JF, Levesque M, Leblanc S, Jacques JF, Al-Saedi HRH, et al. OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. Nucleic Acids Res [Internet]. 2020 Nov 12 [cited 2020 Dec 11];(gkaa1036). Available from: https://doi.org/10.1093/nar/gkaa1036

112. Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. 2019 Jan 8;47(D1):D230–4.

113. Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, et al. A high-resolution map of human RNA translation. Mol Cell. 2022 Aug 4;82(15):2885-2899.e8.

114. Fedorova AD, Kiniry SJ, Andreev DE, Mudge JM, Baranov PV. Thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. Nat Commun. 2022 Dec 23;13(1):7910.

115. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022 Jul;40(7):994–9.

116. Archer SK, Shirokikh NE, Beilharz TH, Preiss T. Dynamics of ribosome scanning and recycling revealed by translation complex profiling. Nature. 2016 Jul;535(7613):570–4.

117. Gao X, Wan J, Liu B, Ma M, Shen B, Qian SB. Quantitative profiling of initiating ribosomes in vivo. Nat Methods. 2015 Feb;12(2):147–53.

118. Wagner S, Herrmannová A, Hronová V, Gunišová S, Sen ND, Hannan RD, et al. Selective Translation Complex Profiling Reveals Staged Initiation and Co-translational Assembly of Initiation Factor Complexes. Mol Cell. 2020 Aug 20;79(4):546-560.e7.

119. Zhao T, Chen YM, Li Y, Wang J, Chen S, Gao N, et al. Disome-seq reveals widespread ribosome collisions that promote cotranslational protein folding. Genome Biol. 2021 Jan 5;22(1):16.

120. Michel AM, Fox G, M. Kiran A, De Bo C, O'Connor PBF, Heaphy SM, et al. GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Res. 2014 Jan 1;42(D1):D859–64.

121. Michel AM, Kiniry SJ, O'Connor PBF, Mullan JP, Baranov PV. GWIPS-viz: 2018 update. Nucleic Acids Res. 2018 Jan 4;46(D1):D823–30.

122. Kiniry SJ, O'Connor PBF, Michel AM, Baranov PV. Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. Nucleic Acids Res. 2019 Jan 8;47(D1):D847–52.

123. Kiniry SJ, Judge CE, Michel AM, Baranov PV. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. Nucleic Acids Res. 2021 Jul 2;49(W1):W662–70.

2. Ribosome decision graphs for the representation of eukaryotic RNA translation complexity

Jack A. S. Tierney[1,2,#], Michał Świrski[3,#], Håkon Tjeldnes[1,4], Jonathan M. Mudge[5], Joanna Kufel[3], Nicola Whiffin[6], Eivind Valen[4,7], Pavel V. Baranov[1].

[1] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland.

[2] SFI Centre for Research Training in Genomics Data Science, University College Cork, Cork, Ireland

[3] Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland.

[4] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway.

[5] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK.

[6] The Big Data Institute and Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK.

[7] Department of Biosciences, University of Oslo, Oslo, Norway.

[#] These authors contributed equally to this work.

2.1 Foreword

In the introduction, we discuss how current representation schemas fail to accurately accommodate known translatomic complexity. In this chapter, we introduce Ribosome Decision Graphs (RDGs), a novel framework that fills this gap. These data structures not only support the representation of each translation event but also their relative relationships. This concept is the product of years of discussion starting when I joined the lab in 2018 for my undergraduate thesis project on the use of graph genomes in Ribo-Seq data analysis, which highlighted the need for more flexible representations in -omics studies. Through hours of conversation and countless iterations between myself, Pasha Baranov and later Michał Świrski of the University of Warsaw we settled on the concept of a Ribosome Decision Graph. By offering a more comprehensive and nuanced approach to depicting translational complexity, RDGs set the stage for improved analysis and understanding of gene expression regulation at the translational level.

2.2 Abstract

The application of ribosome profiling has revealed an unexpected abundance of translation in addition to that responsible for the synthesis of previously annotated protein-coding regions. Multiple short sequences have been found to be translated within single RNA molecules, both within annotated protein-coding and non-coding regions. The biological significance of this translation is a matter of intensive investigation. However, current schematic or annotation-based representations of mRNA translation generally do not account for the apparent multitude of translated regions within the same molecules. They also do not take into account the stochasticity of the process that allows alternative translations of the same RNA molecules by different ribosomes. There is a need for formal representations of mRNA complexity that would enable the analysis of quantitative information on translation and more accurate models for predicting the phenotypic effects of genetic variants affecting translation. To address this, we developed a conceptually novel abstraction that we term Ribosome Decision Graphs (RDGs). RDGs represent translation as multiple ribosome paths through untranslated and translated mRNA segments. We termed the latter 'translons'. Non-deterministic events, such as initiation, re-initiation, selenocysteine insertion or ribosomal frameshifting are then represented as branching points. This representation allows for an adequate representation of eukaryotic translation complexity and focuses on locations critical for translation regulation. We show how RDGs can be used for depicting translated regions, analysis of genetic variation and quantitative genome-wide data on translation for characterisation of regulatory modulators of translation.

2.3 Nascent need for abstract representation of mRNA decoding complexity.

Until relatively recently, available experimental evidence suggested that in eukaryotes each mRNA encoded only a single protein. Since only a single coding region was therefore expected to be translated this region was conventionally termed the CDS (CoDing Sequence). This view has been challenged by the development of the ribosome profiling technique which enables the isolation and sequencing of RNA fragments protected by ribosomes and hence detection of regions being translated (1). In essence, this technique is based on the capture of RNA fragments (footprints) within the ribosomes followed by their sequencing and mapping. Thus it provides information on what sequences are being translated while the densities of mapped footprints are indicative of the frequency with which ribosomes translate these sequences. Numerous ribosome profiling studies carried out in cells from a variety of eukaryotes unexpectedly revealed abundant translation outside of CDS regions. This included the translation of short sequences in the supposedly untranslated regions (UTRs) of mRNAs, as well as in so-called non-coding RNAs, especially long non-coding RNAs (lncRNAs) (2–11). These studies also demonstrated the translation of N-terminally extended CDS regions due to initiation at upstream non-AUG start codons (12) or C-terminally extended due to stop codon readthrough (13). A certain group of eukaryotic organisms (ciliates *Euplotes*) were found to utilize ribosomal frameshifting in thousands of their genes (14). While most of these phenomena were first described prior to the advent of ribosome profiling (15–19), they were considered rare. Certainly, very few cases have been catalogued by reference gene annotation projects, and no conventional abstraction has been developed to represent this translation complexity in annotations, schematic scientific diagrams or analytical workflows. The lack of a formal framework for the representation of this complexity hampers our ability to generate accurate and biologically realistic annotations of translated sequences and to design mathematical models and computer simulations. In its absence, it is difficult or even impossible to quantitatively characterize multiple translation events and define their interrelationships.

To address this challenge, we developed a conceptually novel framework for abstract representation of translation complexity, which we term Ribosome Decision Graphs (RDGs).

RDGs solve many problems, such as the representation of multiple translated regions in the same mRNAs and alternative decoding mechanisms producing multiple proteoforms. We show how RDGs can be used for the accurate depiction of productive and non-productive RNA translation (i.e. translation that does or does not lead to the production of a protein molecule), analysis of quantitative information on translation, and genetic variants affecting mRNA decoding.

2.4 Representation of the complexity of mRNA translation using Open Reading Frames leads to ambiguity.

The development of a conventional abstraction is undermined by the ambiguity of the terms used to define translated regions. For example, while translated regions are often described as Open Reading Frames (ORFs) in literature or scientific discourse, gene annotation projects typically utilise only the term CDS, and only for regions considered to be protein-coding. Instead, an ORF would be regarded by implication as a potential translation that can be identified *in silico*. Here, we in effect consider three concepts in an attempt at unification: 1) that ORFs can be identified *in silico* whether or not they have evidence of translation, 2) that ORFs may undergo translation that does not lead to the production of a stable, functional protein, and 3) that ORFs which are known to be translated into proteins should alone be considered CDSs. In other words, most CDSs are ORFs, but not all ORFs are CDSs. In general, there are two definitions of ORFs, start to stop (Start-Stop) and stop to stop (Stop-Stop) (20), as depicted in Figure 2.1A. Plotting the locations of potential start codons (usually AUGs) and stop codons in three reading frames is undoubtedly highly instrumental for examining potentially translated sequences. However, the common interpretation of nucleic acid sequences in terms of "translated ORFs" is superficial, frequently inaccurate, and often leads to confusion as illustrated in Figure 2.1B-D.

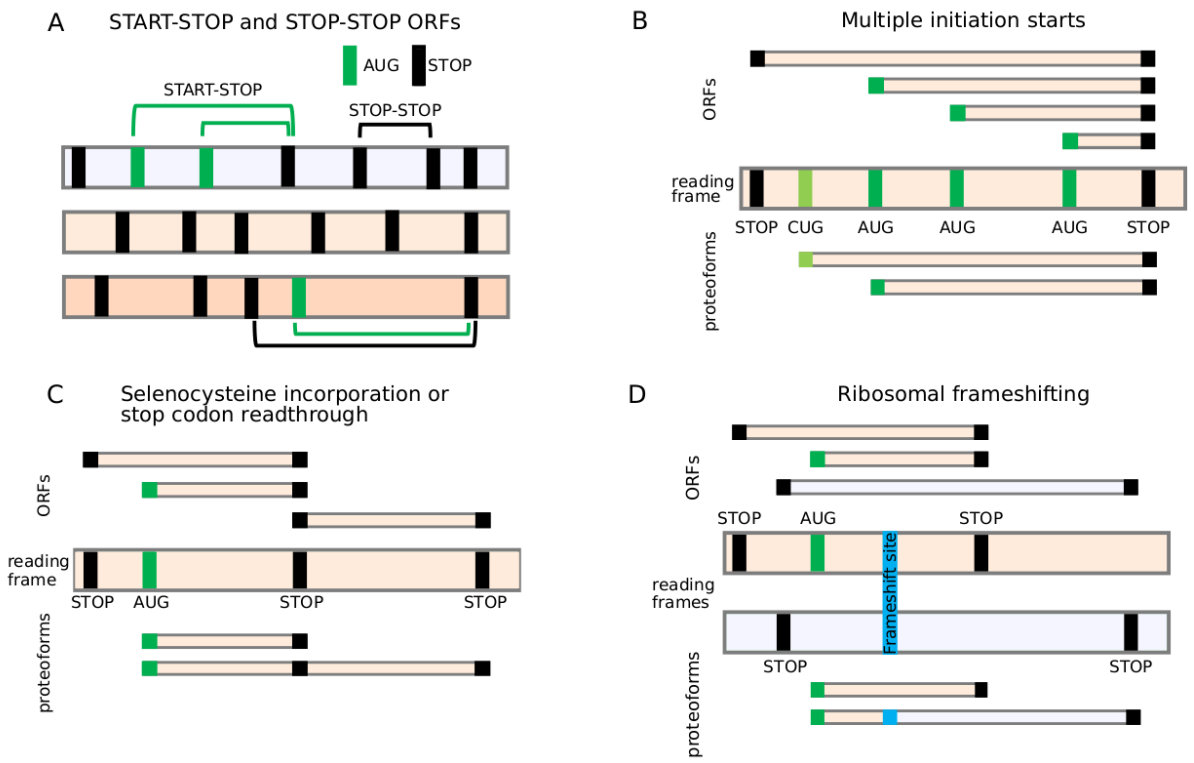


Figure 2.1. ORFs do not adequately represent translational complexity. A. Two formal definitions of ORFs. Three reading frames are shown as horizontal bars with vertical bars corresponding to AUG (green) or stop codons (black). Several examples of Start-Stop (green arcs) and Stop-Stop (black arcs) ORFs are shown. **B-D.** The relationship between ORFs (top) and expressed proteoforms (bottom) for mRNAs with different locations of starts and stops (middle). Only two relevant reading frames are shown for simplicity. **B.** An RNA encoding two proteoforms with alternative N-termini due to utilization of two start codons. Due to multiple potential AUG codons, there are many Start-Stop ORFs whose conceptual translation does not correspond to encoded proteoforms. Stop-Stop ORF does not reflect the existence of alternative proteoforms. **C.** In the case of stop codon readthrough or selenocysteine insertion, ribosomes may read through specific stop codons by incorporating an amino acid yielding a product that cannot be described as a product of a single ORF. **D.** Similarly, ribosomal frameshifting generates a trans-frame protein **(**blue) that does not represent a product of a single ORF.

Perhaps the most frequent source of alternative translation in many eukaryotes, including humans, is the multiplicity of translation initiation sites. It arises predominantly from two common mechanisms involved in the selection of translation initiation sites: leaky scanning and re-initiation. *Leaky scanning* refers to the inefficient recognition of a start codon by the

ribosome resulting in the ribosome scanning complex scanning through the start codon and effectively ignoring it (Kozak 2002). Generally, ribosome scanning complexes assemble at the 5’ cap of RNAs and move along the transcript in the 3’ direction until they encounter a start site and initiate translation (21–23). However, recognition of start sites is a sequence-dependent stochastic process, in which usually only a proportion of scanning complexes finally initiate. Many factors play a role in determining the efficiency with which a ribosome initiates translation at a given codon. These include the identity of the codon and its surrounding sequence (known as the Kozak context) (24), as well as the dwell time of the scanning ribosome at that codon (25). Unless the combination of these factors is strictly optimal for initiation, at least a small fraction of scanning complexes will bypass the potential start site and continue scanning, allowing translation to be initiated further downstream. When a potential initiation site is a non-AUG codon or an AUG in a weak Kozak context, only a small proportion of scanning complexes will initiate translation. Thus, leaky scanning may result in the translation of different coding sequences using numerous initiation sites, while initiation at start codons in the same reading frame can give rise to proteoforms with alternative N-termini (PANTs) (Figure 2.1B). A potentially large number of start codons may be used to initiate translation within the same Stop-Stop ORF, as is the case with the well-explored human *PTEN* gene where functionally distinct extended proteoforms are produced from multiple non-AUG starts (26). Annotating all Start-Stop ORFs is problematic due to the large number of potential start codons, and in certain cases, such as Repeat Associated Non-AUG (RAN) translation (27,28), the exact position of the initiation site cannot even be easily identified.

In the case of stop codon readthrough (13,29) and selenocysteine incorporation (30,31) (Figure 2.1C), the translation products produced by these processes can be defined in computational terms as the fusion of an upstream Start-Stop ORF with a downstream Stop-Stop ORF. Gene annotation projects currently resolve these cases by 're-writing’ the Stop codon or the selenocysteine codon in the protein file, allowing them to code through. This results in the “extended ORF” that overlaps with a “standard ORF” in the same frame. For programmed ribosomal frameshifting (Figure 2.1D), which is common in viruses but also infrequently occurs in cellular genes (32), the description of translation using ORFs would require the introduction of the location of the frameshift site as both start and stop codon. This

could enable the designation of the trans-frame protein product as a fusion of the two such "ORFs". In practice, gene annotation may instead introduce an artificial indel modification of the natural DNA/RNA sequences to yield a single contiguous ORF; for example, the [T] corresponding to human hg38 assembly Chr19:2271440 nucleotide is deleted in both RefSeq (e.g. NM_004152.3) and Ensembl (e.g. ENST00000582888.8). To this end, existing gene annotation of *in silico* trans-frame translation may yield a protein sequence corresponding to the product generated in nature. However, it comes at the expense of producing an incorrect sequence of an mRNA molecule, which does not allow for the regulatory mechanism at play to be accurately represented.

The examples in Figure 2.1 are not exhaustive and there are other translation phenomena that cannot be easily described using ORFs, such as translational bypassing (33–35) and StopGo (also known as StopCarryOn or 2A) (36). Regardless of which ORF definition is used, the concept of a translated ORF is not adequate to represent the complexity of RNA translation.

2.5 RNA translation is segmented

Ribosome profiling has revealed the existence of a large number of short translated sequences, currently termed small or short ORFs (smORFs, sORFs) or Ribo-seq ORFs, as the term CDS is reserved for sequences encoding classical proteins (37). Many Ribo-seq ORFs occur within the same RNA molecules. The lack of appropriate terminology reflecting the complexity of translation becomes even more evident when we consider the relationship between these translation segments. Upstream translation often influences downstream translation and this dependency is known to be utilised to regulate gene expression. For instance, many short translated regions upstream of CDSs (termed upstream ORFs, or uORFs) have been found to regulate translation by blocking ribosomes *via* sensing specific metabolites within the nascent peptide channel (38–42), reviewed in (43). This process is exemplified by translation regulation of the downstream CDSs by a short uORF in vertebrate *AMD1* encoding Adenosyl-Methionine Decarboxylase, a key enzyme in polyamine biosynthesis. The uORF encodes a short peptide MAGDIS that stalls ribosomes through its interactions with the ribosome in the presence of polyamines (44). These stalled ribosomes prevent other ribosomes from binding and scanning downstream to initiate at *AMD1's* CDS. Thus, the uORF provides a negative feedback control mechanism for *AMD1* expression, inhibiting its synthesis when polyamine concentration is high, but allowing for its synthesis when polyamine levels decrease (Figure 2.2A).

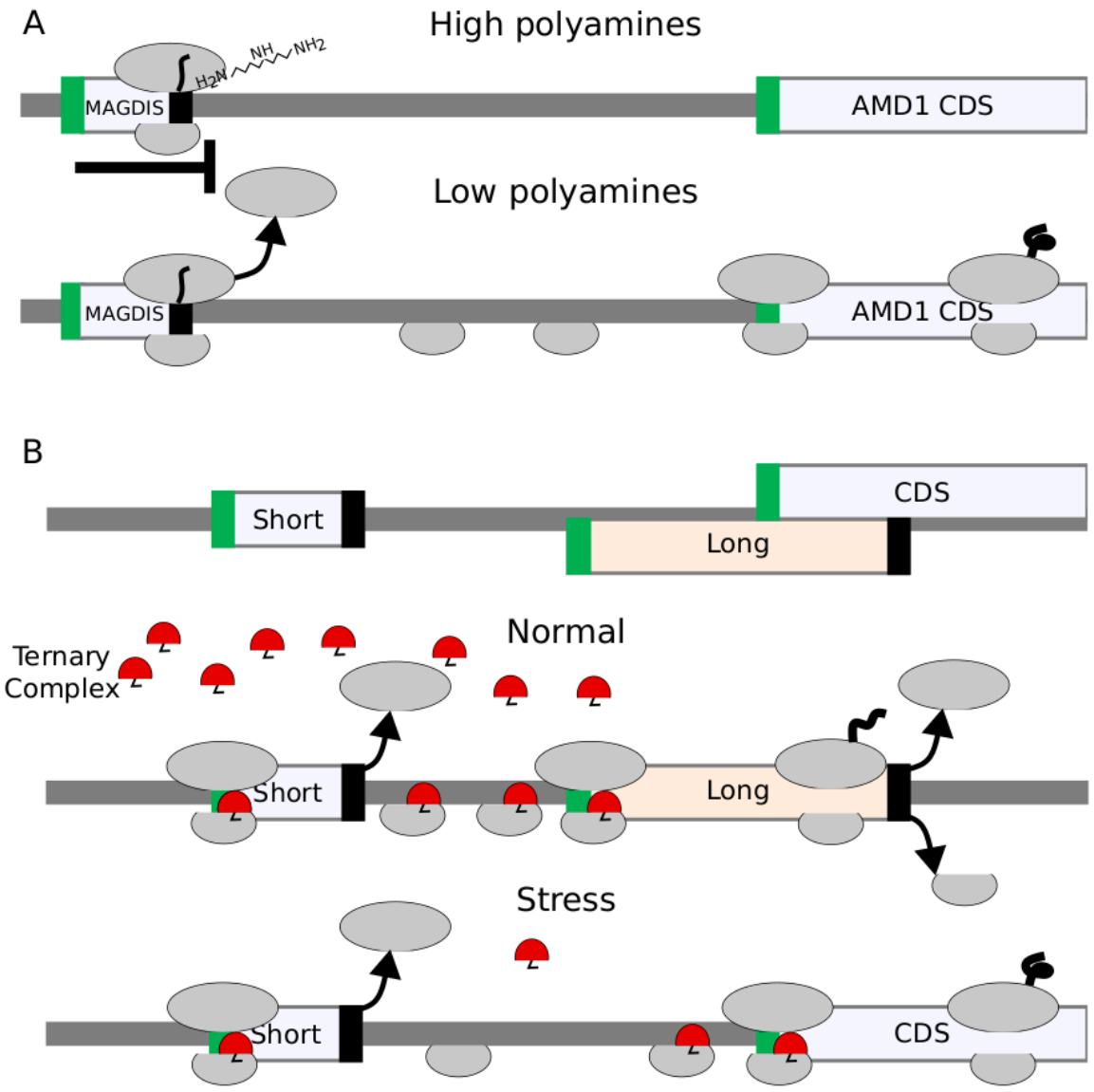


Figure 2.2. Relationship between translated segments within the same RNA. A. A schematic of metabolite-dependent translation regulation *via* ribosome arrest at uORF, such as in *AMD1* mRNA. In the presence of a high concentration of polyamines, ribosomes with MAGDIS peptide stall at the end of the translon. **B.** A schematic of delayed re-initiation mechanism enabling translation of selected mRNAs during global translation suppression caused by eIF2alpha phosphorylation during Integrated Stress Response. When eIF2 is phosphorylated the concentration of the eIF2*tRNAi*GTP ternary complex decreases and it requires a longer time and distance for the scanning ribosome complex to acquire it. As a result, the long uORF is bypassed and initiation occurs at CDS.

In addition to leaky scanning, re-initiation is another process impacting start codon selection. Translation re-initiation occurs when small ribosomal subunits remain bound to the mRNA after translation is complete and reinitiate downstream of the terminating stop codon. This is thought to be common after the translation of short ORFs as it takes time for initiation factors to dissociate from the ribosome. In this way, the ribosome may remain capable of initiation after translating a small number of codons, although other factors are known to contribute to this process, allowing for re-initiation in some instances even after the translation of long ORFs. The detailed molecular mechanisms of these processes are described in dedicated reviews (21,23,45–49). Re-initiation provides a platform for a rapid switch of gene expression on the translational level. Perhaps the most thoroughly studied is the case of delayed re-initiation (50–52) that protects translation of certain mRNAs (e.g. human A*TF4*, yeast *GCN4*) from down-regulation during the Integrated Stress Response (ISR) (53,54). Under this condition, the reduced availability of the ternary complex (tRNAi* eIF2*GTP) increases the time required for post-terminating ribosomes to bind the ternary complex enabling re-initiation. Therefore, the level of stress determines the location of the start codon at which re-initiation occurs. Figure 2.2B provides a schematic illustrating this mechanism.

It Is unclear to what extent the translated products of such regulatory translation contribute to the functional cellular proteome beyond their potential contribution to the antigen pool, as many of them lack conservation at the protein level (37,55,56). Extreme cases of translation regulation without peptide synthesis are represented by minimal ORFs consisting of a start codon immediately followed by a stop codon. While they obviously do not produce any functional peptide, some of them do have regulatory potential as strong ribosome stalling sites (57).

It is clear that translation complexity requires a unified and comprehensive abstraction that would adequately represent all translated regions – not only those that encode classical proteins – and reflect their mechanistic interrelationships. Such representation should be convenient for use by both human scientists when examining individual mRNA sequences and computer agents during programmatic analysis of large datasets.

2.6 Ribosome Paths

The complex nature of translational events and regulatory processes reveals the need to consider the entire passage of an individual ribosomal complex containing the same small ribosomal subunit along the mRNA, from the moment of pre-initiation complex assembly at the 5' cap (or IRES element) to the complete dissociation of both ribosomal subunits from the mRNA as a functional unit. We propose to term such a unit a Ribosome Path (RiboPath). It includes both regions that are scanned and those that are translated. As argued above, ORF is an inadequate descriptor of translated regions, and therefore we want to define and assign a new, unambiguous name to an entity denoting *translated region* as encompassing the entire sequence of RNA translated by a fully assembled elongating ribosome from initiation codon through termination and dissociation of the large ribosomal subunit. We term this region *translon*. See Supplemental Table S1 for the definitions of new terms. It has already been suggested as a term specifying a unit of translation (58) but has not yet been adopted. The main advantage of *translon* over ORF is that it is not constrained by the sequence (specific codons as boundaries). It is based on the process of translation and therefore may incorporate a variety of decoding mechanisms such as ribosomal frameshifting, stop codon readthrough, translational bypassing, etc (18,59). The other term commonly used to indicate translated regions is *cistron*, for example, polycistronic or monocistronic mRNAs. However, this term was originally defined genetically, different cistrons should be responsible for different phenotypes, and it is being used inconsistently in the literature.

To simplify the introduction of the RiboPath concept, for now, we only consider initiation and re-initiation as the mechanisms producing alternative proteoforms. We will exclude other translation mechanisms. Nevertheless, our framework can easily be extended to incorporate other translation mechanisms as we discuss later.

Figure 2.3 illustrates the RiboPath concept with an example of an mRNA encoding two proteoforms arising from alternative CUG and AUG initiation sites in one reading frame (cream) and a single upstream AUG codon in another frame (light lavender) as depicted in the ORF plot at the top. The corresponding translons are shown beneath. Alternative initiation and reinitiation allow the ribosome to pass through five different RiboPaths. The top RiboPath represents the ribosomes that initiate at the first AUG, but fail to reinitiate further

downstream, resulting in a path with a single translon T1. The second path corresponds to the ribosome that successfully re-initiates downstream, thus containing two translons, T1 and T2. In the third RiboPath the ribosomes fail to initiate at the first AUG, but start translation at the CUG allowing for translon T3, which encodes an N-terminally extended proteoform relative to the product of translon T2. The fourth RiboPath corresponds to the ribosomes that fail to initiate at both the first AUG and the CUG but succeed at initiating at the second AUG so that its RiboPath consists of only one translon T2. Finally, the fifth RiboPath is unproductive and represents the ribosomes that have not initiated protein synthesis on this mRNA. The RiboPath presentation makes it clear that certain translons are mutually exclusive as they never occur on the same path, e.g. a single ribosome cannot translate T1 and T3.

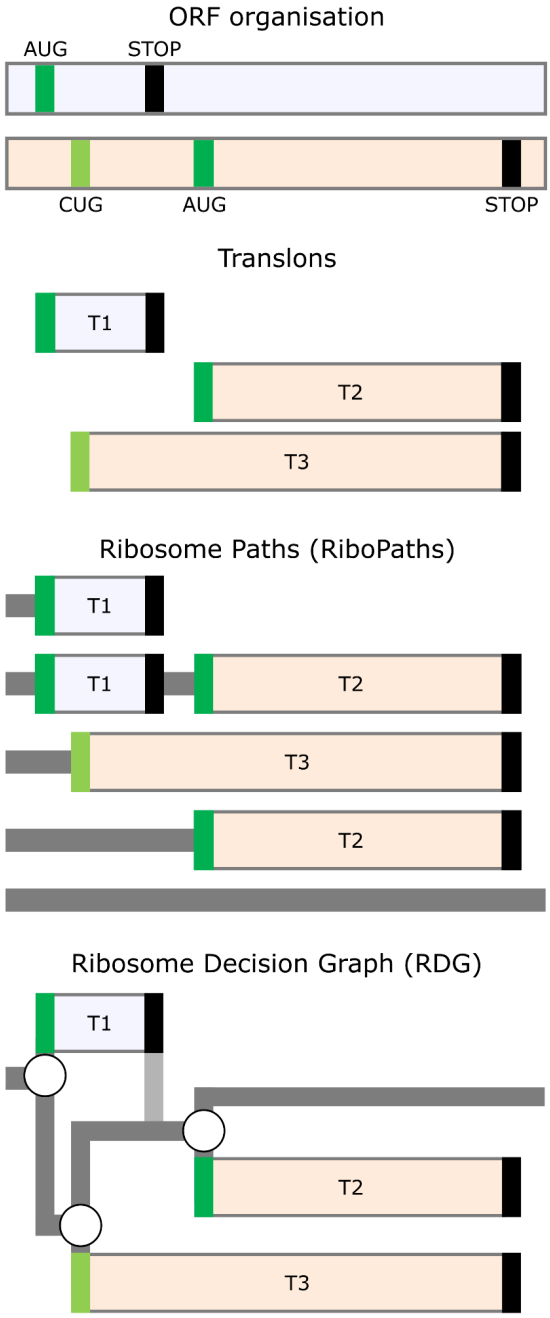


Figure 2.3. Ribosome paths through mRNA. Top: An example of an RNA with three start codons **(**green for AUG and light green for CUG) located in two different reading frames depicted as differentially shaded horizontal bars. The translation of this mRNA is represented as a set of translons **below**, Ribosome Paths **further below** and Ribosome Decision Graphs at the **bottom.** RNA regions scanned by the ribosome are shown in dark grey, vertical path in light grey represents the post-terminating small ribosomal subunit that continues scanning and remains initiation-competent.

2.7 Ribosome Decision Graphs

Once we represent the behaviour of translating ribosomes in terms of paths, it is only natural to further represent these in terms of graphs (Figure 2.3). The three initiation events in Figure 2.3 can be represented as branching points where the ribosome makes a "decision" of whether to initiate or not. We do not imply that ribosomes have free will; the decision is likely determined by the molecular composition and temporal thermodynamics of the local microstate. As in statistical mechanics, for practical purposes, it is appropriate to describe such decisions probabilistically, even if the underlying molecular processes are deterministic. The mRNA region engaged by ribosomes in Figure 2.3 can then be represented as a graph with three branching points. Stop codons in this graph are considered deterministic ends of translons as we exclude the possibility of stop codon readthrough or re-initiation after long translons in our illustrative example. Following this notation, any translated RNA can be represented as a Ribosome Decision Graph (RDG). As in the representation of translation using ORFs/CDSs, RDGs may be either conceptual (representing potential) or real (e.g., experimentally supported). In the case of conceptual RDGs, all potential start codons in mRNAs could be used as branching points, e.g., all AUGs, all CUGs, etc. depending on the specific parameters of the model. Such conceptual RDGs would be very complex graphs with a large number of branching points and possible paths. They are not suitable for evaluation by humans, but provide a straightforward method for generating all theoretically possible products of RNA translation. This can be used for the subsequent mining of mass spectrometry datasets. A set of graphs with branching points sampled from the set of all possible branching points can be used to generate simulated ribosome profiling data. The comparison of simulated and real data would enable the determination of the best RDG fitting the experimental data, thus inferring the real branching points from the data. As exemplified further below, RDGs may also be useful for analysing the impact of genetic variation, because variants that change or introduce new branching points (start and stop codons, frameshifts, etc) would alter the RDG topology.

RDGs could also be used to annotate experimentally validated translations. In this case, only those translation events for which there is experimental evidence will be introduced as branching points. In most cases, these experimentally informed RDGs would be suitable for

manual examination by researchers and would overcome the limitations of the data structures that are currently used for protein coding annotation.

2.8 Implementations of RDGs

To illustrate how RDGs can be used to represent the impact of variation within 5' leader sequences (a.k.a. 5' UTR) on downstream translation we selected the *NF2* variant responsible for neurofibromatosis type 2 (60). The 5' leader sequence of the *NF2* mRNA contains an AUG start codon followed by an in-frame AUG codon in a strong Kozak context. This suggests that few (if any) ribosomes reach the CDS start *via* leaky scanning. It is far more likely that CDS translation involves re-initiation at the CDS start as depicted in Figure 2.4A. A single base insertion variant was identified in two unrelated individuals in a cohort of 1,134 individuals diagnosed with neurofibromatosis type 2 **(**ENST00000338641:−66-65insT; GRCh37:Chr22:29999922A >AT) (60). This insertion causes both a shift in the reading frame and the introduction of another AUG. The shift extends translons T1 and T2, abrogating the initiation of translon T3 corresponding to the *NF2* CDS (Figure 2.4B).

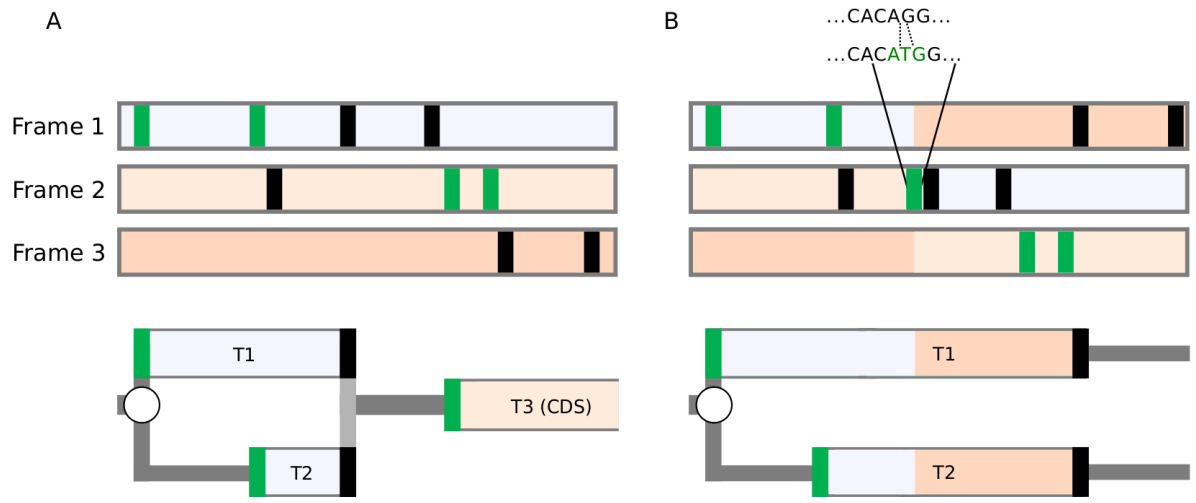


Figure 2.4: Representation of the effect of a genomic variant on CDS translation. Representations of the *NF2* mRNA for the reference sequence (**A**) and in the presence of the insertion variant (**B**). ORF organisation is shown at the **Top** with reading frames shaded differentially according to the reference sequence. AUG and stop codons are represented as green and black vertical bars, respectively. RDGs are shown at the bottom. Given the low probability of leaky scanning through the first two AUGs (see text), it is expected that the translon T3 corresponding to the CDS cannot be translated in the pathogenic variant sequence.

To illustrate how RDGs can be used for the representation of real translation data we chose two simple examples, namely human *NRAS* and *NXT1* mRNAs. The criteria for this selection were the existence of only a single transcript per gene according to GENCODE v.42, and the ribosome profiling supporting translation of only a single AUG-initiated translon in addition to the annotated CDS. Of note, translation of most human 5' mRNA leaders is more complex (see examples in Supplemental Figure S1), and therefore the advantages of using RDG representation for these are even greater, but may not be suitable for introducing this concept as interpretation of ribosome profiling data is more difficult.

Examination of ribosome profiling data in Trips-Viz (61) (Figure 2.5A) for *NXT1* mRNA reveals translation of an upstream region in the -1 frame (blue translon) relative to the CDS (red translon). Similarly, examination of Trips-Viz data indicates translation upstream and in the +1 reading frame (red translon) relative to the annotated *NRAS* CDS (blue translon). For simplicity, the CDS starts are not depicted as a branching point and are considered to be 100% efficient translation initiation sites. As the translated regions in both graphs are overlapping, it is clear that the simultaneous translation of both translons by the same ribosome cannot occur, at least in the absence of 3' to 5' scanning of post-terminating ribosomes (62).

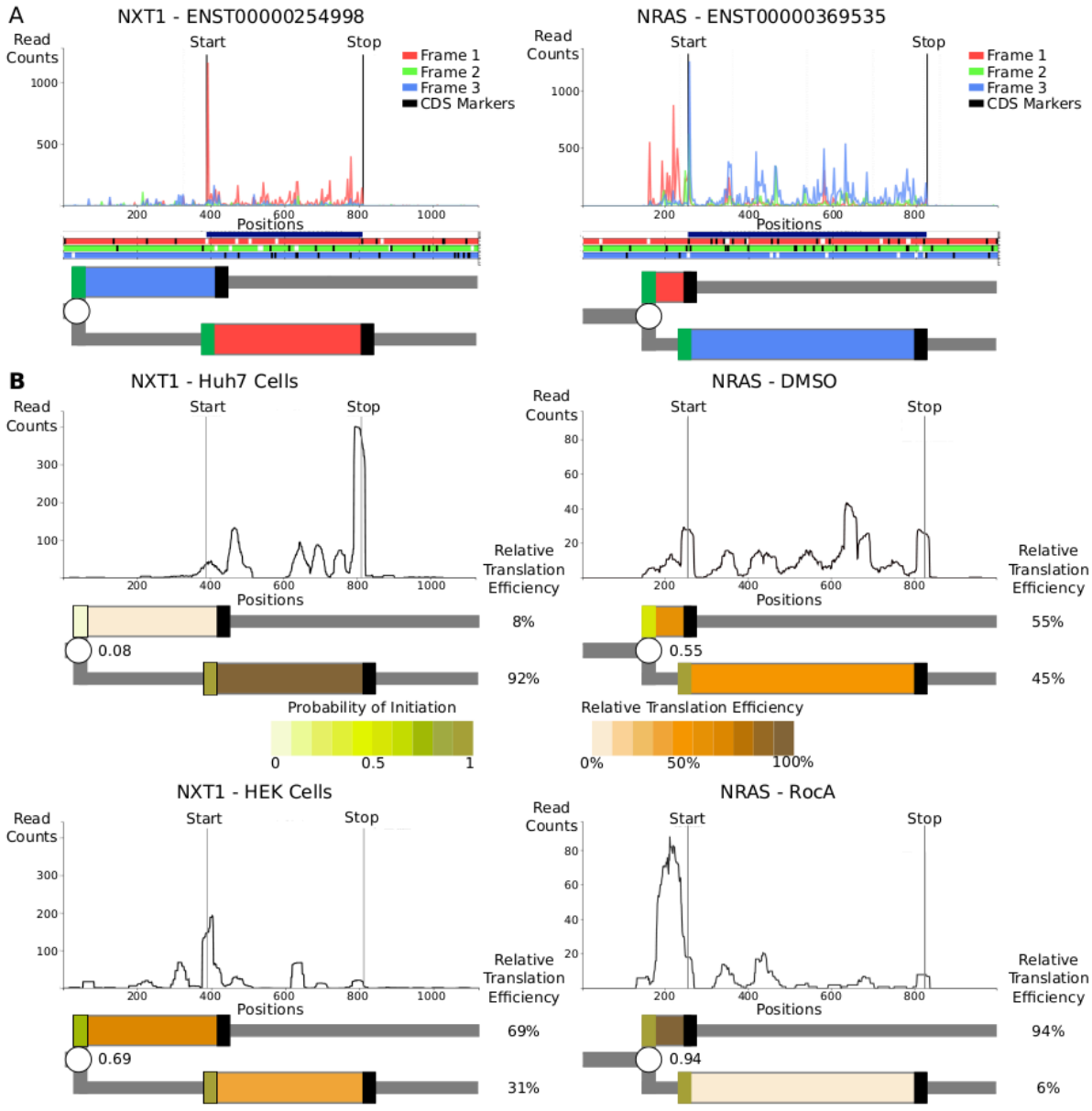


Figure 2.5: Representation of translation of human mRNAs. A. The top plots are A-site subcodon footprint densities for *NXT1* and *NRAS* mRNAs coloured to match the best-supported reading frame below in the ORF plots, in which AUGs are in white and stops are in black. Further below are RDG representations with translons coloured to match the translated frame. The CDS starts are treated deterministically and unproductive RiboPaths are not shown. **B.** Densities of ribosome A-site footprints obtained from ribosome profiling under different conditions (*NRAS*) or from different cells (*NXT1*). RDGs are shown below each density plot with translons coloured as a heatmap reflecting relative translation efficiencies. Branching points (starts) are also coloured as heatmaps reflecting the inferred probability of

their initiation. The probabilities of translation initiation are shown as fractions decimals, while the relative translation efficiencies of each path are shown as percentages (%).

In addition to representing qualitative information, RDGs also enable a quantitative representation of translation regulation. Due to the leaky scanning mechanism of translation initiation, the efficiencies of each CDSs translation in these two examples directly depend on the efficiencies of the upstream starts, e.g., if all ribosomes initiated at the upstream starts, no CDS translation would be observed. The relative translation efficiencies of translons can be used to calculate the probabilities of initiation at the upstream starts (63). These probabilities may vary between different conditions or across different cell types due to a variety of mechanisms, such as global changes in the stringency of start recognition (64,65) or specific regulation of mRNA via ribosome sensing of particular metabolites through interaction with the nascent peptide (38–42). Using RDG representations in this way makes it easier to characterize the relationship between translation events that are regulated (*via* changes in probabilities at branching points) and the relative rates of translons product synthesis.

To illustrate this with real examples, we examined the translation of the above genes using different ribosome profiling datasets (Figure 2.5B). For *NRAS,* we used the dataset from cells treated with Rocaglamide A (RocA) and its untreated control (66). As can be seen in Figure 2.5B the silhouette of ribosome footprint density for the *NRAS* mRNA changes dramatically upon RocA treatment. These footprint densities can be used to calculate the relative translation efficiencies of *NRAS* translons and to derive the probability of translation initiation at the upstream starts (Supplemental Methods). By showing the relative synthesis rates and initiation probabilities as heatmaps, the relationship between these two translons becomes apparent. RocA treatment greatly increases translation initiation probability at the upstream start, most likely *via* the ability of RocA to clamp initiation factor eIF4A to mRNAs containing specific sequence motifs (66), which then reduces the downstream CDS translon. In the case of *NXT1,* we examined data obtained in two different cell lines, HeLa (67) and Huh7 (68). The silhouettes of ribosome footprint densities for *NXT1* mRNA are markedly different, as can be seen in Figure 2.5B. The RDG visualization of these differences in ribosome footprint densities pinpoints the upstream start as the pivotal element of cell-specific

regulation of *NXT1* translation. For HeLa samples, the translation initiation at the upstream start is highly efficient, making the upstream start predominant. In contrast, for Huh7 samples efficiency at this start is much lower and, consequently, the CDS translon is predominant. The reasons for these cell-specific differences are beyond the scope of this work, but several mechanisms may be responsible, including different levels of translation factors that recognize translation initiation starts (69).

RDG figures, such as those presented in this manuscript, are significantly time consuming to produce manually. In order to demonstrate this concept without requiring the manual generation of RDGs we also introduce a supplemental software, RDG-Viewer (available at https://colab.research.google.com/drive/1f5iSgy5DAXeq27Lx1fCyngm4IjinkgC5?usp=sharing) . This Google Collaboratory notebook utilises graph construction functionality from the RDG python package (https://pypi.org/project/RDG/) which is currently under development (https://github.com/JackCurragh/RDG). For more details please see the Supplemental Material.

One of the attractive features of the RDG concept is its expandability. In the RDG examples above, we limited branching points only to starts where initiation and re-initiation events can occur. The most basic information for generating RDGs that allows only leaky scanning would require only locations of starts in a transcript because in-frame stop codons are identifiable from the sequence and are treated deterministically as the ends of translons. However, the concept can be extended to incorporate annotations for any non-deterministic translation events, such as stop codon readthrough or selenocysteine insertion (Figures 2.1C and 2.6A), ribosomal frameshifting (Figures 2.1D and 2.6B), translational bypassing (33) and even yet undiscovered translation phenomena. Annotation Schema in Supplemental Material provides an example of how such annotation could be organised computationally.

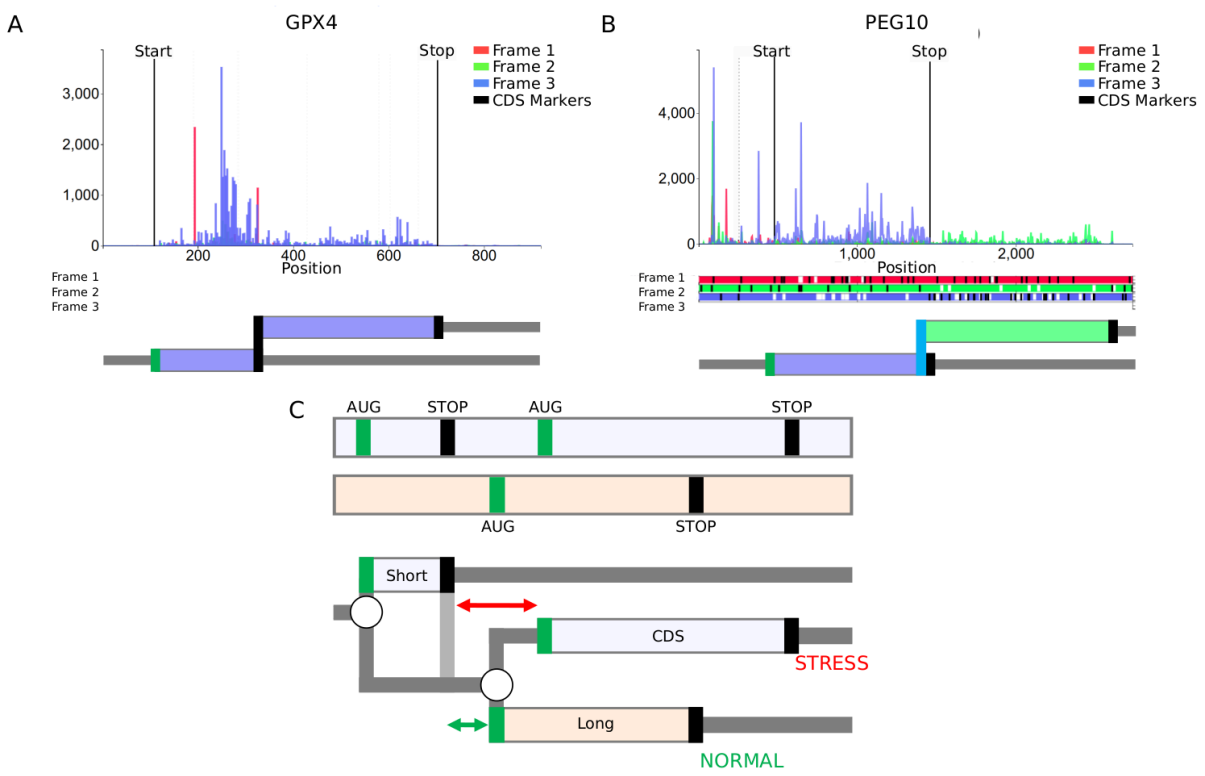


Figure 2.6: RDG representations of special cases. (A) mRNA encoding selenoprotein GPX4 and (**B**) PEG10 mRNA requiring ribosomal frameshifting for its expression. Trips-Viz screenshots of ribosomal profiling density for these mRNAs are at the top with ORF plots beneath and RDG representations further below. **C. Top**: An ORF plot containing three translons (protein-coding CDS and two uORFs, short and long), which is a minimal requirement for the mechanism of delayed re-initiation. **Below**: The corresponding RDG with green and red representing the predominant ribosome paths for normal and stress conditions, respectively. Arrows indicate the distance sufficient for scanning ribosomes to bind ternary complexes under normal (green) or stress (red) conditions.

Despite the apparent simplicity of RDGs notations, it would be naive to expect that it can represent a full range of translational mechanisms. For example, in the case of a delayed re-initiation mechanism (50–52) that makes the translation of certain mRNAs resistant to global down-regulation during Integrated Stress Response (ISR), it is not sufficient to simply add a stop codon as a branching point, allowing either ribosome dissociation from mRNA or re-initiation downstream. This is because the reduced availability of the ternary complex (tRNAi*eIF2*GTP) increases the time required for the post-terminating ribosomes to bind the

ternary complex, thereby enabling re-initiation (Figure 2.2B). Thus, it is not the probability of re-initiation, but the location of the start at which re-initiation will occur that changes during ISR. However, even in this case, the RDG concept can be useful to illustrate the mechanism, as shown in Figure 2.6C for a simplified mock transcript (Supplemental Figure S1 RDG for human *ATF4* that is regulated by delayed reinitiation). It is conceivable to extend the concepts of RDGs with parameters linking scanning distance to re-initiation probability.

An important shortcoming of the presented solution is the difficulty of its application to genomic loci encoding multiple transcript isoforms. The purpose of RDGs is to represent molecular events that take place during the translation of a single mRNA molecule. Therefore, a single RDG can only be applied to a single mRNA sequence. However, the concept of representing biological sequences as graphs is gaining momentum with splice graphs for representing alternative splicing (70) and variation graphs for representing pangenomes (71). Therefore, we envision that the RDG concept will fit into the emerging bioinformatics infrastructure of hierarchical representation of biological sequences as graphs, from genome to transcriptome to translatome.

2.9 Conclusion

The RDG concept has the potential to significantly impact the study of RNA translation complexity. RDGs, in combination with Ribo-seq data may shift the focus of differential translation analysis from changes in translation efficiencies of individual coding regions to the changes in the efficiencies of events regulating their translation. This focus shift will facilitate a mechanistic understanding of RNA translation regulation. Correlating mRNA translation with properties of RDGs (e.g. topology) may open a new possibility for identifying novel common mechanisms of translation regulation. In combination with information on genomic variants, RDGs have the potential to be instrumental in their phenotypic interpretation. Comparison of RDGs across orthologs would allow investigation of evolutionary constraints shaping translation regulation of specific genes.

2.10 References

1. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009 Apr 10;324(5924):218–23.

2. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity of Mammalian Proteomes. Cell. 2011 Nov 11;147(4):789–802.

3. Andreev DE, O'Connor PBF, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, et al. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. eLife. 2015 Jan 26;4:e03971.

4. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a source of new peptides. eLife. 2014 Sep 16;3:e03523.

5. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016 Feb;13(2):165–70.

6. Ji Z, Song R, Regev A, Struhl K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Sonenberg N, editor. eLife. 2015 Dec 19;4:e08890.

7. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. EMBO J. 2016 Apr 1;35(7):706–23.

8. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012 Nov;22(11):2219–29.

9. Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, et al. A high-resolution map of human RNA translation. Mol Cell. 2022 Aug 4;82(15):2885-2899.e8.

10. Wright BW, Molloy MP, Jaschke PR. Overlapping genes in natural and engineered genomes. Nat Rev Genet. 2022 Mar;23(3):154–68.

11. Chong C, Müller M, Pak H, Harnett D, Huber F, Grun D, et al. Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat Commun. 2020 Mar 10;11(1):1293.

12. Fedorova AD, Kiniry SJ, Andreev DE, Mudge JM, Baranov PV. Thousands of human non-AUG extended proteoforms lack evidence of evolutionary selection among mammals. Nat Commun. 2022 Dec 23;13(1):7910.

13. Dunn JG, Foo CK, Belletier NG, Gavis ER, Weissman JS. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. eLife. 2013 Dec 3;2:e01179.

14. Lobanov AV, Heaphy SM, Turanov AA, Gerashchenko MV, Pucciarelli S, Devaraj RR, et al. Position-dependent termination and widespread obligatory frameshifting in Euplotes translation. Nat Struct Mol Biol. 2017 Jan;24(1):61–8.

15. Ivanov IP, Firth AE, Michel AM, Atkins JF, Baranov PV. Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. Nucleic Acids Res. 2011 May;39(10):4220–34.

16. Wethmar K, Smink JJ, Leutz A. Upstream open reading frames: Molecular switches in (patho)physiology. Bioessays. 2010 Oct;32(10):885–93.

17. Namy O, Rousset JP, Napthine S, Brierley I. Reprogrammed Genetic Decoding in Cellular Gene Expression. Mol Cell. 2004 Jan 30;13(2):157–68.

18. Baranov PV, Gesteland RF, Atkins JF. Recoding: translational bifurcations in gene expression. Gene. 2002 Mar 20;286(2):187–201.

19. Ivanov IP, Atkins JF, Michael AJ. A profusion of upstream open reading frame mechanisms in polyamine-responsive translational regulation. Nucleic Acids Res. 2010 Jan 1;38(2):353–9.

20. Sieber P, Platzer M, Schuster S. The Definition of Open Reading Frame Revisited. Trends Genet TIG. 2018;34(3):167–70.

21. Jackson RJ, Hellen CUT, Pestova TV. THE MECHANISM OF EUKARYOTIC TRANSLATION INITIATION AND PRINCIPLES OF ITS REGULATION. Nat Rev Mol Cell Biol. 2010 Feb;11(2):113–27.

22. Hinnebusch AG. The scanning mechanism of eukaryotic translation initiation. Annu Rev Biochem. 2014;83:779–812.

23. Sonenberg N, Hinnebusch AG. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell. 2009 Feb 20;136(4):731–45.

24. Kozak M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol. 1987 Aug 20;196(4):947–50.

25. Kozak M. Context effects and inefficient initiation at non-AUG codons in eucaryotic cell-free translation systems. Mol Cell Biol. 1989 Nov;9(11):5073–80.

26. Tzani I, Ivanov IP, Andreev DE, Dmitriev RI, Dean KA, Baranov PV, et al. Systematic analysis of the PTEN 5' leader identifies a major AUU initiated proteoform. Open Biol. 2016 May;6(5).

27. Cleary JD, Ranum LPW. Repeat associated non-ATG (RAN) translation: new starts in microsatellite expansion disorders. Curr Opin Genet Dev. 2014 Jun;26:6–15.

28. Nguyen L, Cleary JD, Ranum LPW. Repeat-Associated Non-ATG Translation: Molecular Mechanisms and Contribution to Neurological Disease. Annu Rev Neurosci. 2019 Jul 8;42:227–47.

29. Loughran G, Chou MY, Ivanov IP, Jungreis I, Kellis M, Kiran AM, et al. Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Res. 2014 Aug 18;42(14):8928–38.

30. Driscoll DM, Copeland PR. Mechanism and regulation of selenoprotein synthesis. Annu Rev Nutr. 2003;23:17–40.

31. Labunskyy VM, Hatfield DL, Gladyshev VN. Selenoproteins: molecular pathways and physiological roles. Physiol Rev. 2014 Jul;94(3):739–77.

32. Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov PV. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res. 2016 Sep 6;44(15):7007–78.

33. Herr AJ, Atkins JF, Gesteland RF. Coupling of Open Reading Frames by Translational Bypassing. Annu Rev Biochem. 2000 Jul 1;69(Volume 69, 2000):343–72.

34. Nosek J, Tomaska L, Burger G, Lang BF. Programmed translational bypassing elements in mitochondria: structure, mobility, and evolutionary origin. Trends Genet TIG. 2015 Apr;31(4):187–94.

35. Klimova M, Senyushkina T, Samatova E, Peng BZ, Pearson M, Peske F, et al. EF-G–induced ribosome sliding along the noncoding mRNA. Sci Adv. 2019 Jun 5;5(6):eaaw9049.

36. Atkins JF, Wills NM, Loughran G, Wu CY, Parsawar K, Ryan MD, et al. A case for "StopGo": reprogramming translation to augment codon meaning of GGN by promoting unconventional termination (Stop) after addition of glycine and then allowing continued translation (Go). RNA N Y N. 2007 Jun;13(6):803–10.

37. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022 Jul;40(7):994–9.

38. Law GL, Raney A, Heusner C, Morris DR. Polyamine Regulation of Ribosome Pausing at the Upstream Open Reading Frame of S-Adenosylmethionine Decarboxylase *. J Biol Chem. 2001 Oct 12;276(41):38036–43.

39. Hardy S, Kostantin E, Wang SJ, Hristova T, Galicia-Vázquez G, Baranov PV, et al. Magnesium-sensitive upstream ORF controls PRL phosphatase expression to mediate energy metabolism. Proc Natl Acad Sci. 2019 Feb 19;116(8):2925–34.

40. Rahmani F, Hummel M, Schuurmans J, Wiese-Klinkenberg A, Smeekens S, Hanson J.

Sucrose Control of Translation Mediated by an Upstream Open Reading Frame-Encoded Peptide. Plant Physiol. 2009 Jul 1;150(3):1356–67.

41. Laing WA, Martínez-Sánchez M, Wright MA, Bulley SM, Brewster D, Dare AP, et al. An upstream open reading frame is essential for feedback regulation of ascorbate biosynthesis in Arabidopsis. Plant Cell. 2015 Mar;27(3):772–86.

42. Ivanov IP, Shin BS, Loughran G, Tzani I, Young-Baird SK, Cao C, et al. Polyamine Control of Translation Elongation Regulates Start Site Selection on Antizyme Inhibitor mRNA via Ribosome Queuing. Mol Cell. 2018 Apr 19;70(2):254-264.e6.

43. Dever TE, Ivanov IP, Hinnebusch AG. Translational regulation by uORFs and start codon selection stringency. Genes Dev. 2023 Jun 1;37(11–12):474–89.

44. Ruan H, Shantz LM, Pegg AE, Morris DR. The Upstream Open Reading Frame of the mRNA Encoding *S*-Adenosylmethionine Decarboxylase Is a Polyamine-responsive Translational Control Element*. J Biol Chem. 1996 Nov 22;271(47):29576–82.

45. Kozak M. Pushing the limits of the scanning mechanism for initiation of translation. Gene. 2002 Oct 16;299(1–2):1–34.

46. Andreev DE, Loughran G, Fedorova AD, Mikhaylova MS, Shatsky IN, Baranov PV. Non-AUG translation initiation in mammals. Genome Biol. 2022 May 9;23(1):111.

47. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5’-untranslated regions of eukaryotic mRNAs. Science. 2016 Jun 17;352(6292):1413–6.

48. Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 2017 Sep 1;31(17):1717–31.

49. Pestova TV, Kolupaeva VG, Lomakin IB, Pilipenko EV, Shatsky IN, Agol VI, et al. Molecular mechanisms of translation initiation in eukaryotes. Proc Natl Acad Sci U S A. 2001 Jun 19;98(13):7029–36.

50. Baird TD, Wek RC. Eukaryotic initiation factor 2 phosphorylation and translational control in metabolism. Adv Nutr Bethesda Md. 2012 May 1;3(3):307–21.

51. Andreev DE, Arnold M, Kiniry SJ, Loughran G, Michel AM, Rachinskii D, et al. TASEP modelling provides a parsimonious explanation for the ability of a single uORF to derepress translation during the integrated stress response. Sonenberg N, editor. eLife. 2018 Jun 22;7:e32563.

52. Hinnebusch AG. Translational regulation of yeast GCN4. A window on factors that control initiator-trna binding to the ribosome. J Biol Chem. 1997 Aug 29;272(35):21661–4.

53. Costa-Mattioli M, Walter P. The integrated stress response: From mechanism to disease. Science. 2020 Apr 24;368(6489):eaat5314.

54. Pakos-Zebrucka K, Koryga I, Mnich K, Ljujic M, Samali A, Gorman AM. The integrated stress response. EMBO Rep. 2016 Oct;17(10):1374–95.

55. Prensner JR, Abelin JG, Kok LW, Clauser KR, Mudge JM, Ruiz-Orera J, et al. What Can Ribo-Seq, Immunopeptidomics, and Proteomics Tell Us About the Noncanonical Proteome? Mol Cell Proteomics. 2023 Sep 1;22(9):100631.

56. Wacholder A, Parikh SB, Coelho NC, Acar O, Houghton C, Chou L, et al. A vast evolutionarily transient translatome contributes to phenotype and fitness. Cell Syst. 2023 May 17;14(5):363-381.e8.

57. Tanaka M, Sotta N, Yamazumi Y, Yamashita Y, Miwa K, Murota K, et al. The Minimum Open Reading Frame, AUG-Stop, Induces Boron-Dependent Ribosome Stalling and mRNA Degradation. Plant Cell. 2016;28(11):2830–49.

58. Goel SC. Transcription Unit. Nature. 1973 Oct;245(5425):397–397.

59. Rodnina MV, Korniy N, Klimova M, Karki P, Peng BZ, Senyushkina T, et al. Translational recoding: canonical translation mechanisms reinterpreted. Nucleic Acids Res. 2020 Feb 20;48(3):1056–67.

60. Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, et al. Characterising the loss-of-function impact of 5’ untranslated region variants in 15,708 individuals. Nat Commun. 2020 May 27;11:2523.

61. Kiniry SJ, Judge CE, Michel AM, Baranov PV. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. Nucleic Acids Res. 2021 Jul 2;49(W1):W662–70.

62. Gould PS, Dyer NP, Croft W, Ott S, Easton AJ. Cellular mRNAs access second ORFs using a novel amino acid sequence-dependent coupled translation termination-reinitiation mechanism. RNA N Y N. 2014 Mar;20(3):373–81.

63. Michel AM, Andreev DE, Baranov PV. Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. BMC Bioinformatics. 2014 Nov 21;15(1):380.

64. Loughran G, Sachs MS, Atkins JF, Ivanov IP. Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. Nucleic Acids Res. 2012 Apr;40(7):2898–906.

65. Fijalkowska D, Verbruggen S, Ndah E, Jonckheere V, Menschaert G, Van Damme P. eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs. Nucleic Acids Res. 2017 Jul 27;45(13):7997–8013.

66. Iwasaki S, Floor SN, Ingolia NT. Rocaglates convert DEAD-box protein eIF4A into a sequence-selective translational repressor. Nature. 2016 Jun;534(7608):558–61.

67. Park JE, Yi H, Kim Y, Chang H, Kim VN. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. Mol Cell. 2016 05;62(3):462–71.

68. Lintner NG, McClure KF, Petersen D, Londregan AT, Piotrowski DW, Wei L, et al. Selective stalling of human translation through small-molecule engagement of the ribosome nascent chain. PLoS Biol. 2017 Mar;15(3):e2001882.

69. Anisimova AS, Kolyupanova NM, Makarova NE, Egorov AA, Kulakovskiy IV, Dmitriev SE. Human Tissues Exhibit Diverse Composition of Translation Machinery. Int J Mol Sci. 2023 Jan;24(9):8361.

70. Ryan MC, Cleland J, Kim R, Wong WC, Weinstein JN. SpliceSeq: a resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts.

Bioinformatics. 2012 Sep 15;28(18):2385–7.

71. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023 May;617(7960):312–24.

2.11 Supplemental Material

2.11.1 Glossary of Concepts

A number of terms/concepts are introduced in this work. Supplemental Table S1 endeavours to delineate these terms and provide a clear reference resource for those wishing to build on these concepts.

Supplemental Table S1: Glossary of RDG Terms

Term	Definition	Notes
Translon	A Translon is the basic unit of translation. It refers to the entire path of a fully assembled elongating ribosome from initiation codon through termination and dissociation of the large ribosomal subunit.	Differs from the term Open Reading Frame (ORF) as it is not constrained by nucleotide sequence. It is instead only concerned with the activity of 80S ribosomes.
Ribosome Path (RiboPath)	A Ribosome Path refers to the way in which an individual ribosomal complex, containing the same small ribosomal subunit, traverses the mRNA from the moment of pre-initiation complex assembly at the 5' cap to the complete dissociation of both ribosomal subunits from the mRNA.	A RiboPath consists of one or more translons joined by untranslated regions scanned by a ribosomal complex with the same 40S subunit.
Ribosome Decision Graph (RDG)	A Ribosome Decision Graph is a graph representing all RiboPaths pertaining to a single mRNA molecule.	A Ribosome Decision Graph is a graph based representation of the different ways how a specific RNA molecule can be translated. Nodes in these graphs represent the locations of alternative, nondeterministic events. This includes switches between unproductive (e.g. scanning) and productive (elongation) modes of translation as well as switches between reading frames (i.e. during ribosomal frameshifting) as well as yet undiscovered translation phenomena.

2.11.2 RDG-Viewer

The RDG-Viewer is a Python 3 code that is a part of the RGD package (currently under development) that enables the generation and visualisation of RDGs. Specifically, it enables the following functionalities:

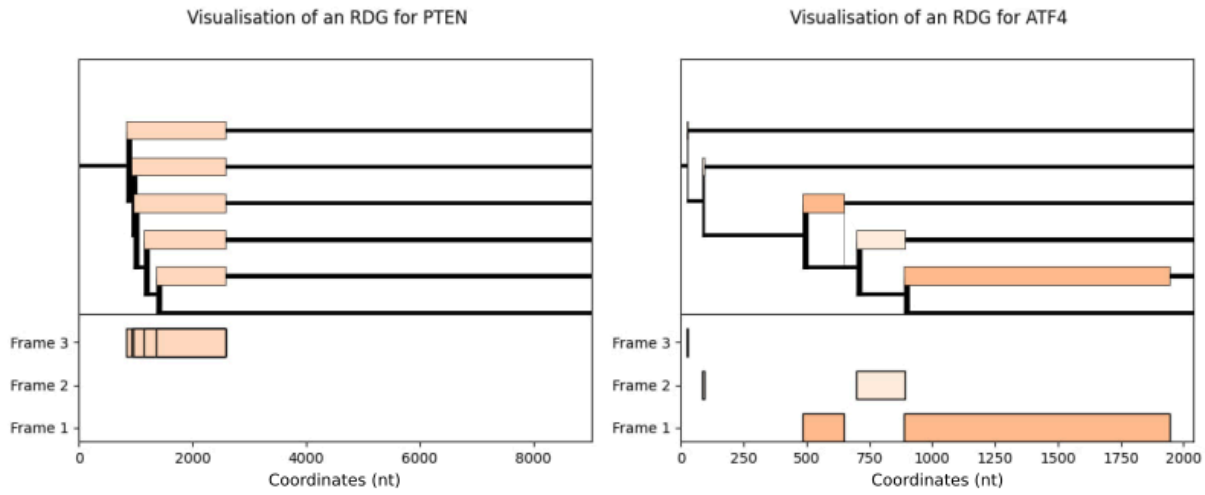
- Generate RDG using annotated translon coordinates as input.
- Generate RDG using the transcript nucleotide sequence and customizable parameters of translation as input.
- Generate images of the plots containing RDGs along with all translons grouped by three different reading frames. Supplemental Figure S1 shows two examples of such plots.

The nucleotide sequences could also be automatically obtained with gget (1) using transcript ENSEMBL ID or gene name symbols, in which case the first transcript isoform is selected from the list of ENSEMBL isoforms for that gene. The gene functionality is currently limited to *H. sapiens*.

The source code of RDG viewer along with its documentation is available at GitHub: https://github.com/JackCurragh/RDG-Viewer

The RDG-Viewer is also implemented in a Google Collaboratory notebook at https://colab.research.google.com/drive/1f5iSgy5DAXeq27Lx1fCyngm4IjinkgC5?usp=sharing and can be used for building RDGs and their visualisations for individual transcripts without a need of installing RDG package scripts locally.

2.11.3 Complex Examples Visualised



Supplemental Figure S1: RDG-Viewer plots. Left: Human *PTEN* transcript (ENST0000037195) with RDG constructed based on the data from (2). **Right**: Human *ATF4* transcript (ENST00000337304) constructed by using all upstream AUGs as branching points. The upper part of these plots contain RDGs, while the bottom part shows translations grouped by three different reading frames.

2.11.4 Annotation Schema

The most basic information for generating RDGs requires only locations of starts in a transcript because in-frame stop codons are identifiable from the sequence and are treated deterministically as the ends of translons. Therefore, the example shown in Figure 2.1B of the main manuscript can be represented as a short array:

starts(x_1;x_2;x_3;x_4)

while the example in Figure 2.3 of the main manuscript as:

starts(x_1;x_2;x_3)

However, the concept can be extended to incorporate annotations for any non-deterministic translation events. For example, to capture such phenomena as stop codon readthrough or selenocysteine insertion (Figures 2.1C and 2.6A in the main manuscript), an addition of a new type of branching point would be necessary: a stop codon, at which the ribosome could either terminate or incorporate an amino acid and continue translation.

starts(x_1); secs(y_1)

A frameshifting site would be yet another type of branching point where the ribosome could either continue translation in the same frame or shift to one of the alternative reading frames. For example, the case of ribosomal frameshifting (Figures 2.1D and 2.6B in the main manuscript) can be represented as:

starts(x_1); shifts(i_1,s_1)

where shifts need to be assigned two coordinates, the coordinate of the last codon in the initial frame $\boldsymbol{i_1}$ and the coordinate of the first codon in the shifted frame $\boldsymbol{s}$. Accordingly, $\boldsymbol{s_1 - i_1} = 1$ would correspond to +1 frameshifting, $\boldsymbol{s_1 - iy_1} = -1$ to -1 frameshifting, and so on. This notation is compatible not only with common types of frameshifting but also with extreme events such as bypassing in bacteriophage T4 gene *60* (34), where 50 nucleotides are skipped by the ribosome without translation, i.e. expressed as ***shifts(i,s)*** where $\boldsymbol{s - i} = 50$.

Algorithms interpreting such annotations will be backward and forward-compatible since older algorithms could simply ignore new types of branching points.

2.11.5 Methods

We used GENCODE v42 (3) as a source of transcriptome annotation when searching for suitable loci and GENCODE v25 when producing visualisations in Trips-Viz (4). Pre-processed alignments from datasets GSE70211, GSE79664, and GSE94454 obtained in relevant studies (5–7) were obtained from Trips-Viz. The translation efficiencies of examined regions were calculated as the number of ribosome footprints uniquely aligning to individual translons normalised over the length of translons used for the mapping. Five first codons of translons were excluded from mapping to avoid the distortions introduced by the high peaks at the starts. The probability of translation initiation was then calculated as $p=R_u/(R_u+R_d)$, where R_u and R_d are translation efficiencies of the upstream and downstream translons.

2.11.6 Supplemental References

1. Luebbert L, Pachter L. Efficient querying of genomic reference databases with gget. Bioinformatics. 2023 Jan 1;39(1):btac836.

2. Tzani I, Ivanov IP, Andreev DE, Dmitriev RI, Dean KA, Baranov PV, et al. Systematic analysis of the PTEN 5' leader identifies a major AUU initiated proteoform. Open Biol. 2016 May;6(5).

3. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. GENCODE 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D916–23.

4. Kiniry SJ, Judge CE, Michel AM, Baranov PV. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. Nucleic Acids Res. 2021 Jul 2;49(W1):W662–70.

5. Iwasaki S, Floor SN, Ingolia NT. Rocaglates convert DEAD-box protein eIF4A into a sequence-selective translational repressor. Nature. 2016 Jun;534(7608):558–61.

6. Park JE, Yi H, Kim Y, Chang H, Kim VN. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. Mol Cell. 2016 05;62(3):462–71.

7. Lintner NG, McClure KF, Petersen D, Londregan AT, Piotrowski DW, Wei L, et al. Selective stalling of human translation through small-molecule engagement of the ribosome nascent chain. PLoS Biol. 2017 Mar;15(3):e2001882.

3. RDG: A Python package for the visualisation and analysis of translation complexity with Ribosome Decision Graphs

Jack A. S. Tierney[1, 2,], Pavel V. Baranov[1]

[1] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland.

[2] SFI Centre for Research Training in Genomics Data Science, University College Cork, Cork, Ireland

3.1 Foreword

Following the introduction of the Ribosome Decision Graph concept, we now present a comprehensive Python package for carrying out operations on these graphs. The RDG Python package was developed alongside the development of the concept and supported the conceptualisation process. This package enables researchers to construct, visualise and analyse RDGs. This package serves as a bridge between the theoretical concept and its practical application, allowing researchers to explore translational complexity in ways that were previously challenging or impossible. The existence of such a package is essential for the wider adoption of this concept by the community.

3.2 Abstract

Motivation: Ribosome Decision Graphs (RDGs) have recently been proposed as a novel concept for the accurate representation of translational complexity. This complexity arises due to the plurality of translation initiation sites, enabled by leaky scanning and reinitiation as well as due to recoding events such as ribosomal frameshifting. The locations of these events are represented as branching points forming directed acyclic graphs containing all potential paths of the ribosome in one data structure. To maximise the potential of this concept a versatile implementation is required.

Results: We introduce RDG, a Python package and command line application for constructing and working with Ribosome Decision Graphs. The RDG Python package provides a toolkit for working with RDGs in the form of a Python class and a set of auxiliary functions that enable users to build, modify, save and visualise RDGs.

Availability and Implementation: The RDG Python package is freely available for installation from the Python Package Index (PyPI) with pip install RDG or directly from the GitHub repository at https://github.com/JackCurragh/RDG.

3.3 Introduction

Since the introduction of the ribosome profiling (Ribo-Seq) technique fifteen years ago, it has become evident that current representations of what is translated from a given mRNA transcript are grossly insufficient. Ribo-seq data has been widely used to identify translated regions (translons) that are translated from open reading frames (ORFs) other than the coding sequence (CDS). Various annotation efforts have been initiated with the view of understanding the extent of the human translatome (1–3) and further research has looked to understand the functional role of these translons as both protein-coding and regulatory regions (4–7).

The mechanisms leading to this expression complexity at the level of translation have been known for decades. Marilyn Kozak proposed the scanning model of translation in a series of publications in the late 1970s/early 1980s (8–11) and more formally defined it in 2002 (12). According to this model, the small ribosomal subunit first binds to the 5` end of the mRNA and begins scanning the RNA sequence in a 5` to 3` direction for an initiation codon in sufficient context. Scanning ribosomes may either scan past potential initiation codons or initiate translation in a probabilistic manner. The probability of successful initiation ("decision" to initiate) depends on a range of factors including sequence context, presence/absence of secondary structure or other roadblocks such as RNA binding proteins and the identity of factors associated with the ribosome. Furthermore, a ribosome that has translated one translon may then reinitiate at a downstream start codon. As not all ribosomes will reinitiate post-termination, the reinitiation can also be seen as a stochastic decision which is likely predominantly based on the persisting association of initiation factors with the elongating ribosomal complex.

As a result of these mechanisms, the proximity of a given initiation codon to the 5` end of a transcript has a major influence on whether it will be initiated as it increases the likelihood of a 40S ribosome scanning that codon. It has become abundantly clear that the majority of mammalian mRNA transcripts are translated, to some extent, 5` of their annotated coding sequence. These, broadly unannotated, translation events impact the expression of the coding sequence whether the product of the upstream translon is functional or not.

Beyond the pervasive translation outside of the canonical CDSs, further translational complexity arises from an array of recoding events where the oft-touted rules of the genetic code and triplet decoding of mRNA are broken (13,14). These events are well documented with numerous examples across the Tree of Life, however, current annotation schemas poorly support the representation of this biology.

We recently proposed a novel concept to accommodate the representation of this translational complexity in one graph-based model, termed Ribosome Decision Graphs (RDGs) (15). RDGs are per-mRNA direct acyclic graph representations of the set of ribosome paths that a given ribosome may take when decoding an mRNA. Representing translation events (initiation/leaky scanning, termination/reinitiation, termination/readthrough and so on) as binary nodes enables the determination of potential event combinations from a set of known translation events. This is the first concept that tries to address the relationships between translons on a given RNA as previous approaches and annotations represent the expression of translons as independent events.

Ribosome decision graphs representing the expression of individual translons as non-independent events is a shift in the paradigm of the analysis of translational complexity. The concept has the potential to have major impacts on the annotation of translated regions, the interpretation of non-coding variants in the 5` leader regions, comparative translational regulation analysis and differential translation analysis.

Maximisation of the potential of this concept across this wide breadth of potential applications requires the development of an accessible toolkit that enables researchers to work with RDG objects without needing to reinvent the wheel. To this end, we introduce the RDG Python package (https://github.com/JackCurragh/RDG).

The main objective of developing the RDG Python package is to provide a comprehensive and user-friendly toolkit for constructing, visualising, and analysing Ribosome Decision Graphs (RDGs). As the first software tool dedicated to working with RDGs, the package aims to facilitate the adoption of this novel approach for representing and studying translational complexity in genomic analyses.

3.4 Methods

The RDG Python package centres around a Python class that defines the ribosome decision graph data structure and facilitates the majority of operations on the RDG. Along with this, a set of auxiliary modules that carry out common applications of the graph structures such as the construction of RDGs from RNA sequence are included. Core functionality for working with RDG objects are included as methods of the RDG class which will be of interest to prospective developers; the majority of users will be interested in the auxiliary functionality such as the construction of graphs from mRNA sequence, graph visualisation and graph storage. This functionality is made accessible as callable functions or via the command line interface satisfying the needs of a broad range of users.

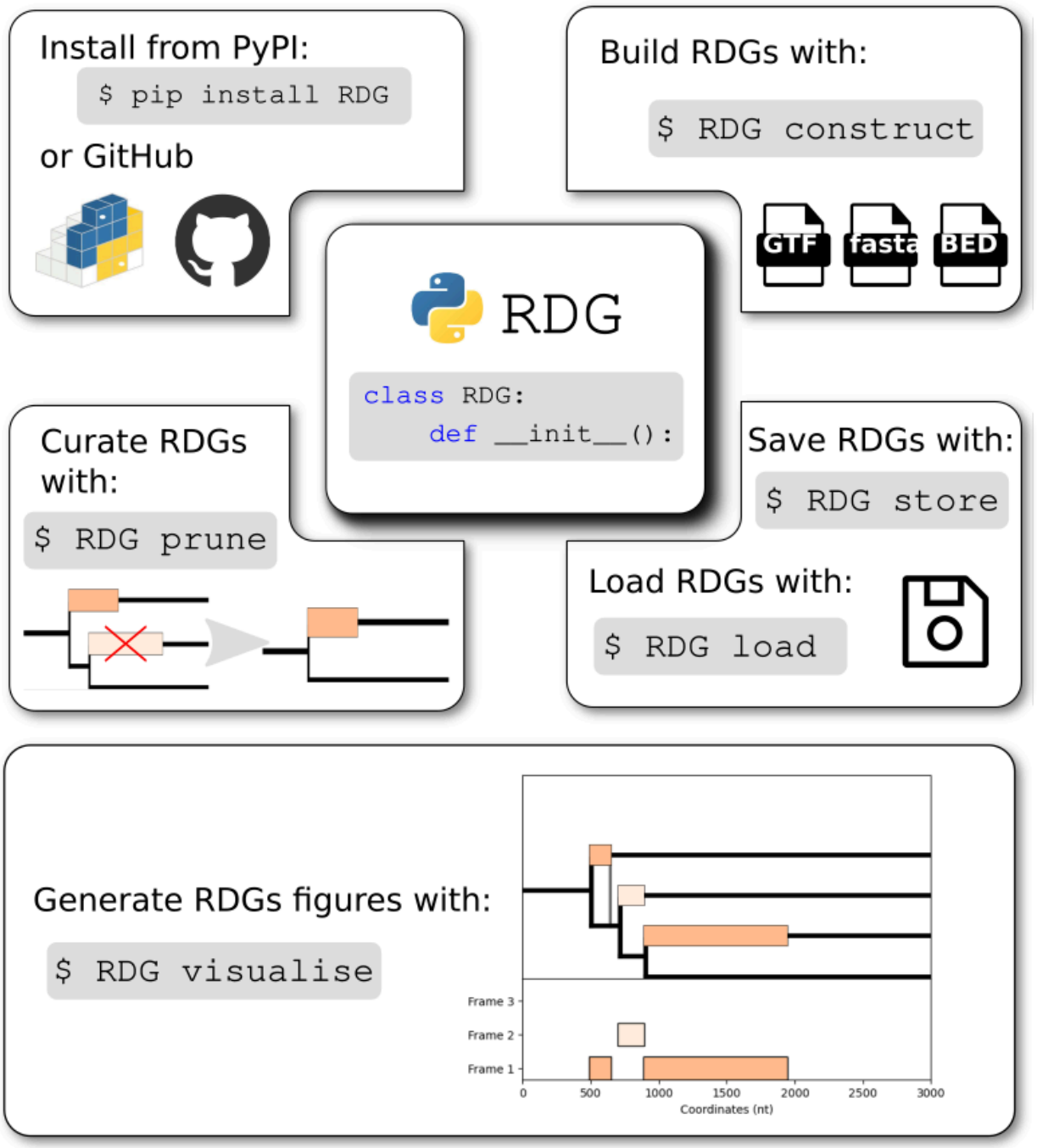


Figure 3.1: Overview of the RDG application and its functionalities.

3.4.1 RDG Class

Each instance of the RDG class is one Ribosome Decision Graph where the nodes, corresponding to translation events (e.g., initiation, termination, recoding event), and edges, representing relations between events, comprising the graph are recorded in RDG properties.

The RDG object has extensive methods for constructing and traversing the graphs as well as querying relationships between translation events. Node and Edge Python classes are used to record properties of individual translation events and relations between them. Nodes of RDGs may be any of the following node types: '5 Prime', '3 Prime', 'Start', 'Stop', 'Frameshift', and 'Readthrough Stop'. While edges can be either ‘Translated’ or ‘Untranslated’. Properties such as probabilities of translation events, Kozak contexts of start nodes and translon PhyloCSF scores can also be recorded as node properties.

3.4.2 RDG Construct Functionality

The RDG Python package supports the construction of graphs in a variety of manners. Graphs can be constructed from annotation files (BED and GTF formats), from sequence files (FASTA format) along with a set of parameters, such as start codons to consider, or from user-specified coordinates from the command line. These options are enacted by the auxiliary ‘RDG construct’ module making use of various RDG class methods.

The construction of RDGs straight from sequence is an attractive option however due to the wide range of known near-cognate start codons, graphs of long sequences often become unwieldy. Allowing all near-cognates ensures that the most commonly traversed paths will be included in the graph; however the more nodes included, the harder it is to determine which path is most likely. To address this, the RDG construct module supports the specification of various ‘principles of graph construction’ such as allowing reinitiation to be dependent on distance from the stop codon or the length of the preceding translon and the number of starts passed which the ribosome can scan. The most appropriate principles of graph construction will vary depending on the desired application.

In the RDG Python package, graphs are constructed by the progressive addition of translons to the graph accounting for the various specified principles upon translon addition. The process begins with the initialisation of a graph with just one 5` and one 3` terminal node. The insertion of a translon then follows the following steps:

1. Identification of existing edges of type ‘Untranslated’ with which the start node intersects.
2. For each clashing branch:

2.1. Initialise new Node objects for translon start and translon stop nodes using specified start and stop codon positions.
2.2. Create a new ‘Translated’ Edge object to link translon start to stop.
2.3. Update clashing Edge to terminate at the translon start site.
2.4. Update upstream Node to record the translon start as an out node.
2.5. Create a new ‘3 Prime’ Node object as an endpoint for the case where translation does not occur.
2.6. Add new ‘Untranslated’ Edge from translon start to new 3` terminal node.
2.7. Create another ‘3 Prime’ Node object as an endpoint for the case where translation did occur
2.8. Connect this ‘3 Prime’ node to translon stop with another new ‘Untranslated’ Edge.

The creation of the additional ‘3 Prime’ node and ‘Untranslated’ path following the occurrence of translation is important to ensure progressive addition of downstream translations is still supported. Without this reinitiation would not be possible as reinitiation support involves checking these ‘Untranslated’ Edges to determine if they permit translation according to the ‘principles of graph construction’.

The addition of any recoding event (stop codon readthrough or ribosomal frameshifting) requires the canonical coding path to be represented in the graph already.

Stop codon readthrough is supported by the conversion of a stop codon, that is specified by their transcript coordinate, to a boolean node with both ‘Translated’ and ‘Untranslated’ outgoing edges that represent the case of readthrough and the case of termination respectively. Again these edges will terminate at either nodes that were already downstream of the stop or new ‘3 Prime’ nodes.

The addition of a ribosomal frameshifting event involves the splitting of the ‘Translated’ edge within which the event occurs in two with a binary node of type ‘Frameshifting’. This node is linked to the original stop node representing the case where frameshifting did not occur and to a new stop node that is in a different frame than the original stop node. Again the addition of

‘Untranslated’ edges and corresponding terminal nodes is maintained to ensure reinitiation support is maintained.

Constructed graphs can then be used for analysis, be visualised or stored in a file for later visualisation and analysis.

3.4.3 RDG Visualise Functionality

The RDG Python package comes with its own novel plotting functionality developed using matplotlib (16). It is possible to conceive a number of different styles in which ribosome decision graphs can be depicted, however, currently only a cladogram inspired representation is supported. In these plots the horizontal axis represents the transcript coordinates from zero to the length of the transcript. The layout of the individual translons across the three different reading frames is presented in a subplot below the ribosome decision graph. The colours corresponding to each of the three frames can be customised as can ratios of edge sizes allowing the user to customise their RDG figures.

The graph layout is a horizontal cladogram with the root on the left and the leaves on the right. The positioning of nodes on the horizontal axis is straightforward as it is determined by the coordinate, on the level of the transcript, at which the translation event occurs. The vertical axis positioning is then more complicated as it must ensure there are no intersecting edges. The algorithm for vertical positioning starts by assigning the positions of all ‘3 Prime’ nodes followed by the iterative assignment of positions to branch points. This is done by averaging the vertical positions of their downstream nodes so that the branching occurs half way between the two downstream nodes. Following this any non-branch nodes can be assigned a vertical position so that they all run parallel based on the vertical position of their upstream branch node.

3.4.4 RDG Read and Write Functionality

RDGs can be stored for later use using SQLitedict or Pickle. In Python, pickling is the process of serialising Python objects into a character stream with all of the information required to fully recreate the Python object. Pickling RDGs is the preferred method for storing an individual graph, however, SQLitedict based storage is recommended for large sets of

ribosome decision graphs as it allows random access. This means even in a set of RDGs the size of the human transcriptome individual ribosome decision graphs can be easily obtained based on the locus name.

As ribosome decision graphs consist of a set of binary branch points as do phylogenetic trees. The RDG package also supports the writing of RDGs to Newick format. This file format/syntax is commonly used to store phylogenies in text files. It uses paired parentheses and commas to record the child and parent node relationships and floating point numbers after node IDs to signify branch lengths. This enables users to repurpose many phylogeny visualisation software for the visualisation of RDGs.

3.4.5 RDG Prune Functionality

Basic functionality supporting the ‘pruning’ of branches from ribosome decision graphs based on certain properties of nodes and branches is also supported. This involves the removal of nodes and their daughter branches while maintaining graph integrity. There are a broad range of potential applications of this feature, however, in the current version of the package the only auxiliary functionality supported is the pruning of RDG branches at any branch points that are not sufficiently supported by ribosome profiling reads. This serves as a proof of principle for future annotation projects utilising RDGs.

3.4.6 Installation and Accessibility

The RDG Python package is freely available for installation from the Python Package Index (PyPI) with pip install RDG or directly from the GitHub repository at https://github.com/JackCurragh/RDG. Furthermore, a Google Collaboratory notebook was previously released as a supplementary method for the original concept paper (https://colab.research.google.com/drive/1f5iSgy5DAXeq27Lx1fCyngm4IjinkgC5?usp=sharing).

3.5 Discussion

Current annotation schemas, and the bioinformatic applications for working with them, have limited functionality for working with the plurality of translation initiation sites, the relationships between translated regions and various recoding events. The RDG Python package provides the community with a toolkit for addressing these biological phenomena via an implementation of the RDG concept.

The development of the RDG package is an ongoing process with a range of enhancements in the pipeline. A core focus going forward is the refinement of the branch pruning functionality enabling users to trim RDGs based on a broader set of parameters such as conservation scores or experimental evidence. We also aim to expand the customisation options for RDG visualisation, empowering researchers to generate high-quality, publication-ready figures that capture the complexity of their transcript's translational landscape. Integration of the RDG package with existing bioinformatics tools and pipelines is a further priority to maximise the reach of this approach.

We foresee a wide range of potential applications of the RDG Python package across multiple domains of genomic research. In the context of genome annotation, this tool may streamline the identification of novel translons as it can drastically reduce the set of candidates based on the ‘principles of graph construction’. These improved annotations will in turn enable better characterisation of the effect of variants in non-coding regions such as 5' leaders including how variants in these regions can influence the translation of other translons. Researchers aiming to carry out cross-species comparisons of translational regulation will be able to leverage this tool to identify conserved ribosome decision graph structure as a proxy for conserved translational regulation. The package may also prove valuable for research into differential translation, as RDGs support the assignment of probabilities of translation events. The detection of changes in these probabilities across conditions has the potential to uncover novel insights into translational regulation.

3.6 References

1. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022 Jul;40(7):994–9.

2. Chothani SP, Adami E, Widjaja AA, Langley SR, Viswanathan S, Pua CJ, et al. A high-resolution map of human RNA translation. Mol Cell. 2022 Aug 4;82(15):2885-2899.e8.

3. Martinez TF, Chu Q, Donaldson C, Tan D, Shokhirev MN, Saghatelian A. Accurate annotation of human protein-coding small open reading frames. Nat Chem Biol. 2020 Apr;16(4):458–68.

4. Chothani S, Ho L, Schafer S, Rackham O. Discovering microproteins: making the most of ribosome profiling data. RNA Biol. 2023 Dec 31;20(1):943–54.

5. Chen J, Brunner AD, Cogan JZ, Nuñez JK, Fields AP, Adamson B, et al. Pervasive functional translation of noncanonical human open reading frames. Science. 2020 Mar 6;367(6482):1140–6.

6. Couso JP, Patraquim P. Classification and function of small open reading frames. Nat Rev Mol Cell Biol. 2017 Sep;18(9):575–89.

7. Zhang H, Wang Y, Wu X, Tang X, Wu C, Lu J. Determinants of genome-wide distribution and evolution of uORFs in eukaryotes. Nat Commun. 2021 Feb 17;12(1):1076.

8. Kozak M. How do eucaryotic ribosomes select initiation regions in messenger RNA? Cell. 1978 Dec 1;15(4):1109–23.

9. Kozak M. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. Nature. 1984 Mar 15;308(5956):241–6.

10. Kozak M. Evaluation of the “scanning model” for initiation of protein synthesis in eucaryotes. Cell. 1980 Nov;22(1 Pt 1):7–8.

11. Kozak M. Possible role of flanking nucleotides in recognition of the AUG initiator codon

by eukaryotic ribosomes. Nucleic Acids Res. 1981 Oct 24;9(20):5233–52.

12. Kozak M. Pushing the limits of the scanning mechanism for initiation of translation. Gene. 2002 Oct 16;299(1–2):1–34.

13. Atkins JF, Gesteland RF. Recoding: Expansion of Decoding Rules Enriches Gene Expression. 2009.

14. Rodnina MV, Korniy N, Klimova M, Karki P, Peng BZ, Senyushkina T, et al. Translational recoding: canonical translation mechanisms reinterpreted. Nucleic Acids Res. 2020 Feb 20;48(3):1056–67.

15. Tierney JAS, Świrski M, Tjeldnes H, Mudge JM, Kufel J, Whiffin N, et al. Ribosome decision graphs for the representation of eukaryotic RNA translation complexity. Genome Res. 2024 Apr 1;34(4):530–8.

16. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007 May;9(3):90–5.

4. Ribo-Seq Data Portal: Empowering Large-Scale Translation Reanalysis through Access to Pre-processed Samples and Standardised Metadata

Jack A. S. Tierney [1,2,#], Michał Świrski [3,#], Håkon Tjeldnes [1,4,#], Anmol M. Kiran [1], Gionmattia Carancini [1,2], Stephen J. Kiniry [5], Joanna Kufel [3], Eivind Valen [4,6*], Pavel V. Baranov [1*]

[1] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

[2] SFI CRT in Genomics Data Science, University of Galway, Galway, Ireland

[3] Institute of Genetics and Biotechnology, Faculty of Biology, University of Warsaw, Warsaw, Poland

[4] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

[5] EIRNA Bio, Cork, Ireland

[6] Department of Biosciences, University of Oslo, Oslo, Norway.

[#] Denotes equal contribution

* Denotes corresponding author(s)

4.1 Foreword

In the journey toward the understanding of translatomic complexity, accessible and standardised translatomic data rivals the need for accurate representations of translation events for pertinence. To address this need, we introduce the RiboSeq Data Portal which provides access to thousands of pre-processed Ribo-Seq samples along with standardised sample metadata. The development of this resource marks a major advance forward in making Ribo-Seq data accessible to the wider community, potentially accelerating discoveries through translatomic and multi-omic studies.

Much of the metadata cleaning was carried out in collaboration with Håkon Tjeldnes and Michał Świrski, particularly when they both visited University College Cork during the summer of 2023. Disorganised metadata is arguably the biggest barrier to leveraging the wealth of existing -omic data that is available.

Anmol Kiran was invaluable in making rdp.ucc.ie live to the public through his systems administration wizardry.

4.2 Abstract

Ribosome profiling (Ribo-Seq), invented in 2009, has consistently grown in popularity and public repositories now contain more than 13800 samples. Despite its prevalence, large-scale reanalysis of Ribo-Seq data lags behind other assays such as RNA-seq and ChIP-seq. This partly owes to a lack of data processing and metadata standardisation. Here, we introduce the Ribo-Seq Data Portal encompassing the entirety of public Ribo-Seq datasets and presents unified, standardised and pre-processed data together with its curated metadata. Beyond data retrieval, its integration with several existing visualisation tools enables large-scale analysis of Ribo-Seq data for researchers irrespective of their computational infrastructure access.

4.3 Background

Ribosome profiling, more commonly referred to as Ribo-Seq, involves the arrest of translating ribosomes while they decode mRNA into amino acids. Following the arrest of ribosomes, a nuclease is used to digest RNA that is not protected by bound ribosomes. Approximately 30nt long Ribosome-protected fragments (RPFs or footprints) are then isolated and sequenced. The alignment of their sequences to a reference genome or transcriptome enables genome-wide analysis of translation at sub-codon resolution (1–3).

Since the initial publication of this technique in 2009 (4), several variants targeting ribosome complexes at different stages of translation have been developed. In addition to cycloheximide which targets ribosomes in the pre-translocation conformation, several other drugs have been used to arrest ribosomes in other conformations of elongating ribosomes (5,6). Furthermore, ribosomes can be preferentially arrested during the initiation of translation with drugs such as lactimidomycin or harringtonine (7–9) allowing mapping of translation initiation sites, while formaldehyde crosslinking can be used to isolate scanning complexes (10). Variations of these techniques enable the assessment of compartmentalised translation (11,12) and the dynamics of co-translational factors bound to the ribosomes (13–15). As a result, numerous applications of ribosome profiling data have emerged over the past decade. Most commonly ribosome profiling is used for the identification of translated regions outside annotated CoDing Sequences (CDSs) (16–21) and for investigating the regulation of gene expression at the level of translation (4,22,23).

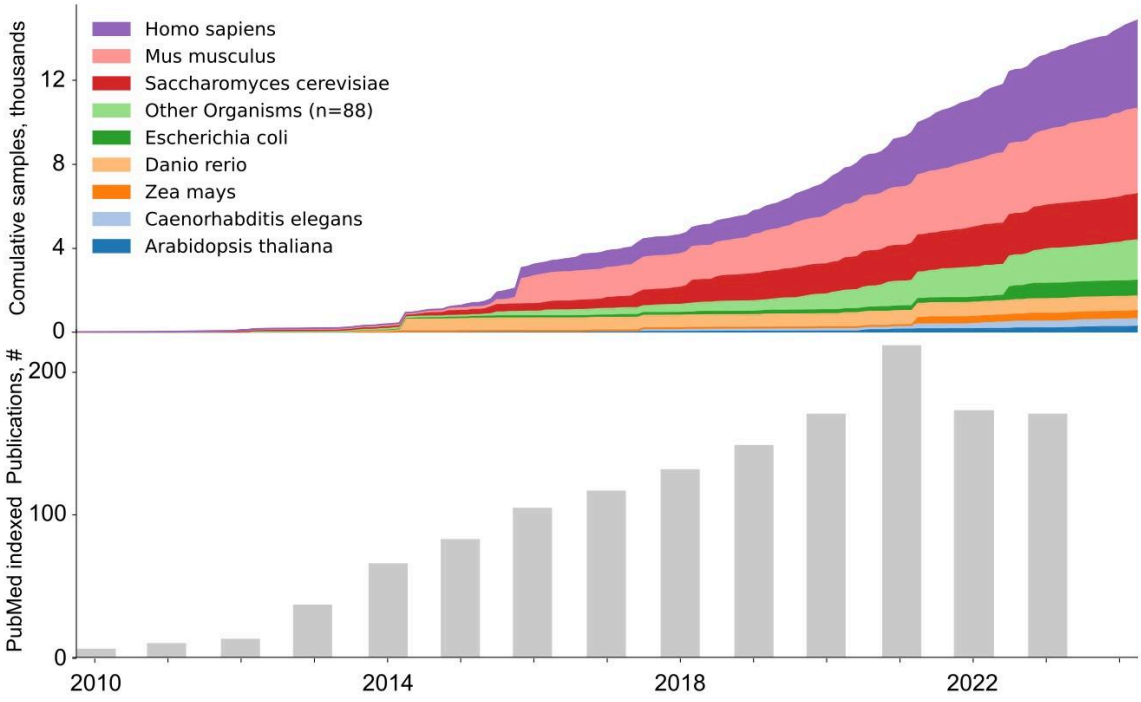


Figure 4.1: Growth of Ribosome Profiling Data and Literature

Top: Cumulative line plot illustrating the increasing number of publicly available ribosome profiling datasets across different organisms over time. Each curve represents a distinct organism, with the area under the curve filled to facilitate visual comparison. **Bottom:** Bar chart displaying the yearly trend in the number of publications related to ribosome profiling or Ribo-Seq, as indicated by the presence of these terms in the title or abstract of PubMed entries. The y-axis represents the count of relevant publications for each corresponding year on the x-axis. The decrease in the number of publications after 2021 likely reflects the drop in ribosome profiling being mentioned in the abstracts rather than a decrease in its use.

Before the advent of this technique, the extent to which mRNA is translated was widely underestimated. Ribosome profiling confirmed earlier predictions that approximately half of all human protein-coding mRNA transcripts undergo translation upstream of the protein-coding sequence due to the presence of AUG codons (24,25). Several bioinformatic approaches have been developed to predict translated regions from Ribo-Seq data (26–31). The identification of numerous short translated regions prompted a community-driven effort

to catalogue them (Mudge *et al.*, 2021) as well as independent efforts to utilise large numbers of datasets to generate annotations and make them available in the browser such as OpenProt (17,18).

Ribosome profiling has enabled genome-wide analysis of gene expression regulation at the translational level when paired with RNA-Seq to account for mRNA abundance levels. This approach can be applied to report changes in ribosome occupancy under different conditions or across different tissue types. Several methods have been developed for discriminating gene expression changes due to translation regulation and alterations of RNA levels, also allowing the identification of genes in which changes in RNA levels are buffered by reciprocal changes in translation (23,32).

Due to the insights this technique provides, the popularity of Ribo-Seq has grown as has the number of datasets deposited in public repositories such as the European Nucleotide Archive (ENA) (33) and the Sequence Read Archive (SRA) (34) (Figure 4.1A). However, these data continue to be underutilised, partly due to inconsistencies in the accompanying metadata. Ribosome Profiling Database (RPFdb) was a specialised database designed to store, organise, and provide access to ribo-seq data, however, it has become inaccessible (35). This database served as a valuable resource for researchers studying translation by offering a centralised platform where they could access ribosome footprinting data, translated ORF annotations and data quality information.

As the Ribo-Seq technique gained traction and an increasing number of datasets were being published, it became clear that most studies produced data for a single purpose and did not explore the full potential of their datasets. Consequently, there is much to be learned from reprocessing and reanalysing these datasets. Any reanalysis of publicly available Ribo-Seq data will typically repeat the same preprocessing steps to obtain ‘clean’ ribosome-protected fragments prior to carrying out alignment steps. This redundancy in data processing incurs an accumulation of costs, both financial and environmental. RiboSeq.Org (https://riboseq.org) endeavours to provide publicly available tools and data resources, including GWIPS-viz (36), RiboGalaxy (37), Trips-Viz (38) and RiboCrypt (39), which have been developed over the course of the last decade. GWIPS-viz, Trips-Viz and RiboCrypt are populated with processed public Ribo-Seq datasets and support visualisation and analysis in a browser (explored further

below). However, due to the independent development of these platforms, they are populated by different pipelines that share the aforementioned redundant pre-processing steps. Consequently, it became our goal to standardise these data processing steps across RiboSeq.Org resources and make the resulting processed data public, to reduce the cost of reanalysis to the community.

As there are only minimal rules governing the description of sequenced samples when depositing in public databases, there is considerable variability in the resolution and accuracy of the metadata provided by the depositors. The distribution of relevant information across synonymously named data fields, along with the use of variable spellings and synonyms within these fields, makes it challenging to identify related samples originating from different projects. Consequently, it is intractable to carry out large-scale analyses of translation using publicly available Ribo-Seq datasets without significant manual intervention.

Previous efforts and methods have been developed to characterise metadata from public repositories for a similar purpose (40–42). However, these methods do not cater to the specifics of ribosome profiling metadata effectively. To our knowledge, no prior method accurately characterises ribosome profiling metadata and makes it available to the wider community.

The aggregation of samples with a similar protocol design can be used to alleviate some of the intrinsic noise associated with Ribo-Seq data.overage gained from such aggregation enables methods to identify translated regions that may be missed when analysing a single sample. The aggregation of samples with a similar protocol design can also be used to alleviate some of the intrinsic noise associated with Ribo-Seq data.

4.4 Results

4.4.1 Ribo-Seq Data Portal - A Platform for the Exploration of Ribosome Profiling Datasets

The Ribo-Seq Data Portal is home to most of the publicly available ribosome profiling datasets preprocessed for downstream analysis accompanied by organised and manually curated information on the origins and experimental conditions. From this single online resource, a user can identify samples of interest, download processed data and associated metadata for their own downstream data analysis or analyse the data directly within their web browser using RiboSeq.Org tools (GWIPS-viz, Trips-Viz and RiboCrypt).

The current version of the Data Portal contains 14,840 samples, 4040 of which have been further validated manually by RiboSeq.Org contributors. These samples originated from a total of 969 Ribo-Seq studies across more than 96 different species/strains.

Each sample on the Ribo-Seq Data Portal has undergone a series of metadata curation steps using a controlled vocabulary consisting of a set of standard subject headings (e.g. Cell Line, Tissue, Inhibitor), paired with their own thesauri that enables the homogenisation of terms across studies. For instance, the terms *'chx', 'Cyclohex', and 'cycloheximide'* are converged to the term 'Cycloheximide'. This process resulted in the list of global metadata field names collapsing from a total of 813 entries to a total of just 88. Of these, 31 core fields come directly from the metadata table downloaded from the SRA for each sample. Information is extracted from these core fields to identify Ribo-Seq samples and subsequently to populate other fields when they are not ascribed to their own field for that sample. An example of this is where the data depositor has not specified the cell line used in its own metadata column as it was specified in the sample's name. By searching for keywords in these core columns the correct cell line for this sample can be identified. However, priority is given to information specified explicitly in their own column. This standardised metadata enables coherent search and filter functionality across the database and is also made available for download in CSV format. Downloaded CSVs can then be used to specify the batches and conditions to be used in large-scale differential expression reanalysis.

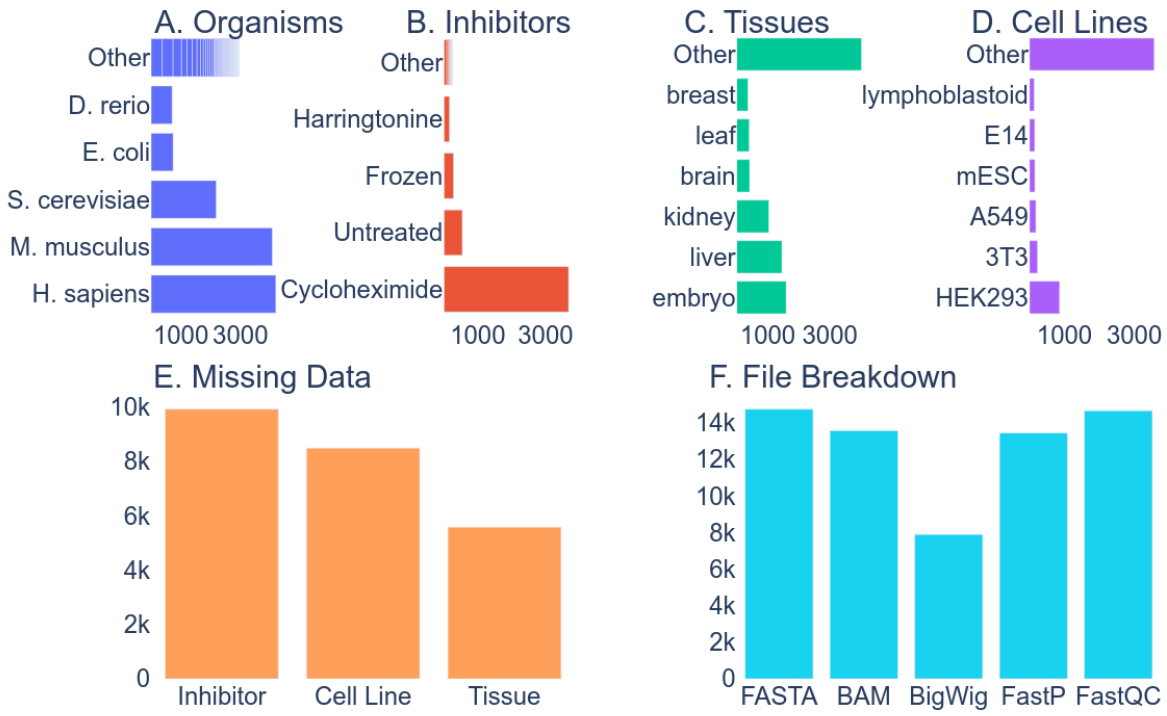


Figure 4.2: Metadata Landscape of RiboSeq Data Portal (RDP)

A-D. Horizontal bar charts displaying the distribution of organisms **(A)**, inhibitors **(B)**, tissues **(C)**, and cell lines **(D)** represented in the RiboSeq Data Portal Database Each bar segment corresponds to a distinct category within the respective metadata field, with the segment width indicating the relative proportion of datasets associated with that category.

E. Bar chart illustrating the extent of missing metadata across various fields. The bars represent the proportion of datasets lacking information for each metadata field, with taller segments indicating a greater degree of missing data.

F. Bar chart depicting the availability of different file types within the ribosome profiling datasets. Each bar corresponds to a specific file type, such as raw data files, processed data files, or metadata files, with the segment height reflecting the relative frequency of datasets containing that file type.

Furthermore, sequence read files for each sample are obtained from the SRA, and are processed using standardised approaches, ensuring that no additional batch effects are

introduced due to variable data processing steps. Although it may not be the optimal way to process each individual sample it enables large-scale reprocessing. These resulting processed files are then made available for download *via* the Data Portal. The primary file type made available is a collapsed read file in a collapsed FASTA format where unique reads are stored once per sample and the number of occurrences of that read is stored in the sequence description. Other file types made available include BigWig and binary alignment maps (BAM) along with the corresponding quality control reports from tools such as FastQC and FASTP. The provision of these processed data files makes it possible for other researchers to carry out their own downstream analysis with varied degrees of flexibility without having to repeat redundant preprocessing steps when accessing the datasets from the SRA or ENA.



Figure 4.3: A graphical schema of the RiboSeq Data Portal workflow.

Samples selected within the portal (above) can be readily viewed, analysed or downloaded *via* the various means shown (below).

The development of the Ribo-Seq Data Portal greatly improved the interoperability of decentralised web resources such as GWIPS-viz, Trips-Viz and RiboCrypt allowing parallel exploration of translated regions at the genome and transcriptome levels with minimal effort.

4.4.2 Programmatic Access Facilitates Large-Scale Analysis Projects

The value of the data made accessible *via* the Ribo-Seq Data Portal would be greatly diminished if manual or custom steps were required to integrate this data repository into an analysis workflow. To this end, we also introduce RDP-Tools and the Data Portal REST API. As RiboSeq Data Portal has adopted non-canonical variants of standard file formats the interoperability of this platform with external tools requires the provision of extended tooling. RDP-Tools is a Python-based command-line application that enables users to handle data files obtained from the Ribo-Seq Data Portal in their own workflow without the need for custom scripting. Primarily, it facilitates the collapsing and de-collapsing (inflating) of identical Ribo-Seq reads in various common file formats (FASTA, FASTQ, BAM).

Furthermore, RDP-Tools enables programmatic interaction with the Data Portals application programming interface (API). Queries *via* the Ribo-Seq Data Portal API return the requested metadata in JSON format. This may include download links that can easily be used to fetch the files from the portal as part of a data processing workflow. As a result, a user can identify, download and process Ribo-Seq samples for analysis without manual intervention.

4.4.3 Ribo-Seq Data Portal enables Large-Scale Interrogation of Translation with RiboSeq.Org

After selecting the data of interest, users can follow links to relevant RiboSeq.Org resources to visualise and analyse their data selection. GWIPS-viz, Trips-Viz and RiboCrypt produce visualisations of ribosome profiles, but in different contexts. GWIPS-viz is based on the UCSC genome browser (43)and enables the visualisation of genomic alignments (44). As

these alignments are carried out against unannotated genomic sequences, translation support for each annotation can be manually identified using GWIPS-viz. Another advantage of this platform is that translation outside annotated regions, such as in unannotated exons, retained introns and intergenic regions, can also be identified. Moreover, genomic alignments can identify the transcript isoform with the best translational support for further investigation on Trips-Viz or RiboCrypt.

On Trips-Viz, users can visualise ribosome footprints mapped directly to annotated RNA transcript sequences. Consequently, it is possible to assign each aligned read to a specific reading frame based on the position of their inferred A-site relative to the transcript start site (TSS). This frame information is key to the identification of translated regions (translons) (45) and may help identify translation mechanisms such as frameshifting or stop codon readthrough (46). Furthermore, Trips-Viz can be used to calculate global parameters of the datasets, compare footprint density profiles of individual transcripts between sample sets, and carry out differential gene expression analysis among many other functionalities (38).

RiboCrypt's visualisation functionality aims to combine the aforementioned positive aspects of GWIPS-viz and Trips-Viz by providing translated reading frame information within genomically aligned ribosome profiles. This is an R Shiny app that supports both genomic and transcriptomic views from these genomic alignments. The use of a splice-aware aligner (STAR) minimises the effects of the reduced mapping around splice junctions that occur on GWIPS-viz due to the use of a non-splice-aware aligner. The heatmap visualisation functionality provided by RiboCrypt enables the simultaneous visualisation of hundreds of Ribo-Seq samples.

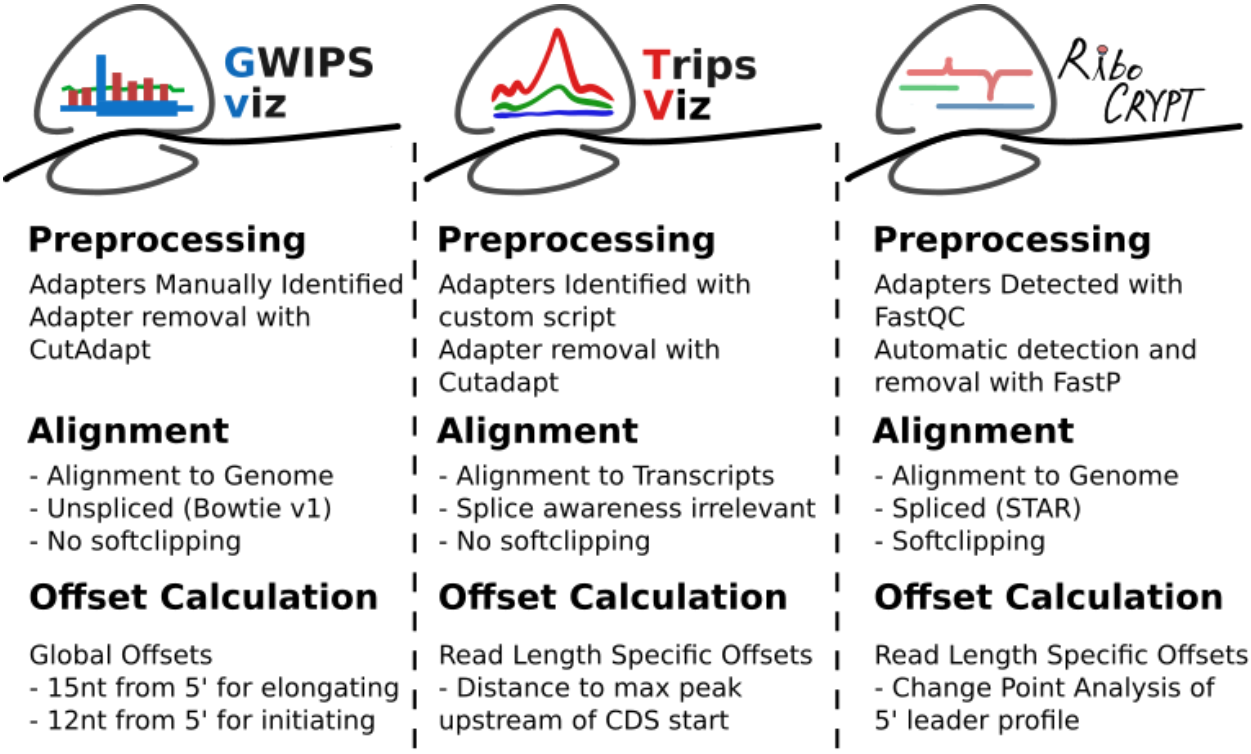


Figure 4.4. Comparison of three RiboSeq.Org browsers.
Breakdown of differences in data processing workflows between RiboSeq.Org resources covering preprocessing, alignment and offset calculations.

Similar to Trips-Viz, RiboCrypt also provides in-browser downstream analysis on preprocessed publicly available datasets. RiboCrypt offers differential expression analysis using either FPKM ratios or DESeq2 and analysis of ribosome dwell times per codon. Its 'MetaBrowser' enables the visualisation of all available samples for a given organism clustered based on the samples' metadata, for example, clustering based on the tissue or cell origin of the sample.

The combined use of these platforms, guided by the characterised metadata available on the data portal, enables researchers to explore the translation dynamics at play in publicly available ribosome profiling samples.

4.5 Discussion

Large-scale reanalysis of public data has resulted in significant insights across many different data types. A recent striking example of data reanalysis is a search of more than 20 petabases of sequencing data from over 5.7 million samples with the goal of identifying novel RNA viruses. As well as identifying over a million novel viruses, this project illustrated how such large-scale reanalysis can be carried out economically (47).

Reanalysis of publicly available datasets also allows researchers to uncover potentially confounding factors that would otherwise be undetectable when analysing data from a single source. Fasterius *et al.* characterised the degree of genetic heterogeneity of various cell line populations identifying notable examples of stable and unstable cell lines (48). This reanalysis project identified important considerations for future research that carries out same-cell line comparisons. Similarly, McCabe *et al.* reanalysed RNA-Seq data from epithelial ovarian cancer to measure the suitability of cell lines that are commonly used as *in vitro* models of major subtypes of this cancer (49). These examples show the potential for reanalysis leading to improvements in our experimental approaches.

In relation to ribosome profiling, the reanalysis of 49 available datasets obtained by multiple laboratories from cells treated with RocA revealed previously unnoticed effects of this drug on elongating ribosomes (50), while reanalysis of aggregated Ribo-Seq datasets from 13 publications revealed the existence of thousands of human proteoforms with extended N-termini due to initiation at non-AUG codons (Fedorova et al 2022). Aggregating Ribo-Seq datasets has been effectively used to improve the signal-to-noise ratio of the data on the RiboSeq.Org visualisation platforms. This provides clearer insights into the regions of RNA that are being translated on a given transcript. Translon calling with aggregated data as an input is employed by Trips-Viz (38) The benefits of this aggregation are highlighted in a recent review (Chothani *et al.*, 2023). The aggregation described here is a pan-tissue/cell line approach, however, with accurately annotated sample metadata it is now tangible to characterise tissue-specific translatomes in a manner similar to how GTEx characterise transcript expression across tissues (‘The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans’, 2015).

A large amount of valuable sequencing data remains under-utilised post-publication. This can be attributed to a number of factors, such as the requirement for significant computational infrastructure and technical expertise to reprocess datasets of interest. Furthermore, high-quality automated large-scale data processing and analysis efforts are not tractable in the absence of high-quality sample level metadata. Sometimes this information is missing entirely from the sample metadata. Even in cases where the metadata are provided by the submitter, this information is not stored in standardised data fields and ambiguous terms are often used to describe the samples making it difficult to reliably process in an automated manner. The RiboSeq Data Portal aims to remedy these metadata inconsistencies, especially those of published ribosome profiling datasets as some key metadata fields are required for this data type. However, the approaches described in this manuscript could be applied to any other set of published sequencing datasets.

For more than a decade, RiboSeq.Org has been making thousands of published ribosome profiling samples available for visualisation and analysis in the browser to lessen the computational requirements of exploring ribosome profiling data. Furthering this goal, the data portal makes processed datasets available for download to reduce the burden of data reanalysis for the community. Each user who carries out a reanalysis of Ribo-Seq data that they sourced from a public repository will be required to carry out many of the same data processing steps regardless of the downstream analysis they are constructing. These steps are time-consuming and potentially expensive depending on the computational environment used.

Each run of a bioinformatic workflow has not only a financial cost but also a cost to the environment. Many groups conducting the same standard preprocessing and quality control steps contribute to unnecessarily increasing the carbon footprint of such reanalysis projects. Greally et al. (2022) conducted a similar, but more elaborate, workflow for cleaning and conducting quality control for 392 RNA-seq samples and calculated the carbon footprint of the data processing. Preparing these samples for analysis with their pipeline and computational environment was estimated to produce 54.97 kilograms of CO2 equivalent units (kgCO2e). This is stated to be the equivalent of the amount of CO2 that a mature tree would sequester over 5 years or that a car would emit driving 314.11 kilometres. Processing

these files once and subsequently making the files available for use by the community has the potential to greatly reduce the incurred environmental burden of reprocessing data.

In the development of this resource, a number of decisions had to be made that need to be clearly stated as they may be seen as limitations. We endeavour to provide ‘clean’ Ribo-Seq reads that consist solely of the ribosome-protected fragment, however, unique molecular identifiers (UMIs) are not handled as part of the preprocessing pipeline. This will only be an issue for some downstream analysis such as the application of aligners that do not support soft clipping of the alignments. Should a user wish to use an older aligner such as Bowtie (v1) that does not support soft clipping they would need to investigate the needs to handle UMIs themselves. Furthermore, just the top identified adapter is removed from each sample rather than the specific adapter used in the study. The specific adapter cannot be obtained for each sample as this information is very inconsistently reported in the metadata. Users can find the information on the adapter trimming process in the FASTP report that is made available for each sample.

The RiboSeq Data Portal’s metadata cleaning approach described above is ever-adapting and improving to better accommodate existing datasets. Some metadata that has been ‘cleaned’ by the current workflow will inevitably be flawed and will require updating. Furthermore, we are exploring the use of large language models for metadata standardisation which may result in major updates to the metadata published on this platform. To enable users to continue to interact with these updates regular snapshots of the database are made available for reference.

4.6 Conclusions

In this work, we present RiboSeq.Org's latest addition, the RiboSeq Data Portal which hosts well-characterised metadata for a large set of ribosome profiling samples and greatly improves the interoperability of RiboSeq.Org resources. Furthermore, alongside these accurate metadata, the processed data files made available via the data portal improve the tangibility of large-scale data analysis of translation while greatly reducing the computational requirements and environmental costs incurred by reanalysis efforts.

4.7 Methods

4.7.1 Ribo-Seq Study and Sample Identification

Publicly available ribosome profiling studies are identified via a multi-step process that runs monthly to ensure the database is up to date. First, a search is carried out against the NCBI-BioProject database for the keywords: *"Ribosomal footprinting"[All Fields] OR "Ribosome footprinting"[All Fields] OR "Ribosome profiling"[All Fields] OR "Ribo-Seq"[All Fields]*. Then, the SRA Run meta information is downloaded for each study that was identified. This metadata is aggregated and a classification step is performed where individual runs are identified as Ribo-Seq or Not Ribo-Seq. This classification is based on the presence or absence of the following terms in the sample name or sample title: *"ribo", "footprint", "RPF", "RFP", "80S", "fp", "rp", "rf"*. This search process has been iteratively updated to identify as many genuine ribosome profiling samples as possible, however, many instances are likely missed. Where this is identified, we either add the sample(s) manually or extend the search parameters with the new term. Notably, this laborious step could be made unnecessary if data depositories supported a 'RIBOSEQ' library type upon upload rather than forcing inaccurate labelling such as 'RNASEQ' or 'OTHER'.

4.7.2 Metadata Curation

Following the identification steps, we possess a set of Ribo-Seq sample accessions and the associated meta information available on SRA/ENA. This run information consists of core columns that are available and populated for all the samples in the database and variable/open columns that are project- or sample-specific. As variable columns are populated by different individuals for each study, the formatting and content for each vary greatly. This variability in column titles and column contents makes it difficult to standardise the metadata across studies. To address this issue, we designed a controlled vocabulary that consists of standard subject headings and thesauri of terms, which has been iteratively developed to ensure it optimally describes the metadata obtained. This controlled vocabulary enables drastic homogenisation of the metadata which will also aid search and filter functionality down the line.

To improve the controlled vocabulary and to ensure the accuracy of the metadata curations, a manual QC step is carried out where each study metadata is reviewed and corrected if necessary. This introduces a top echelon on Ribo-Seq samples within RiboSeq.Org that have 'verified' metadata.

Scripts for the identification and curation of these metadata can be found on GitHub at https://github.com/Roleren/riboseq_metadata.

4.7.3 File Processing

Our goal was to process and make as much available data as possible for each identified Ribo-Seq sample. The minimal goal is to download and preprocess Ribo-Seq sample FASTQ files to generate collapsed read files and make them available for download in FASTA format. First, for each Ribo-Seq run accession in our metadata database, we attempt to download the SRA file for that run from the Amazon Web Services (AWS) hosted SRA repository using the aws-cli program v1.27.157 (https://aws.amazon.com/cli/). Successfully downloaded SRA files are converted to FASTQ format using fasterq-dump v3.0.5 (51). Where this AWS download fails, the command line application fastq-dl v2.0.0 (52) is used to obtain the FASTQ files from the ENA.

FastQC v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is run on all the obtained read files using a curated list of commonly found adapters in ribosome profiling libraries. After identifying the adapters, the report was preserved. The identified sequence is then trimmed from all reads using FastP v0.23.4 (53) with a minimum length set to 20 nucleotides. The FastP report was obtained in JSON format. The FastP and FastQC reports are compiled into one standalone report with MultiQC v1.12 (54).

Trimmed Ribo-Seq samples then undergo a collapsing step where the set of nonredundant read sequences is identified and the occurrence of each individual sequence is counted. These reads are then written to a FASTA file where the unique read identifier and read count are recorded in the sequence description followed by the sequence of the read. Although this was carried out with a standalone script, this functionality has since been wrapped into a command

line tool, RDP-Tools. The code for RDP-Tools is now available at https://github.com/riboseqorg/RDP-Tools.

The Nextflow pipeline for the download and collapse of Ribo-Seq samples data files can be found on GitHub at https://github.com/JackCurragh/Collapse-FASTQ.

4.7.4 Data Portal Development

The RiboSeq.Org data portal was developed in Python v3.10.10 using the Django framework v4.2 and SQLite3 3.31 databases. It facilitates the search and retrieval of Ribo-Seq studies and samples of interest, maximising the implementation of our controlled vocabulary to return coherent results. The database can be searched and filtered across all metadata fields including the presence of samples on other RiboSeq.Org platforms and availability for download. It is also possible to filter studies that have been manually verified for accuracy by a RiboSeq.Org developer. The RiboSeq Data Portal API was developed using the Django REST Framework (https://www.django-rest-framework.org/).

The source code for this data portal can be found on GitHub at https://github.com/riboseqorg/RiboSeqOrg-DataPortal.

4.8 Author contributions

J.A.S.T, M.S and H.T developed the workflow for metadata cleaning. J.A.S.T, M.S, A.K and G.C fine tuned the controlled vocabularies used for metadata cleaning led by H.T who constructed the initial set. J.A.S.T and A.K developed the web interface, API and database. J.A.S.T developed the pipeline for read file collapsing and produced the collapsed read files and associated reports. As part of RiboCrypts development M.S and H.T produced all BAM and BigWig files made available in the RiboSeq Data Portal. J.A.S.T developed the manuscript while E.V, P.V.B, A.K, G.C, M.S and H.T reviewed and edited the manuscript. J.K, E.V and P.V.B advised and supported the work via the provision of funding.

4.9 References

1. Baudin-Baillieu A, Hatin I, Legendre R, Namy O. Translation Analysis at the Genome Scale by Ribosome Profiling. Methods Mol Biol Clifton NJ. 2016;1361:105–24

2. Andreev DE, O'Connor PBF, Loughran G, Dmitriev SE, Baranov PV, Shatsky IN. Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. Nucleic Acids Res. 2017 Jan 25;45(2):513–26.

3. Ingolia NT, Hussmann JA, Weissman JS. Ribosome Profiling: Global Views of Translation. Cold Spring Harb Perspect Biol. 2019 May 1;11(5):a032698.

4. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009 Apr 10;324(5924):218–23.

5. Lareau LF, Hite DH, Hogan GJ, Brown PO. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. eLife. 2014 May 9;3:e01257.

6. Wu CCC, Zinshteyn B, Wehner KA, Green R. High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. Mol Cell. 2019 Mar 7;73(5):959-970.e5.

7. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity of Mammalian Proteomes. Cell. 2011 Nov 11;147(4):789–802.

8. Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2012 Sep 1;109(37):E2424-32.

9. Meydan S, Marks J, Klepacki D, Sharma V, Baranov PV, Firth AE, et al. Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome. Mol Cell. 2019 May 2;74(3):481-493.e6.

10. Archer SK, Shirokikh NE, Beilharz TH, Preiss T. Dynamics of ribosome scanning and recycling revealed by translation complex profiling. Nature. 2016 Jul;535(7613):570–4.

11. Jan CH, Williams CC, Weissman JS. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. Science. 2014 Nov 7;346(6210):1257521.

12. Williams CC, Jan CH, Weissman JS. Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. Science. 2014 Nov 7;346(6210):748–51.

13. Oh E, Becker AH, Sandikci A, Huber D, Chaba R, Gloge F, et al. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell. 2011 Dec 1;147(6):1295–308.

14. Bohlen J, Fenzl K, Kramer G, Bukau B, Teleman AA. Selective 40S Footprinting Reveals Cap-Tethered Ribosome Scanning in Human Cells. Mol Cell. 2020 Aug 20;79(4):561-574.e5.

15. Wagner S, Herrmannová A, Hronová V, Gunišová S, Sen ND, Hannan RD, et al. Selective Translation Complex Profiling Reveals Staged Initiation and Co-translational Assembly of Initiation Factor Complexes. Mol Cell. 2020 Aug 20;79(4):546-560.e7.

16. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016 Feb;13(2):165–70.

17. Olexiouk V, Van Criekinge W, Menschaert G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D497–502.

18. Brunet MA, Lucier JF, Levesque M, Leblanc S, Jacques JF, Al-Saedi HRH, et al. OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. Nucleic Acids Res. 2021 Jan 8;49(D1):D380–8.

19. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al.

Standardized annotation of translated open reading frames. Nat Biotechnol. 2022 Jul;40(7):994–9.

20. Chothani S, Ho L, Schafer S, Rackham O. Discovering microproteins: making the most of ribosome profiling data. RNA Biol. 2023 Dec 31;20(1):943–54.

21. Prensner JR, Abelin JG, Kok LW, Clauser KR, Mudge JM, Ruiz-Orera J, et al. What Can Ribo-Seq, Immunopeptidomics, and Proteomics Tell Us About the Noncanonical Proteome? Mol Cell Proteomics. 2023 Sep 1;22(9):100631.

22. Stumpf CR, Moreno MV, Olshen AB, Taylor BS, Ruggero D. The Translational Landscape of the Mammalian Cell Cycle. Mol Cell. 2013 Nov 21;52(4):574–82.

23. Chothani S, Adami E, Ouyang JF, Viswanathan S, Hubner N, Cook SA, et al. deltaTE: Detection of Translationally Regulated Genes by Integrative Analysis of Ribo-seq and RNA-seq Data. Curr Protoc Mol Biol. 2019 Dec;129(1):e108.

24. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci. 2009 May 5;106(18):7507–12.

25. Andreev DE, Loughran G, Fedorova AD, Mikhaylova MS, Shatsky IN, Baranov PV. Non-AUG translation initiation in mammals. Genome Biol. 2022 May 9;23(1):111.

26. Chew GL, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Dev Camb Engl. 2013 Jul 1;140(13):2828–34.

27. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, et al. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Rep. 2014 Sep 11;8(5):1365–79.

28. Zhang P, He D, Xu Y, Hou J, Pan BF, Wang Y, et al. Genome-wide identification and differential analysis of translational initiation. Nat Commun. 2017 Nov 23;8(1):1749.

29. Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP, et al.

Improved Ribo-seq enables identification of cryptic translation events. Nat Methods. 2018 May;15(5):363–6.

30. Calviello L, Hirsekorn A, Ohler U. Quantification of translation uncovers the functions of the alternative transcriptome. Nat Struct Mol Biol. 2020 Aug;27(8):717–25.

31. Choudhary S, Li W, D. Smith A. Accurate detection of short and long active ORFs using Ribo-seq data. Bioinformatics. 2020 Apr 1;36(7):2053–9.

32. Oertlin C, Watt K, Ristau J, Larsson O. Anota2seq Analysis for Transcriptome-Wide Studies of mRNA Translation. Methods Mol Biol Clifton NJ. 2022;2418:243–68.

33. Cummins C, Ahamed A, Aslam R, Burgin J, Devraj R, Edbali O, et al. The European Nucleotide Archive in 2021. Nucleic Acids Res. 2022 Jan 7;50(D1):D106–10.

34. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 2022 Jan 7;50(D1):D387–90.

35. Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. 2019 Jan 8;47(D1):D230–4.

36. Michel AM, Fox G, M. Kiran A, De Bo C, O'Connor PBF, Heaphy SM, et al. GWIPS-viz: development of a ribo-seq genome browser. Nucleic Acids Res. 2014 Jan 1;42(D1):D859–64.

37. Fedorova AD, Tierney JAS, Michel AM, Baranov PV. RiboGalaxy: A Galaxy-based Web Platform for Ribosome Profiling Data Processing - 2023 Update. J Mol Biol. 2023 Jul 15;435(14):168043.

38. Kiniry SJ, Judge CE, Michel AM, Baranov PV. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. Nucleic Acids Res. 2021 Jul 2;49(W1):W662–70.

39. Swirski M, Tjeldnes H. RiboCrypt: Interactive visualization in genomics. BioC. 2023;

40. Guo Z, Tzvetkova B, Bassik JM, Bodziak T, Wojnar BM, Qiao W, et al. RNASeqMetaDB: a database and web server for navigating metadata of publicly available mouse RNA-Seq datasets. Bioinformatics. 2015 Dec 15;31(24):4038–40.

41. Kumar P, Halama A, Hayat S, Billing AM, Gupta M, Yousri NA, et al. MetaRNA-Seq: An Interactive Tool to Browse and Annotate Metadata from RNA-Seq Studies. BioMed Res Int. 2015 Aug 25;2015:e318064.

42. Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. Bioinformatics. 2017 Sep 15;33(18):2914–23.

43. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2023 update. Nucleic Acids Res. 2023 Jan 6;51(D1):D1188–95.

44. Michel AM, Kiniry SJ, O'Connor PBF, Mullan JP, Baranov PV. GWIPS-viz: 2018 update. Nucleic Acids Res. 2018 Jan 4;46(D1):D823–30.

45. Tierney JAS, Świrski M, Tjeldnes H, Mudge JM, Kufel J, Whiffin N, et al. Ribosome decision graphs for the representation of eukaryotic RNA translation complexity. Genome Res. 2024 Apr 1;34(4):530–8.

46. Baranov PV, Atkins JF, Yordanova MM. Augmented genetic decoding: global, local and temporal alterations of decoding processes and codon meaning. Nat Rev Genet. 2015 Sep;16(9):517–29.

47. Edgar RC, Taylor B, Lin V, Altman T, Barbera P, Meleshko D, et al. Petabase-scale sequence alignment catalyses viral discovery. Nature. 2022 Feb;602(7895):142–7.

48. Fasterius E, Al-Khalili Szigyarto C. Analysis of public RNA-sequencing data reveals biological consequences of genetic heterogeneity in cell line populations. Sci Rep. 2018 Jul 25;8(1):11226.

49. McCabe A, Zaheed O, McDade SS, Dean K. Investigating the suitability of in vitro cell lines as models for the major subtypes of epithelial ovarian cancer. Front Cell Dev Biol. 2023 Feb 13;11:1104514.

50. Li F, Fang J, Yu Y, Hao S, Zou Q, Zeng Q, et al. Reanalysis of ribosome profiling datasets reveals a function of rocaglamide A in perturbing the dynamics of translation elongation via eIF4A. Nat Commun. 2023 Feb 2;14(1):553.

51. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021 Dec 1;50(D1):D20–6.

52. Petit III RA, Hall MB, Tonkin-Hill G, Zhu J, Read TD. fastq-dl: efficiently download FASTQ files from SRA or ENA repositories [Internet]. Available from: https://github.com/rpetit3/fastq-dl

53. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinforma Oxf Engl. 2018 Sep 1;34(17):i884–90.

54. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinforma Oxf Engl. 2016 Oct 1;32(19):3047–8.

5. Metric-Based Analysis of Ribosome Profiling Data Quality with RiboMetric

Jack A. S. Tierney[1,2] , Lukas Wierdsma, Pavel V. Baranov[1].

[1] School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland.

[2] SFI Centre for Research Training in Genomics Data Science, University College Cork, Cork, Ireland

5.1 Foreword

When large amounts of ribosome profiling data is readily accessible in standardised formats the question moves from ‘How can we get the datasets?’ to ‘Which datasets should we get?’. The quality and characteristics of Ribo-Seq data can vary considerably which can have significant impacts on our ability to draw meaningful conclusions from our analysis. Current data assessment approaches have limited applicability to large-scale studies as they require manual human assessment of quality assessment plots. This chapter introduces RiboMetric, a novel tool for the comprehensive assessment of Ribo-Seq dataset properties. By providing a standardised, metric-based approach to evaluating key features such as periodicity, uniformity, and read length distribution, RiboMetric enables researchers to make more informed decisions when selecting datasets for analysis. This tool complements the resources introduced in earlier chapters, further enhancing our ability to conduct robust and reliable translatomic research.

5.2 Abstract

Background: Ribosome profiling (Ribo-Seq) has become the predominant method for the genome-wide assessment of translation and as a result, its popularity has grown and increased numbers of datasets has been made available in public depositories. Analysis of these datasets have unveiled considerable variability in data quality often attributed to complexities of the nuanced protocol. Existing tools for the assessment of Ribo-Seq data properties do not readily support command-line execution hindering their integration pipelines and furthermore do not enable the comparison of large numbers of datasets simultaneously.

Method: We introduce RiboMetric, a command-line application for the assessment of properties of Ribo-Seq data that formalises metric-based assessment of these dataset properties. Triplet periodicity, profile uniformity, terminal nucleotide bias and the read length distribution are evaluated using a variety of metrics novelly implemented in RiboMetric. These metrics were assessed using both simulated and read data in order to understand their performance. Alongside these metrics, the tool generates comprehensive reports akin to those produced by other quality assessment tools such as FastQC.

Results: RiboMetric is shown to characterise dataset properties using metrics successfully. Datasets exhibiting high and low periodicity and uniformity were effectively identified by the implemented metrics while the terminal nucleotide bias and read length distributions metrics provided more technical insights into aspects of the library preparation steps carried out. This work sets the foundation for future optimisations of metric-based Ribo-Seq data property assessment. By executing RiboMetric on 1,322 human Ribo-Seq samples, we show the potential of applying this method to identify correlations between metrics and potential outliers.

Conclusion: RiboMetric serves as a comprehensive tool for assessing Ribo-Seq data quality via a command-line executable promoting the reproducibility and reliability of Ribo-Seq data. RiboMetric is freely available at https://pypi.org/project/RiboMetric/ and https://github.com/JackCurragh/RiboMetric

5.3 Introduction

Ribosome profiling (Ribo-Seq) is a foundational technique that has revolutionised the study of translation by providing a global snapshot of ongoing translation in a sample via a combination of ribosome arrest, nuclease digestion and the sequencing of the resulting ribosome-protected fragments. Introduced in 2009, Ribo-Seq enables researchers to obtain insights into translation *in vivo* genome-wide at subcodon resolution (1). This allows for a deeper understanding of global changes in translation under various conditions (2). Additionally, Ribo-Seq can be used to identify which regions of RNAs are translated (3–5) and even determine which reading frame they are translated in (6,7).

Before the advent of Ribo-Seq, the extent of the known translation outside of annotated coding regions and the study of genome-wide translation regulation was limited. Mapping the positions of translating ribosomes to the transcriptome has enabled extensive annotation of translated regions (translons (8)), valuable mechanistic insights and further insights into gene expression regulation at the level of translation (9,10).

Variants of the ribosome profiling protocol aim to capture ribosomes during different phases of translation, namely initiation, where the full ribosome complex assembles at a start codon, and elongation, where the nascent peptide is synthesised. The original and most common variant involves the immobilisation of translating ribosomes with cycloheximide which blocks the pre-translocation stage of translation elongation (1,11,12). The alignment of the reads originating from this protocol results in a ribosome profile where the mapped footprints are distributed across the entire translated region. As a result, these datasets are particularly suited to differential translation analysis and the identification of translated regions.

Treatment with lactimidomycin or harringtonine has been shown to enrich the presence of fragments protected by ribosomes that are actively initiating translation (13–15). The presence of these initiating ribosome-protected fragments (RPFs) can be further enriched in the protocol by the prior addition of puromycin as this decreases the number of elongating ribosomes captured (16,17). The resulting profiles from these initiating ribosome profiling protocols show profiles enriched at translation initiation sites (TISs) which has been shown to

be useful for the detection of start codons and for studying start site selection and rates of initiation (18,19).

Various protocol modifications have been made to the original protocol since 2009 to reduce the time, cost and protocol-dependent distortions of the data. The use of different inhibition methods can have significant impacts on the resulting ribosome profiles. The initial protocol called for preincubation of the sample with cycloheximide however this was later deemed unnecessary when cycloheximide was in the lysis buffer (20,21). In *Saccharomyces cerevisiae,* it was noted that the use of cycloheximide to inhibit translation led to an artificial build-up of ribosomes at the beginning of the coding sequence (CDS) (22). However, this bias was not shown to affect cultured human cells (23).

The nuclease digestion step is the next major source of variation in ribosome profiling protocols. The goal of this step is to break down any RNA fragments not protected by a ribosome converting polysomes to monosomes (1). RNase I is the most abundantly used nuclease for this purpose as it results in comparatively low sequence bias and strong 3nt periodicity. However, RNase I has been found to also digest the ribosomal RNA (rRNA) leading to increased abundance of rRNA in the sequenced library. Alternatively, micrococcal nuclease (MNase) has been employed as it better preserves the ribosome structure at the cost of single nucleotide resolution (24). P1 nuclease has recently been proposed as a sequence-independent single-strand nuclease that preserves ribosome integrity better than RNase I (25). Following nuclease digestion the resulting RNA fragments are then size-selected, targeting fragments ~30nt long in order to enrich ribosome-protected fragments, and rRNA depletion to reduce the number of RPF-sized fragments of rRNA being sequenced.

Although not a protocol variation, the size selection step is a major source of variability between datasets. This step involves cutting a section from a gel that is approximately the expected size of the RNA fragments. Depending on the width of the band that is cut from the gel the read length distribution of the dataset can vary greatly.

Library construction is the final variable step in Ribo-Seq protocols. Depending on the approach used in this step various biases can be introduced. For example, ligation-free template switching approaches are reported to have a bias favouring reads with G at their 5’

end. Furthermore, the circularisation of reads has been shown to introduce a ligation bias where the number of reads in the sequenced sample is influenced by ligation adapter affinity for specific bases at read termini. Unique Molecular Identifiers (UMIs) have gained popularity in recent years in order to combat the effects of PCR amplification biases.

Ribosome profiling has inspired the development of a number of related protocols for the investigation of different aspects of translation at a genome-wide level. Scanning ribosomal complexes can be profiled by derivatives of the original Ribo-Seq protocol. Translation Complex Profile Sequencing (TCP-Seq) and Ribosome Complex Profiling (RCP-Seq) involve nuclease digestion following formaldehyde crosslinking (26,27). The resulting 40S and 80S associated fragments are then separated using sucrose gradients. This enables insights into scanning dynamics from the aligned 40S RPFs along with elongating dynamics from the 80S fraction.

Selective TCP-Seq (Sel-TCP-Seq) provides more nuanced insights into scanning ribosome dynamics via the immunoprecipitation of targeted proteins to profile scanning ribosomes that are in complex with specific initiation factors (28,29).

Disome sequencing or Disome-Seq enables the procurement of RPFs that originate from collided ribosomes. Collided ribosomes can be identified as a peak that is heavier than the predominant monosome peak when digested RNA is subjected to sucrose gradient centrifugation (30). The sequencing and subsequent analysis of disome-seq reads provide genome-wide insights into the locations of ribosome collisions.

Sequenced ribosome-protected fragments are commonly aligned against the genome or transcriptome followed by an assessment of the quality of the aligned data prior to further analysis. These quality assessments commonly look at levels of contaminating RNAs, the read length distribution, the bias the aligned reads have to the annotated coding reading frame, and the distribution of reads across mRNAs. Variation in these properties of Ribo-Seq datasets may be introduced as previously mentioned during library preparation but also during data processing.

A number of computational approaches have been developed to carry out checks on Ribo-Seq data however it is also common for these checks to be carried out with custom scripting. For

the most part, these methods focus on the generation of visualisations that a researcher may use to determine the properties of each dataset. Of these tools, just five have been developed with a specific focus on quality assessment. Two of which are published, mQC (31) and RiboStreamR (32), while Ribo-SeQC remains in preprint (33). The other two tools designed specifically for quality assessment of Ribo-Seq data are RiboScore (https://github.com/keanejm/RiboScore) and RibosomeProfilingQC (34) which are available as Python and R packages respectively. Many further tools are not quality assessment specific but support much of the same functionality (see Table 5.1)

Table 5.1 - Ribo-Seq Quality Tools

Tool Name	Tool Type	Code	Reference	Language	Access/Availability
Quality Assessment Tools					
Ribo-SeQC	Quality Assessment	GitHub	Preprint	R	Git Clone
RibosomeProfilingQC	Quality Assessment	Bioconductor	None	R	Bioconductor
RiboStreamR	Quality Assessment	GitHub	Published	R	Git Clone
RiboScore	Quality Assessment	GitHub	None	Python	Git Clone
mQC	Quality Assessment	GitHub	Published	Perl, Python, R	Conda
Tools w/ Quality Assessment Functionality					
Trips-Viz	Browser App	GitHub	Published	Python; Flask	trips.ucc.ie; Git Clone
RiboCrypt	Browser App	Bioconductor	In Prep.	R; R Shiny	ribocrypt.org; Bioconductor
RiboGraph	Browser App	GitHub	Preprint	Python; JavaScript	Git Clone
ORFik	Pipeline	Bioconductor	Published	R; R Shiny	Bioconductor
RiboViz	Pipeline	GitHub	Published	Nextflow	Git Clone; Scripts Provided
RiboToolKit	Pipeline	None	Published	R, Perl, JavaScript	No longer available (Dead link)

Although a large variety of bioinformatic applications are capable of carrying out quality assessment of ribosome profiling data they produce highly similar varieties of quality assessment plots. For the most part, these plots are sufficient for understanding the data however they do not facilitate large-scale comparisons of data properties across large numbers of datasets as found in the RiboSeq Data Portal. RPFdb, a platform for exploring processed Ribo-Seq datasets, notes they make quality reports available for each processed sample (35). However, assessing dataset quality at the database scale is significantly time consuming. We propose that a metric-based assessment of Ribo-Seq data properties will enable this comparison at scale. Metrics for assessing the periodicity of Ribo-Seq data are used in various applications including Trips-Viz and RiboStreamR but not for other Ribo-Seq data properties (32,36).

To address these limitations in existing tools and facilitate large-scale comparisons, we have developed RiboMetric, a comprehensive quality control and analysis tool for ribosome profiling data. RiboMetric generates detailed sample reports and quantitative metrics that assess critical properties of Ribo-Seq datasets, including periodicity, uniformity, terminal nucleotide bias, distribution across mRNA transcripts, and read length distribution. By leveraging these evaluations, RiboMetric enables nuanced quality assessment and facilitates large-scale comparisons of ribosome profiling datasets. In this work, we investigate various metrics for each critical property by assessing their appropriateness and behaviour using thousands of publicly available datasets and simulated datasets. Through case studies and examples, we discuss the scenarios where metrics succeed and fail highlighting their strengths and weaknesses. Based on these findings we make recommendations for metric selection for quality assessment and implement them as default settings in RiboMetric.

5.4 Materials and Methods

5.4.1 Datasets

Processed publicly available Ribo-Seq datasets were sourced from the RiboSeq Data Portal (RDP) from the riboseq.org collection of resources (Chapter 4). These datasets originate from the European Nucleotide Archive or Sequence Read Archive and have undergone standardised preprocessing and genomic alignment. No filtering was applied and all available BAM files of 1386 human Ribo-Seq samples were sourced along with their associated metadata.

We also simulated data scenarios in order to benchmark the metrics and understand their behaviour under edge-case conditions. Rather than simulating entire Ribo-Seq datasets, simulated inputs for the various Ribo-Seq property metric functions were generated.

For periodicity metrics, we exhaustively computed all possible read frame distributions summing to 100, representing the full spectrum from extreme bias to perfect balance. For example:

- (98, 1, 1): Strong bias towards the first frame
- (2, 98, 0): Strong bias towards the second frame
- (33, 33, 34): Approximately equal distribution across frames

Mathematically to simulate metagene profiles from these frame ratios we did:

for each position i in the metagene:

$$C_i = B * R_{(i \bmod 3)} + \varepsilon \quad (1)$$

Where: C_i is the count at position *i,* B is the base count, R_j is the ratio for frame j (j $\in$ {0, 1, 2}) and $\varepsilon \sim N(0, \sigma^2)$: random noise.

For metrics requiring read counts per frame as input, we aggregated these simulated metagene counts by frame. This approach allowed us to test our metrics across a wide range of theoretically possible scenarios, ensuring robust performance under various conditions.

A number of uniformity assessment functions also take coding sequence metagenes as inputs. The aforementioned metagenes were updated to introduce variability across the metagene rather than just in the three frames. To do this, the metagene was split into 6 bins of 20 nucleotides which could then be assigned a weight from 0 to 6 pertaining to how variable the expression of this region would be. We permuted a set of combinations to cover the full range of variability of region uniformity. Examples from this set include (0, 0, 0, 0, 0. 0) and (6, 6, 6, 6, 6, 6). These bin weights were then used to adjust the metagenes accordingly. In order to assess global uniformity the counts in each bin were multiplied by the bin weight resulting in varied degrees of expression across the profiles. In order to assess local uniformity a random selection of positions within each bin were multiplied by the bins weight resulting in varied degrees of random peaks in the profile.

To simulate terminal nucleotide bias, we developed a bias function that modifies dinucleotide frequencies. Let f_i be the initial frequency of dinucleotide i, b be the bias degree parameter, and S be the set of favoured dinucleotides. We apply an exponential bias function:

$$f_i' = f_i * e^{b} \text{ for } i \in S \text{ (favoured dinucleotides)} \quad (2)$$

$$f_i' = f_i * e^{-b} \text{ for } i \notin S \text{ (non favoured dinucleotides)} \quad (3)$$

These frequencies are then normalised to sum to 1. This approach allows us to systematically vary both the number of over-represented dinucleotides and the degree of bias introduced.

Finally, to assess our read length distribution metrics, we simulated diverse Ribo-Seq read length distributions. For each simulated dataset, we randomly selected a mean read length μ from a uniform distribution between 18 and 35 nucleotides (the typical range for Ribo-Seq data). We then chose a standard deviation σ randomly between 1 and 5 to represent varying levels of dispersion. Using these parameters, we generated 1000 read lengths from a normal distribution $N(\mu, \sigma^2)$. To ensure realistic read lengths, we clipped the values to the range [18,

35] and rounded to the nearest integer. This process was repeated 100 times to create a diverse set of simulated distributions. Mathematically, each read length L in a distribution is generated as:

$$L = N(\mu, \sigma^2) \tag{4}$$

where distribution N is clipped so that $18 < L < 35$ and $\mu \sim U(18, 35)$ and $\sigma \sim U(1, 5)$. This approach allowed us to systematically evaluate our metrics across a wide range of plausible Ribo-Seq read length distributions.

5.4.2 RiboMetric Implementation

RiboMetric is a command-line application written in Python 3. Command line argument parsing is handled with argparser and command-line reporting with Rich. BAM file parsing is handled with a combination of PySAM and Oxbow. Annotation file parsing is conducted with GFF Pandas. The majority of the modules and metrics rely on Pandas, Numpy and SciPy. Plots are generated with Plotly which are compiled into a HTML report using Jinja2. PDF reports are generated from HTML reports with xhtml2pdf.

Stable versions of RiboMetric are made available via PyPI at https://pypi.org/project/RiboMetric/.

5.4.3 Quality Control Metrics

Periodicity

Two types of triplet periodicity assessment metrics are implemented in RiboMetric. The first variety looks to quantify the degree to which reads are assigned to one reading frame relative to the others. The first of this variety of metrics is termed the 'Reading Frame Dominance' metric and is calculated as the proportion of reads the map to the main reading frame:

$$Periodicity_{Dominance} = \frac{max(F_0, F_1, F_2)}{(F_0 + F_1 + F_2)} \tag{5}$$

Where F_0, F_1, F_2 are the number of reads mapping to each of the three reading frames.

The next is termed the ‘Reading Frame Score - Trips-Viz’ as this is the method of periodicity assessment implemented in Trips-Viz. This metric is the difference between the top supported reading frame and the second top reading frame as a proportion of the top reading frames support.

$$Periodicity_{Trips} = 1 - \frac{F_1 - F_2}{F_1} \quad (6)$$

Where F_1 is the number of reads in the most supported frame and F_2 is the number of reads in the second most supported frame.

The third metric of this variety is introduced as a novel metric and is described as the information content of the read frame proportions of a given region. The score is calculated as the square root of the "bits of information" in the read frame distribution. The square root is used to normalise the score so that it scales better between 0 and 1. This is done by first calculating the difference between the maximum Shannon entropy of the triplet distribution and the Shannon entropy of the given probability distribution. It is calculated using the formula:

$$Periodicity_{Info} = \sqrt{\frac{log_2 3 - \sum_{i=0}^{2} p_i \, log_2(p_i)}{log_2 3}} \quad (7)$$

Where p_i is the proportion of reads in frame *i.* The score is then calculated as the square root of the difference between the maximum entropy and the actual entropy, divided by the maximum entropy. This gives a value between 0 and 1, where higher values indicate a stronger triplet periodicity in the input probabilities. A higher $Periodicity_{Info}$ score indicates that the distribution of reads among the three possible frames is more regular and less random, suggesting a higher degree of periodicity.

The second variety of triplet periodicity assessment metrics are spectral metrics that look to detect the periodic nature of signals. In this case, we take the frequency of A-site counts across annotated coding sequences as the signal and use these approaches to decompose the signal into its constitutive frequencies. The first of these metrics is a Fourier transform-based metric used to quantify the strength of the triplet periodicity signal in the metagene profile. This metric leverages the Fourier transform, which decomposes a signal into its constituent frequencies. The key idea is that if a strong triplet periodicity exists in the metagene profile, it should manifest as a prominent frequency component at 1/3 cycles per position, corresponding to a period of 3 nucleotides. We quantify the strength of triplet periodicity using a Fourier Transform-based metric:

$$Periodicity_{Fourier} = 1 - \left(\frac{A_3}{\Sigma A_i}\right) \quad (8)$$

Where A_3 is the amplitude of the Fourier coefficient corresponding to a frequency of 1/3 cycles per position $\Sigma\ A_i$ is the sum of amplitudes of all Fourier coefficients. This metric ranges from 0 (no periodicity) to 1 (perfect periodicity), with higher values indicating stronger triplet periodicity.

Second, an autocorrelation-based periodicity metric is another approach to quantify the strength of the triplet periodicity signal in the metagene profile. This metric relies on the autocorrelation function, which measures the similarity between a signal and a lagged version of itself. The key idea is that if a strong triplet periodicity exists in the metagene profile, the autocorrelation function should exhibit a prominent peak at a lag of 3 nucleotides.

We also employ an autocorrelation-based metric to assess triplet periodicity:

$$Periodicity_{auto} = \frac{(R_3 - \bar{R})}{\bar{R}} \quad (9)$$

Where R_3 is the autocorrelation value at lag 3 and $\bar{R}$ is the mean autocorrelation value (excluding lag 0).

This metric can range from negative values (anti-periodic signal) to positive values, with higher positive values indicating stronger triplet periodicity. A value of 0 suggests no apparent periodicity.

Codon Coverage Uniformity of Metagenes

Autocorrelation-based metrics are also used in RiboMetric to quantify the uniformity of the Ribo-Seq signal in a similar method to the above. This uniformity metric leverages the autocorrelation function to measure the self-similarity of the Ribo-Seq signal across different lags. A highly uniform signal should exhibit high autocorrelation values across all lags, as it lacks significant periodic patterns or fluctuations. Conversely, a non-uniform signal with strong periodic components or localised peaks will have lower autocorrelation values at certain lags. In order to avoid a low score due to triplet periodicity alone this is calculated based on codon counts rather than nucleotide counts.

First, the Ribo-Seq metagenes counts per codon are calculated. The autocorrelation of these counts is calculated as the cross-correlation between this sequence and itself. The uniformity score is then calculated as the mean of the first 5 lags of the cross-correlation.

This score ranges from 0 to 1 where 1 represents the perfectly uniform sequence. The uniformity metric provides a complementary perspective to the periodicity metric, as it quantifies the overall smoothness and lack of periodic patterns in the Ribo-Seq signal, rather than specifically focusing on the triplet periodicity component.

We also include an entropy-based metric that assesses the degree of randomness in the samples coding metagene. A sample with maximum entropy across codon counts is considered to be perfectly uniform. Again, codon counts are used in order to prevent the degree of periodicity impacting the uniformity score. For each read length, we calculated the entropy of the proportion of codon counts per position across the coding metagene and normalised these values by the maximum possible entropy returning an entropy-based uniformity score between 0 and 1.

$$Uniformity_{entropy} = \frac{-\sum_{i=0}^{i=n} p_i \log_2(p_i)}{\log_2(n)} \quad (10)$$

Where p_i is the proportion of A-site counts at codon *i* of the coding metagene profile and *n* is the number of codons in the metagene.

Alternatively, we look at two inequality-based metrics to assess the uniformity of the metagene profile of a sample. These metrics are commonly used to assess economic inequality that is commonly used in biology (37,38) but we repurpose them to assess the non-uniformity of a ribosome profile. Firstly the Gini coefficient, which is a measure of the statistical dispersion within a distribution is used to represent the inequality of a profile. The Gini coefficient assesses inequality, so a maximum score of 1 indicates a non-uniform profile and a minimum score of 0 indicates perfect uniformity. In order to apply this coefficient to a ribosome profile we consider each codon in the metagene as an 'individual' and based on the density of Ribo-Seq reads we attribute them 'wealth'. The Gini coefficient is then calculated as the ratio of the area between the line of perfect uniformity and the observed distribution curve over the area under the line of perfect uniformity. Finally, in order to make this a metric of uniformity rather than non-uniformity we invert the value.

$$Uniformity_{Gini} = \frac{\sum_i x_i (2i - n - 1)}{n \sum_i x_i} \quad (11)$$

Where x_i is the counts at position *i* in a list of all counts across the metagene sorted in ascending order and *n* is the number of positions in the metagene.

The second metric taken from the study of economic inequality is the Theil index which is a value also derived from the concept of entropy. It is the sum of the products of the proportion of Ribo-Seq reads at each codon position and the logarithm of the reciprocal of this proportion.

We calculate a uniformity metric-based on the Theil index, originally developed to measure economic inequality:

$$Uniformity_{Theil} = 1 - \left(\frac{1}{n\,ln(n)}\right)\sum p_i ln(n * p_i) \quad (12)$$

Where n is the number of codons in the metagene profile and p_i is the proportion of reads at codon position i. This metric ranges from 0 (complete non-uniformity) to 1 (perfect uniformity), with higher values indicating more uniform distribution of reads across codon positions. This metric differs from $Uniformity_{Entropy}$ as the argument of the logarithm function is the reciprocal of that codon's proportion. This results in greater weight being given to the less frequently occupied codons.

Terminal Nucleotide Bias

Terminal nucleotide biases are calculated for both 5' and 3' read termini as preferential cleavage or ligation may occur at one terminus and not the other. From the read sequences provided in the BAM file we record just the dinucleotide frequencies observed in order to limit the memory requirements of the application. In order to obtain expected dinucleotide frequencies without requiring the provision of reference transcript sequences RiboMetric records the dinucleotide frequencies across each entire unique Ribo-Seq read provided. The first terminal nucleotide bias metric is simply the complement of the maximum difference in observed and expected proportion of any dinucleotide observed at the read terminus. A higher value of this metric indicates that there is an overrepresentation of at least one dinucleotide at that terminus.

$$Terminal\ Nucleotide\ Bias_{Max\ Abs} = max_i \left|O_{ij} - E_{ij}\right| \quad (13)$$

Where i represents the read terminus (5' or 3') and O_{ij} denotes the observed frequency of dinucleotide j at terminus i. E_{ij} represents the expected frequency of dinucleotide j at terminus i.

Alternatively, the Kullback-Leibler divergence is calculated as a measure of dissimilarity between the observed and expected distributions of dinucleotide frequencies at each terminus using the same observed and expected frequency values as specified above.

$Terminal\ Nucleotide\ Bias_{Max\ Abs}$ highlights the most significant discrepancy while the $Terminal\ Nucleotide\ Bias_{KL\ Divergence}$ is a more comprehensive measure of the dissimilarity across all dinucleotides.

$$Terminal\ Nucleotide\ Bias_{KL}\left(O_i \,||\, E_i\right) = \sum O_i log\frac{O_{ij}}{E_{ij}} \quad (14)$$

Where the terms used match equation 13 above.

Read Length Distribution

It is commonly stated that Ribo-Seq reads are expected to be 28-32 nucleotides in length however it is also acknowledged that depending on the components of the ribosomal complex, ribosome conformation, and laboratory protocol the expected read length distribution may change. As a result, a wide variety of metrics are required in order to attempt to characterise the potential read length distributions observed in Ribo-Seq data. In RiboMetric these read length distribution metrics can be classified as either assessing dispersion or shape of the distribution.

The first dispersion-focused metric is the coefficient of variation of the read length distribution which enables an assessment of the dispersion relative to the mean. This metric is the standard deviation of the data divided by the mean.

$$Read\ Length\ Distribution_{CV} = \frac{\sigma}{\mu} \quad (15)$$

Where μ is the mean read length of the sample and σ is the standard deviation.

The second dispersion metric is an interquartile range normalised metric that aims to assess the dispersion of the read length distribution while accounting for potential outliers. This metric is the ratio of the interquartile range to the range between the 10th and 90th percentiles.

$$Read\ Length\ Distribution_{IQR\ Normalised} = \frac{Q3 - Q1}{P_{90} - P_{10}} \quad (16)$$

Where *Q1* and *Q3* are the first and third quartiles respectively and P_{10} and P_{90} are the tenth and ninetieth percentiles.

The first shape related metric is the proportion of reads in the sample in the most common read length. The relative height of the peak of the distribution gives important information of the dominance of the most frequent read length which can be taken in conjunction with other read length distribution properties to gain insights about a given dataset.

$$Read\ Length\ Distribution_{Max\ Prop} = \frac{max(RLD)}{\sum RLF} \tag{17}$$

The normality of the read length distribution is tested using the D'Agostino-Pearson test using the scipy implementation. This test is a combination of the outputs of skewness and kurtosis tests.

$$Kurtosis = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{18}$$

$$Skewness = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 \tag{19}$$

Where *n* is the number of read lengths, $\bar{x}$ is the mean read length frequency and *s* is the standard deviation.

$$Read\ Length\ Distribution_{normality} = \frac{S^2 + K^2}{6} \tag{20}$$

Where *S* is the skewness metric and *K* is the kurtosis metric.

Finally, we also calculate a bimodality metric that is also a combination of the same skew and kurtosis metrics. Bimodality in read length distribution of Ribo-Seq data is not uncommon and this property can originate from a number of sources.

$$Read\ Length\ Distribution_{bimodality} = \frac{(S^2+1)}{K+3\left(\frac{(n-1)^2}{(n-2)(n-3)}\right)} \quad (21)$$

5.4.4 Data Processing

A Nextflow pipeline was developed to enable the application of RiboMetric to the genomically aligned Ribo-Seq samples from the RiboSeq Data Portal which can be found at https://github.com/JackCurragh/RiboMetric-nf. This pipeline takes the genomically aligned read files in BAM format and transcriptome annotation files in GTF format. It first converts genomic BAMs to transcriptomic BAMs with Mudskipper (https://github.com/OceanGenomics/mudskipper) and then prepares the transcriptomic BAMs for RiboMetric by sorting and indexing it with SAMtools (39). The aggregation of RiboMetric outputs and subsequent analyses can also be found in the RiboMetric-nf repository.

5.4.5 Implementation and Reproducibility

Comparative analysis of RiboMetric outputs was carried out using Python, and the Plotly and Sci-Kit Learn packages. All scripts required to reproduce the analysis of RiboMetric outputs can be found on GitHub. Simulated input data-based comparisons can be found at https://github.com/JackCurragh/Metric-Analysis. The notebook for the analysis of the human dataset RiboMetric outputs can be found in the aforementioned pipeline repository (https://github.com/JackCurragh/RiboMetric-nf).

5.4.6 Metric Decile Calculation

In order to facilitate the interpretation of RiboMetric results in the context of other real datasets we provide a decile categorisation for each resulting score. The decile cut-offs were calculated based on maximum-minimum normalised distributions of RiboMetric results calculated on the 1,322 human Ribo-Seq datasets to produce 10 equisized deciles. As more of the RiboSeq Data Portal is populated with RiboMetric reports this will be extended to include metrics from other organisms bolstering the applicability of these deciles.

5.5 Results

5.5.1 RiboMetric

The primary result presented in this work is RiboMetric, a command-line bioinformatic application for assessing various properties of Ribo-Seq data written in Python. RiboMetric takes transcriptome aligned Ribo-Seq data in BAM and "collapsed BAM" formats. Collapsed BAMs are a version of a BAM file where the number of occurrences of each unique read in a sample is recorded in the reads name. Thus allowing for each read to be stored in the BAM file just once. To our knowledge, this was first implemented in RiboToolKit (40). To enable the generation of the full suite of reports and metrics the user must also provide a transcriptome annotation file in GTF or GFF3 format.

Following the parsing of RiboMetrics input files and configuration parameters a number of modules are run that prepare the data for the generation of the metrics and the generation of the various reports. A key module in this process is the prediction of A-sites of Ribo-Seq reads by change point analysis as seen in Shoelaces (41). However, RiboMetric supports the provision of per-read length offsets in tab-delimited format to enable the user to run other offset calculating tools instead.

The primary output of RiboMetric is an interactive report in HTML format which can also be exported as a PDF. This report contains various quality assessment plots as well as a summary table containing the metrics calculated for that sample. The outputs of the aforementioned modules can be outputted in JSON format to enable the regeneration of reports using different plotting parameters without rerunning the full workflow again.

As the majority of tools that carry out quality assessment of Ribo-Seq data do so as part of a larger workflow, they often require specific preprocessing steps (load ORFik Experiment) or file formats (.ribo) (42–44). Although these are not necessarily negative points against the tools themselves they may be seen as limitations in terms of their implementation within modular workflows such as those developed by the nf-core community, without further development (45). RiboMetric aims to fill this functionality gap as it is designed to take common file formats via standard command-line arguments.

5.5.2 Assessment of Metrics for Data Properties

There are a wide range of conceptually viable metrics that may be used to assess each property of ribosome profiling data that may be of interest to a researcher. During the development of RiboMetric, we aimed to implement a diverse variety of each metric type in an effort to get a comprehensive picture of what describes each property in a representative manner. It was not feasible nor sensible to attempt to implement all possible metrics. Here, we explore the performance of each implemented metric using both real human data and simulated inputs.

Periodicity

As described previously, the periodic nature of ribosome profiling data is assessed using a number of metrics. Namely, these are $Periodicity_{max}$, the proportion of reads in the highest frame, $Periodicity_{fourier}$, a fourier transform based metric, $Periodicity_{auto}$ an autocorrelation based metric, $Periodicity_{info}$ an information content based metric and $Periodicity_{Trips}$ which is based on the proportion of reads in each reading frame and originates from the Trips-Viz browser (36,46) (for more details on all metrics see Materials and Methods).

When these metrics are evaluated using the simulated metadata there is a clear correlation with the proportion of reads in frame 0 indicating they are all indeed measuring related properties, however, there are key differences between metric types. The metrics concerned with read counts per frame ($Periodicity_{max}$ and $Periodicity_{Trips}$) and metrics concerned with signal frequency ($Periodicity_{Auto}$ and $Periodicity_{Fourier}$) look to capture different aspects of the data. Signal frequency metrics capture the true strength or amplitude of the periodic signal whereas read count per frame metrics capture the bias towards the coding frame, termed ‘Frame Dominance’. These features of the data are related but differ as exemplified in Figure 5.1.

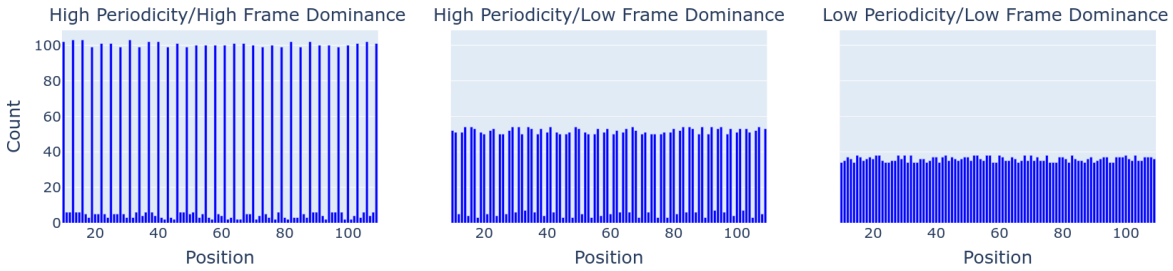


Figure 5.1: Periodicity Versus Frame Dominance. Simulated ribosome profiles with varied degrees of periodicity and frame dominance. **Left:** High periodicity and high frame dominance. Majority of the signal comes from one frame at regular intervals of 3. **Middle:** High periodicity but low frame dominance. Signal adheres strongly to 3nt repetitions but no one frame dominates. **Right:** Low frame dominance and low periodicity. The signal is not biassed to any one frame and does not repeat at regular intervals.

Simulated ribosome profiles (as described in Materials and Methods - Datasets) were used to assess the performance of each periodicity metric under varying degrees of both periodicity and frame dominance. The strongest correlations between metrics are seen within the signal frequency metrics subgroup. These metrics have perfect Spearman correlations meaning that samples sorted by these metrics would have the exact same order.

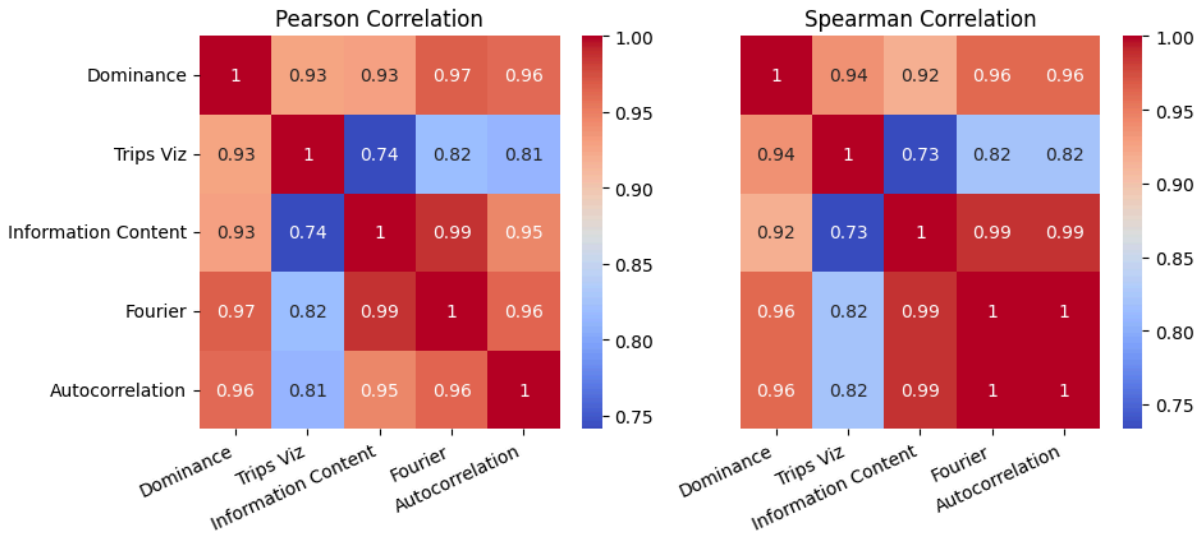


Figure 5.2: Pearson and Spearman Correlations of Periodicity Metrics of Simulated Ribosome Profiles. Left: Pearson correlations show strong positive linear correlations between all metrics, indicating that the metrics broadly agree on profile periodicity. Heatmap colouring ranges from ~0.75 to 1.00, with particularly high correlations between Dominance, Information Content, Fourier, and Autocorrelation. **Right:** Spearman correlations reveal that signal frequency metrics (Fourier and Autocorrelation) have perfect monotonic concordance, as evidenced by their correlation coefficient of 1.00. Additionally, while most correlations remain strong, Spearman correlations are generally slightly lower than Pearson correlations for some pairs, particularly between Trips Viz and other metrics. Heatmap colouring ranges from ~0.75 to 1.00.

The $Periodicity_{Trips}$ metric was developed to maximise the clarity of plotted ribosome profiles in the Trips-Viz transcriptome browser. This focus on visualisation led to the $Periodicity_{Trips}$ metric being the most dissimilar to all other metrics, as reflected in both the Pearson and Spearman correlation matrices. In response to the limitations of the $Periodicity_{Trips}$' score, the Information Content metric was subsequently developed to address these shortcomings. The goal of the Information Content metric was to provide a more accurate measure of ribosome profile periodicity while maintaining some level of visual clarity. This refinement is evident in the correlation matrices, where the $Periodicity_{Info}$ metric shows stronger alignment with other periodicity metrics, particularly $Periodicity_{Fourier}$ and $Periodicity_{Auto}$.

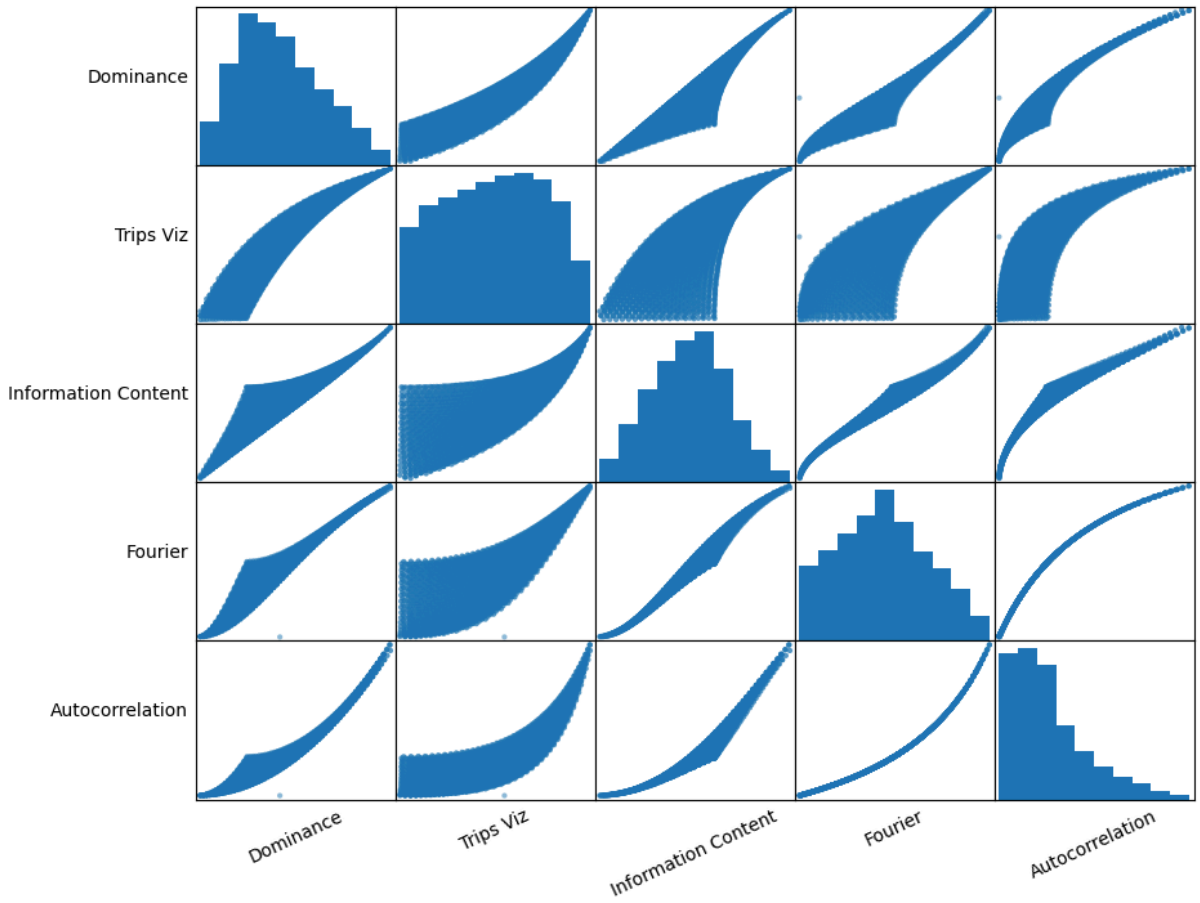


Figure 5.3: Pairwise Scatter Plot and Histogram Matrix of Periodicity Metrics of Simulated Ribosome Profiles. Figure shows a pairwise scatter plot and histogram matrix of the periodicity metrics. The plots on the diagonal show the distribution of each individual metric, showing each metric's frequency and spread. The others are scatter plots of each pairwise combination of the metrics. These scatter plots highlight the relationships between the metrics, with nonlinear patterns visible in several pairs, particularly those involving the Trips Viz metric. The curvature in these scatter plots indicates nonlinear relationships and differences in how each metric quantifies periodicity in ribosome profiles. Each metric has been Max-Min scaled so all axes range from 0 to 1.

All subplots of the scatter matrix show a degree of curvature as only metrics with both Pearson and Spearman correlations equal to 1.00 would fall directly on a line of equality ($y=x$). The $Periodicity_{Trips}$ scatter plots show the greatest degree of curvature as a result of its unique bias towards profiles that would create clear single transcript plots. $Periodicity_{Trips}$ scatter plots also have a wider spread. This category is referred to as this

reflects the fact that the $Periodicity_{Trips}$ scores profiles that have a single dominant frame higher even though this is not necessarily an intrinsic property of periodic signals. As a result, there are many instances where signal frequency metrics score profiles with low Trips-Viz scores much higher.

Based on the simulated data that contains a comprehensive range of potentially valid metagene profiles, all 5 periodicity metrics correlated well with our perception of 3nt periodic signal.

Uniformity

The uniformity of ribosome profiling data is commonly discussed as a property of the dataset often linked to sample quality. Changes in the uniformity of the Ribo-Seq signal can be attributed to a variety of both technical and biological factors. The uniformity of a ribosome profile can be considered in two ways, global and local uniformity.

The global uniformity of the Ribo-Seq signal refers to the pattern of expression across translated regions. Due to differences in translation elongation rates across different sections of coding sequences a gradual 'ebb and flow' of Ribo-Seq signal is typically observed. This uniformity is not necessarily a property of either high or low quality data although outliers may be suspect.

Local uniformity of Ribo-Seq signal refers to the presence or absence of significant peaks in the profile localised to specific codons. These sharp and random peaks may occur due to ribosomal stalling or pausing that may be of biological significance, however, these may also occur as technical artefacts such as PCR duplicates that were not properly handled.

Uniformity metrics implemented in RiboMetric can be grouped into two categories. The first of these are inequality based metrics that originate from the field of economics and the study of wealth inequality. Namely, these are the Gini Index ($Uniformity_{Gini}$) and Theil Index ($Uniformity_{Theil}$). The Gini Index is a popular metric for the inequality of a distribution that is typically sensitive to changes in the middle of the distribution. Whereas the Theil Index is based on entropy and is known to be more sensitive to changes in values at the tails (i.e very

high or very low expression at a subset of positions). The second category of uniformity metric approximates 'randomness' and include an autocorrelation based metric ($Uniformity_{Auto}$) and an entropy based metric ($Uniformity_{Entropy}$).

Initial assessment of selected simulated metagenes show all uniformity metrics score the uniformly distributed metagene profiles highly (Figure 5.4). However, Theil index and the entropy metric score all three selected profiles relatively highly. The autocorrelation metric best differentiates between the most and least globally uniform profiles followed by the Gini index. In the context of these selected profiles from this simulation the Gini and Theil indices were not likely to score any profile particularly lowly as the use of bins ensures there is not set of codons that are expressed significantly unequally.

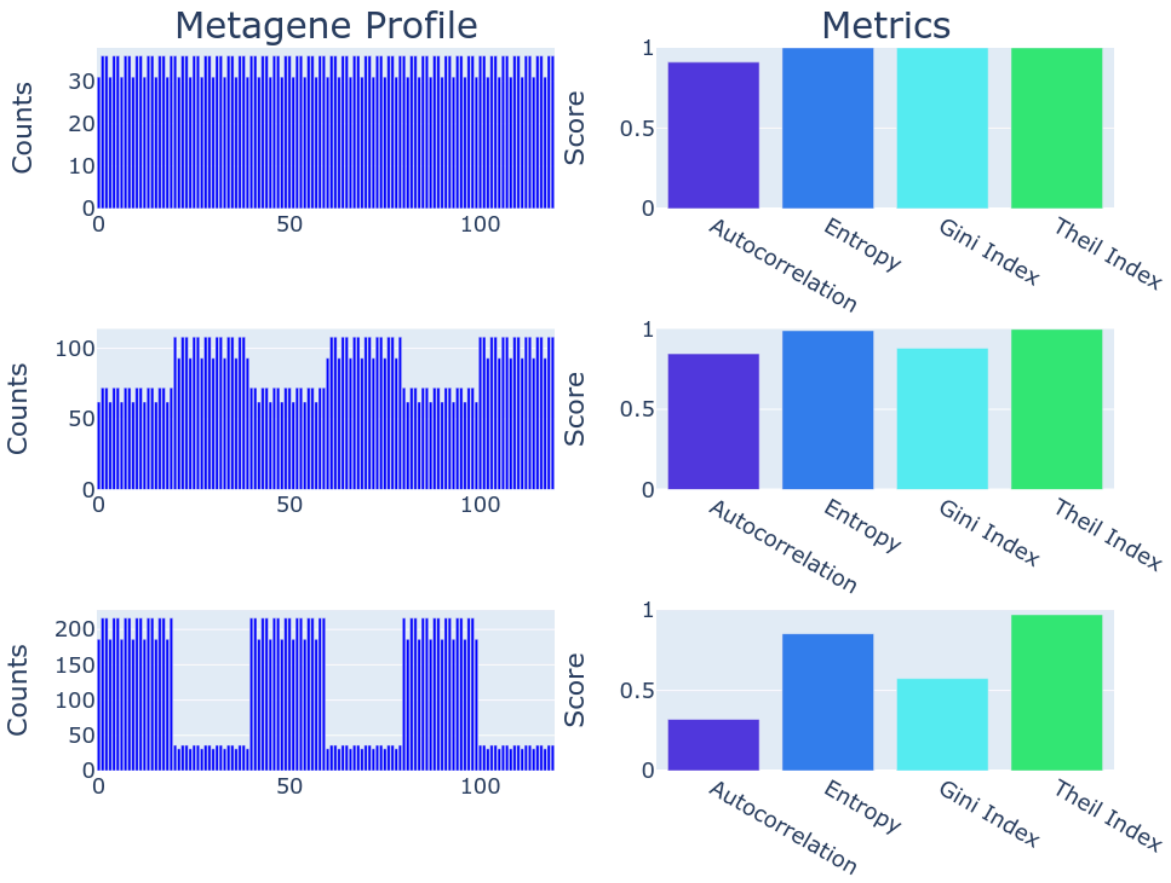


Figure 5.4: Performance of uniformity metrics on global uniformity. **Left:** Simulated metagene profiles ordered from top to bottom by the degree of global uniformity in

decreasing order. Simulated counts are shown on the y-axis and position in nucleotides on the x-axis. The reads of all three profiles are distributed across the three reading frames at a ratio of 30:35:35. **Right:** Uniformity scores corresponding to the simulated metagene profile in the same row. Scores are max/min normalised based on the simulated data described in Materials and Methods.

Consistent with these initial observations the correlations between all uniformity metrics on the simulated dataset for the assessment of global uniformity show that $Uniformity_{Entropy}$ and $Uniformity_{Theil}$ have the strongest correlation. $Uniformity_{Auto}$ correlates poorly with the other three metrics indicating that, at least for these simulations, $Uniformity_{Auto}$ is the only metric that captures the variability in global uniformity of Ribo-Seq profiles.

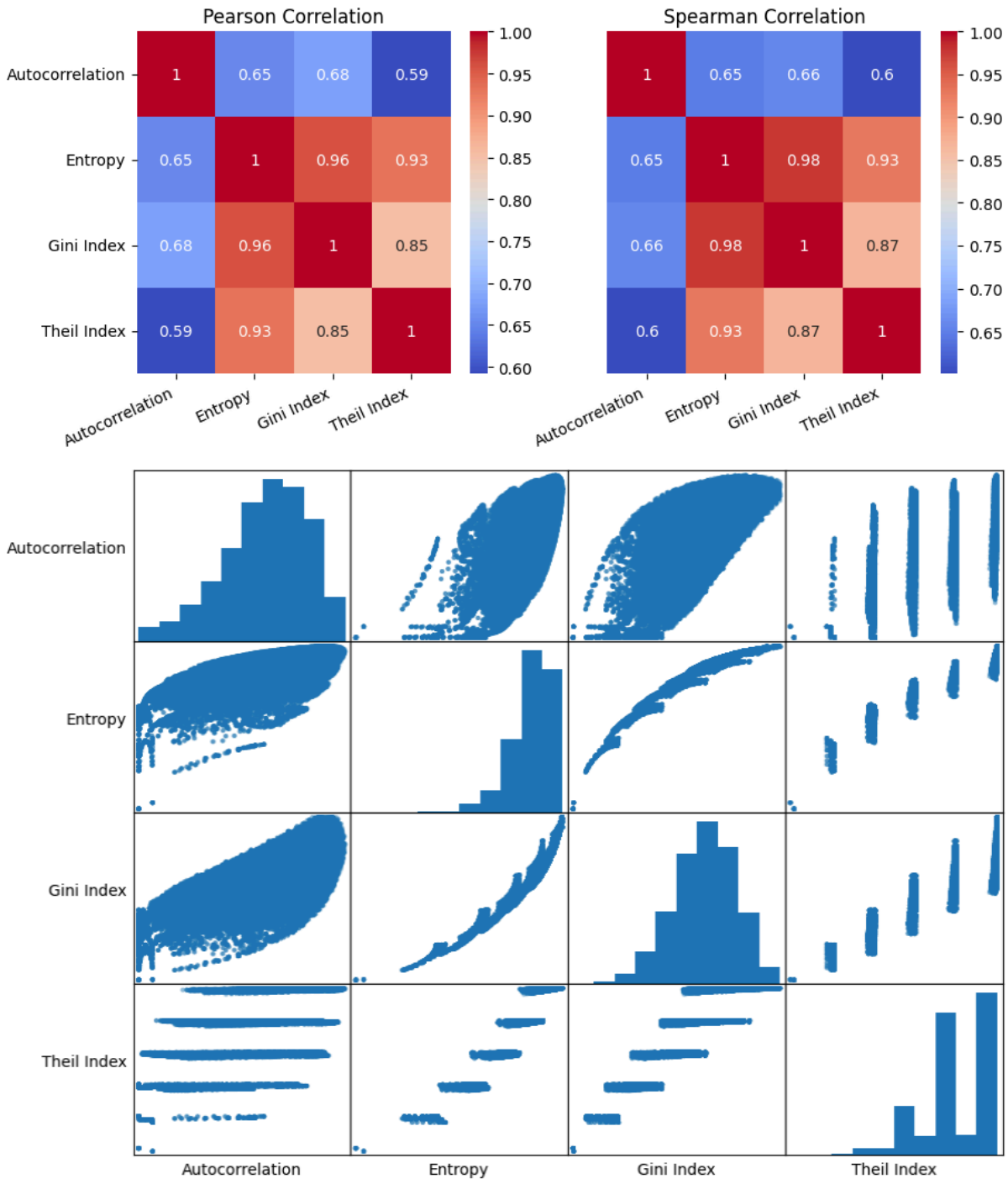


Figure 5.5: Correlations of uniformity metrics on simulated data with varying global uniformity. **Top left:** Pearson correlation matrix. **Top right:** Spearman correlation matrix. **Bottom:** Scatter matrix.

Contrastingly, all uniformity metrics implemented in RiboMetric capture changes in local uniformity in selected simulated Ribo-Seq profiles well. The most uniform profile is scored highly by all metrics and the least uniform lowly by all metrics.

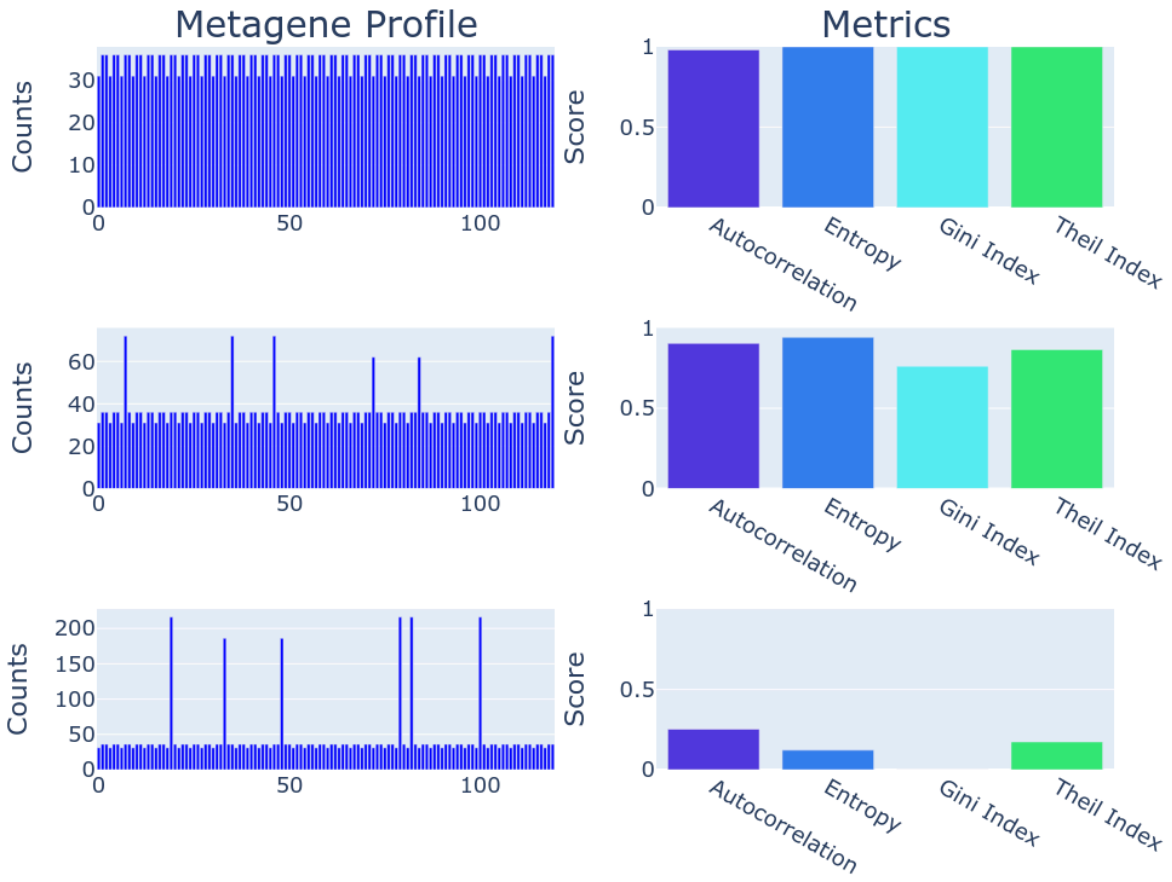


Figure 5.6: Performance of uniformity metrics on local uniformity. **Left:** Simulated metagene profiles ordered from top to bottom by the degree of local uniformity in decreasing order. Simulated counts are shown on the y axis and position in nucleotides on the x axis. The reads of all three profiles are distributed across the three reading frames at a ratio of 30:35:35. **Right:** Uniformity scores corresponding to the simulated metagene profile in the same row. Scores are max/min normalised based on the simulated data described in Materials and Methods.

However, the correlations between these metrics across the full local uniformity simulated dataset show that the Theil index correlates poorly with the other metrics implying that the selected profiles that were individually examined are not representative of the profiles that the Theil index considers uniform or not. The other three metrics do correlate well. Combined with the individually examined examples this implies the Theil index is not a good metric for the assessment of this property .

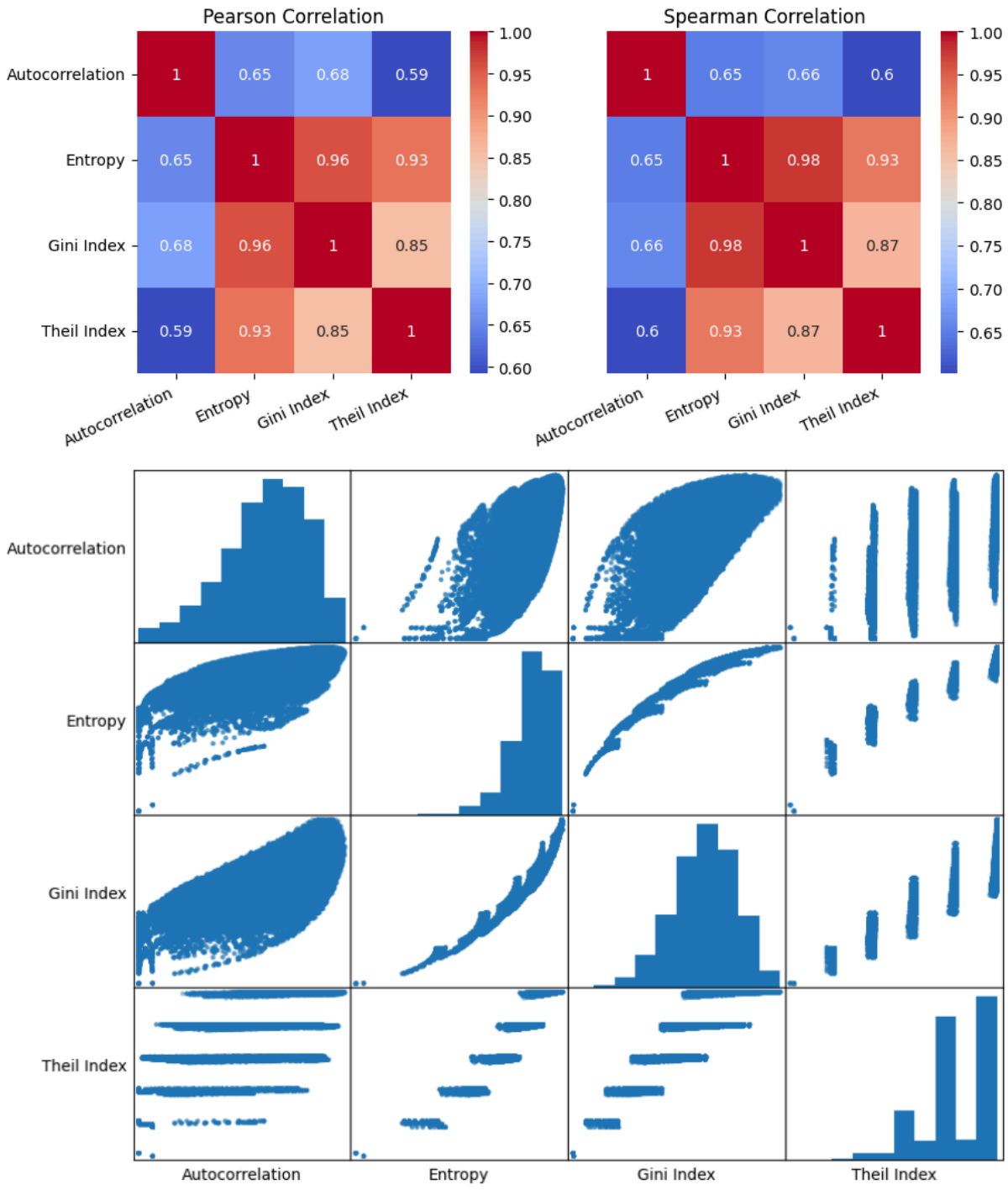


Figure 5.7: Scatter matrix of uniformity metrics on simulated data with varying local uniformity. **Top left:** Pearson correlation matrix. **Top right:** Spearman correlation matrix. **Bottom:** Scatter matrix.

Terminal Nucleotide Bias

Terminal nucleotide bias refers to the over-representation of specific nucleotides at the end of Ribo-Seq reads. There are a number of potential technical origins of this bias including the preferential ligation of certain ligases or selective digestion by certain nucleases. These technical biases can lead to the overrepresentation of certain reads in samples which may need to be corrected prior to downstream analysis in order to prevent the confounding of results (47). Two metrics are implemented in RiboMetric to assess this property of a given dataset. A naive metric, termed $Terminal\ Nucleotide\ Bias_{Max\ Abs}$ looks to assess to what degree the dinucleotide with absolute highest representation relative to its expected frequency is represented. A more nuanced metric, $Terminal\ Nucleotide\ Bias_{KL}$ calculated the Kullback Leibler divergence between the observed dinucleotide frequency distribution and the expected dinucleotide frequency distribution.

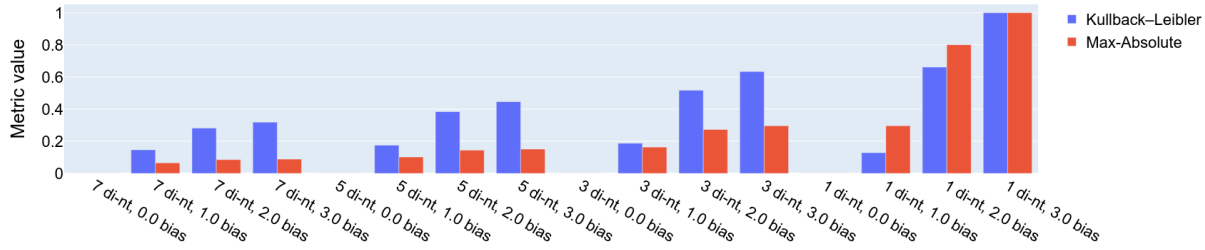


Figure 5.8: Terminal Nucleotide Bias scores across different degrees of bias and numbers of overrepresented dinucleotides. Grouped bar chart compares two Terminal Nucleotide bias metrics, Kullback-Leibler (blue) and Max Absolute (red). X-axis shows combinations of differing numbers of overrepresented dinucleotides (1nt, 3nt, 5nt and 7nt) and varying degrees of bias (0.0, 1.0, 2.0 and 3.0).

Data simulated to assess these two metrics vary the number of biassed dinucleotides and the degree to which these are biassed. Figure 5.8 shows that both metrics correlate with the degree of bias regardless of the number of biassed dinucleotides. However, the $Terminal\ Nucleotide\ Bias_{Max\ Abs}$ metric is less sensitive when a greater number of

dinucleotides. $Terminal\ Nucleotide\ Bias_{KL}$ effectively differentiates between the varying degrees of terminal dinucleotide biases even when ~50% of the dinucleotides are biassed making it a suitable metric for measuring this property of ribosome profiling samples.

Read Length Distribution

As previously discussed, the distribution of read lengths in a Ribo-Seq sample can provide valuable insights into the properties of that sample. The read length distribution metrics implemented in RiboMetric look to capture a variety of aspects of this distribution with the hope that a combination of these metrics will provide a clear picture of the distribution from metrics alone. The first aspect of the distribution is the dispersion or spread of the observed read lengths. Depending on the size of the band cut from the gel during the preparation of the Ribo-Seq sample the expected range of read lengths will vary. This dispersion is measured by $Read\ Length\ Distribution_{CV}$, $Read\ Length\ Distribution_{IQR\ Normalised}$ and $Read\ Length\ Distribution_{MaxProp}$.

The shape of the read length distribution is typically near-normal due to the central limit theorem as the variability in enzymatic digestion of the RNA introduces a small random effect to each read length. However, the conformation of the ribosome at the time of nuclease digestion may result in the presence of different populations of ribosome-protected fragments in the sample with significantly different read lengths. As a result it is not uncommon for a bimodal read length distribution to be observed. The shape of the distribution is assessed by the $Read\ Length\ Distribution_{Normality}$ and $Read\ Length\ Distribution_{Bimodality}$ metrics.

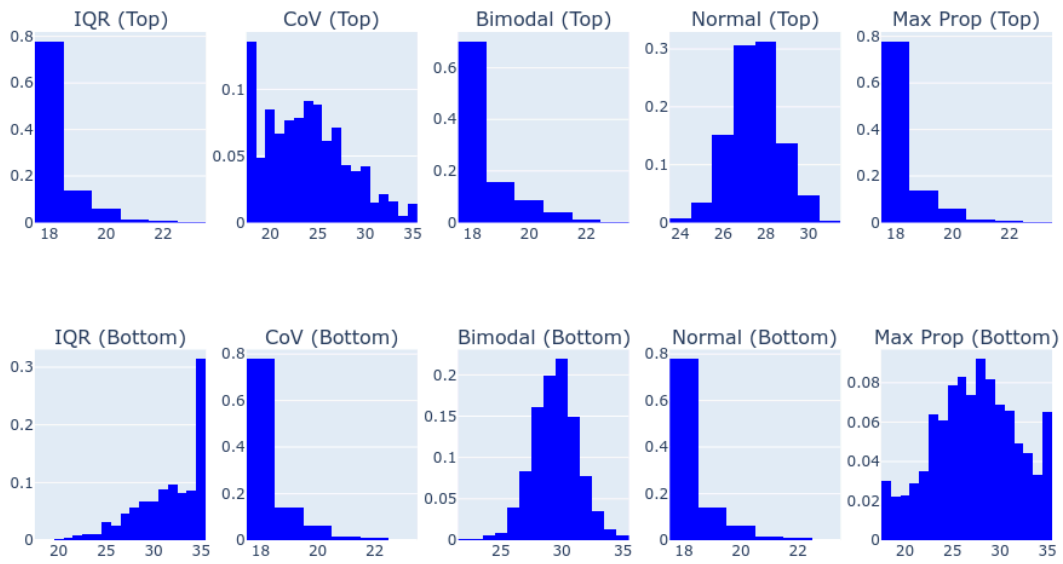


Figure 5.9: Extremes of Read Length Distribution Metrics. Proportions of reads of each length in simulated read length distributions. The highest (top row) and lowest (bottom row) scores for each metric are shown. For example, the top left plot titled 'IQR (Top)' shows the read length distribution of the highest scoring distribution according to the $Read\ Length\ Distribution_{IQR\ Normalised}$ metric.

Although $Read\ Length\ Distribution_{CV}$, $Read\ Length\ Distribution_{IQR\ Normalised}$ and $Read\ Length\ Distribution_{MaxProp}$ are all metrics that assess dispersion it is of note that $Read\ Length\ Distribution_{IQR\ Normalised}$ and $Read\ Length\ Distribution_{MaxProp}$ agree on the distribution with the least dispersion whereas $Read\ Length\ Distribution_{CV}$ scores this metric lowest. A high coefficient of variance suggests a greater degree of dispersion which is seen in its top ranked distribution. It appears that each of these dispersion metrics perform sufficiently for top and bottom ranked distributions, however, from the scatter matrix shown below it is clear that there is not a strong correlation between these metrics for these simulated distributions.

In terms of the shape of these distributions, the top scoring distribution on the normality score is indeed normally distributed and the bottom distribution is heavily skewed implying this metric can successfully delineate normal and non-normal distributions. However, the scatter matrix shows that the majority of the distributions in this simulated set score lowly on this metric. This implies that near-normal distributions are not delineated from non-normal distributions. As a result, the $Read\ Length\ Distribution_{Normality}$ metric is a better classifier of normal and non-normal distributions rather than it being a good metric of the relative-normality of the distribution. The $Read\ Length\ Distribution_{Bimodality}$ metric assigns its highest score to the same distribution as the maximum proportion and IQR metrics, however, it is unclear whether this is down to there being a lack of truly bimodal distributions in the simulated set or if there is an issue with the metric.

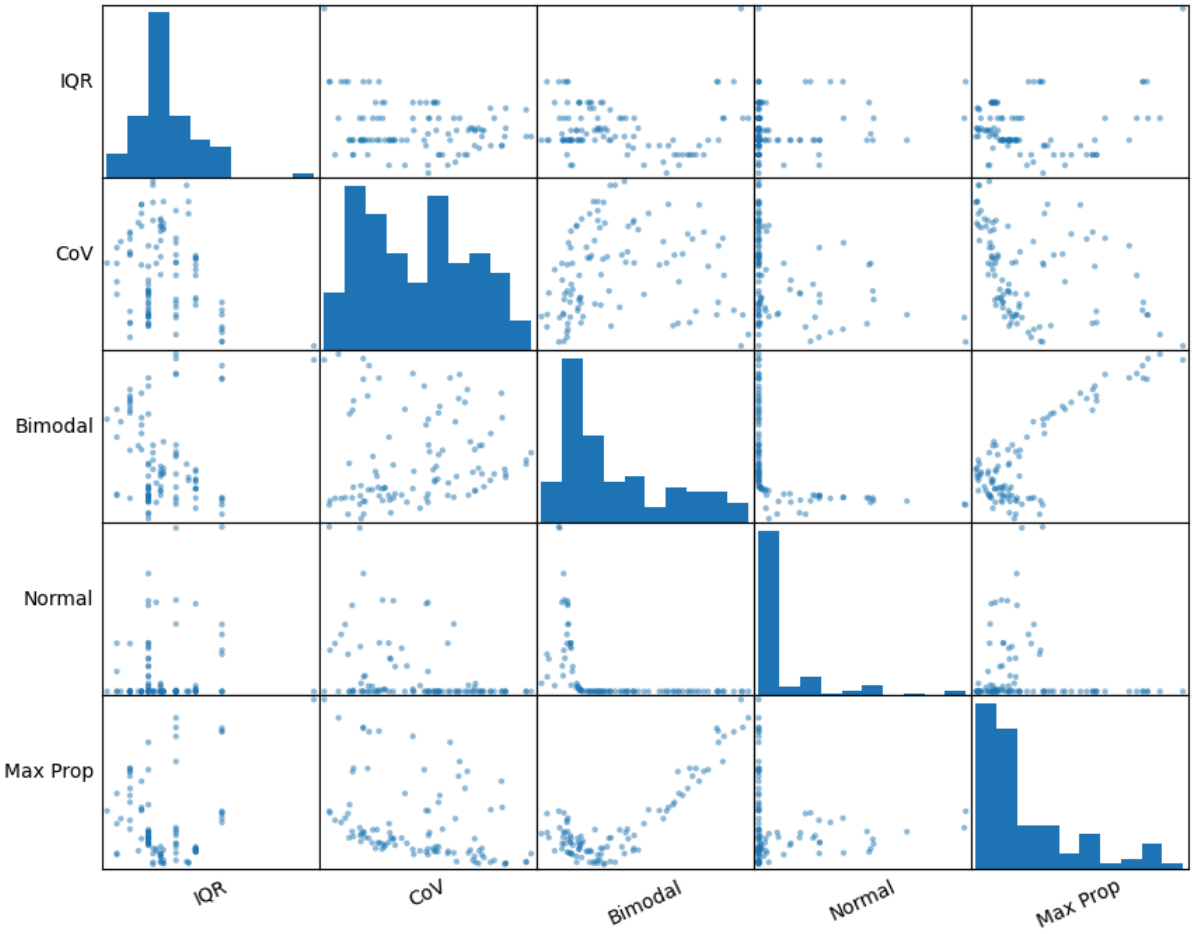


Figure 5.10: Scatter matrix of read length distribution metrics.

The correlations between all read length distribution metrics are not expected to be strong as they are assessing different properties of the data. There is however an incidental correlation between the bimodality and maximum proportion metrics.

5.5.3 Relationships Between Metrics

The correlations between metrics calculated for each of the 1,322 samples showed varying degrees of concordance with the results observed for the simulated data. This was broadly to be expected as the data was simulated in order to evaluate the performance of the metrics across the range of potential values rather than accurately capture what would be observed in real datasets.

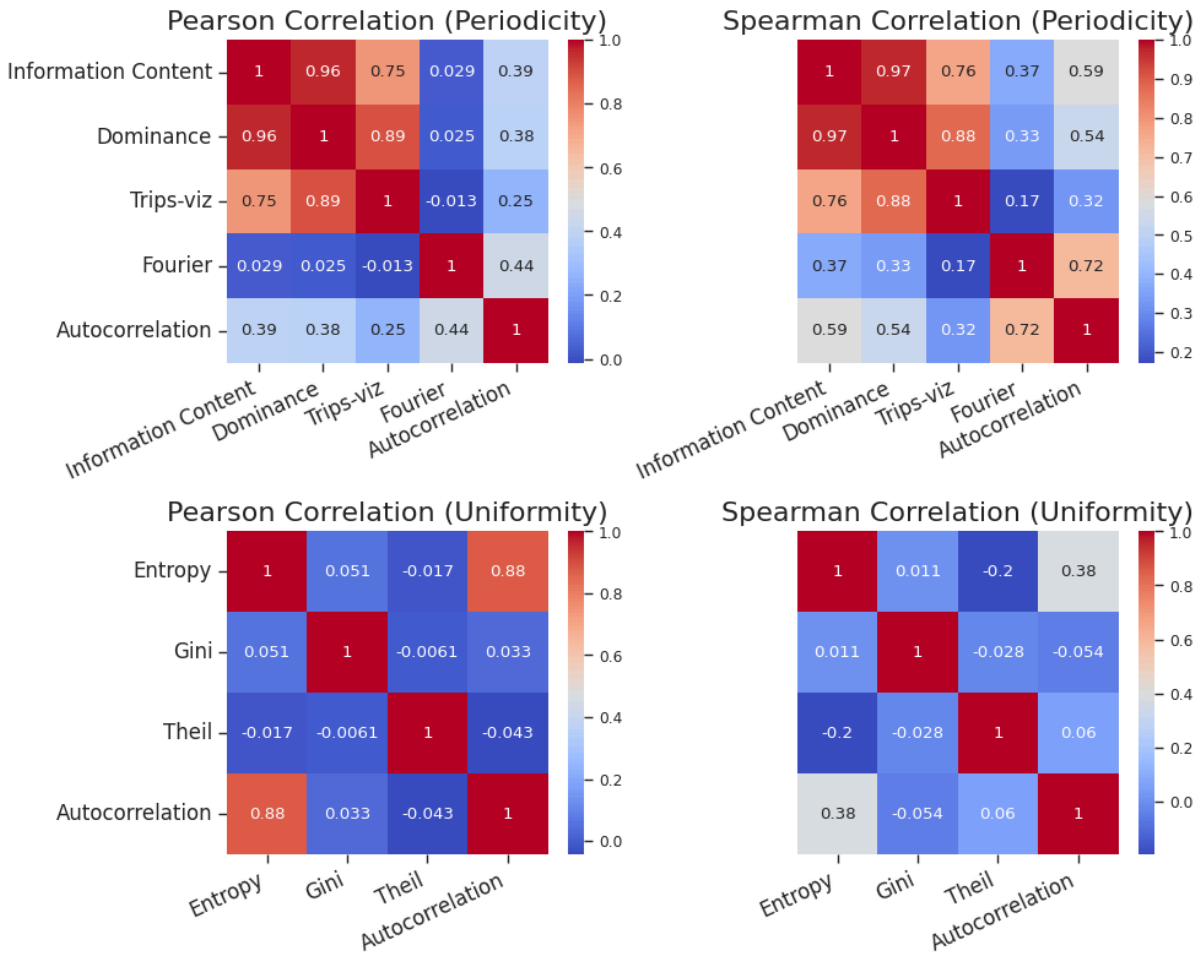


Figure 5.11: Correlation matrices of **Top:** Periodicity and **Bottom:** Uniformity (global) scores across 1,322 human Ribo-Seq samples.

As observed in the simulated data there is also a strong correlation between the proportion of signal metrics in the real human datasets. However, there are discrepancies between the Spearman and Pearson correlations of the $Periodicity_{Fourier}$ and $Periodicity_{auto}$ periodicity metrics and to some extent the proportion of signal metrics. This implies that the relationships between these metrics is not linear, resulting in poor Pearson correlation, but has a greater degree of monotonic concordance evidenced by the higher Spearman correlation. The inequality metrics correlate universally weakly with the other metrics in the real data, however, the Autocorrelation and Entropy metrics show a strong linear relationship that is not necessarily monotonic. Where these correlations between metrics are observed across the human datasets we can infer that these metrics are measuring similar aspects of the data properties.

The discrepancies between Spearman and Pearson correlations can be attributed to either a non-linear relationship between the metrics (resulting in low Pearson correlation) or a non-monotonic relationship resulting from the presence of outliers or changes in variability across the range of values. Variability and noisiness is a universal feature of the calculated metric distribution across all metric types. This makes it difficult to interpret each resulting metric on a real dataset without the context of the rest of the real data.

To remedy this and aid in the interpretation of RiboMetric results we deduced decile cutoffs to categorise each resulting metric-based on where it would rank among the 1,322 human Ribo-Seq datasets. These deciles can help make high-level observations about patterns across the metrics that can be lost in the variability of the raw data. To further aid these observations, metrics were averaged within their metric-type category excluding any metrics that were previously shown to be poorly performing on real data metrics. From this approach, we observe that low scoring samples in terms of periodicity are often high scoring in terms of terminal nucleotide bias.

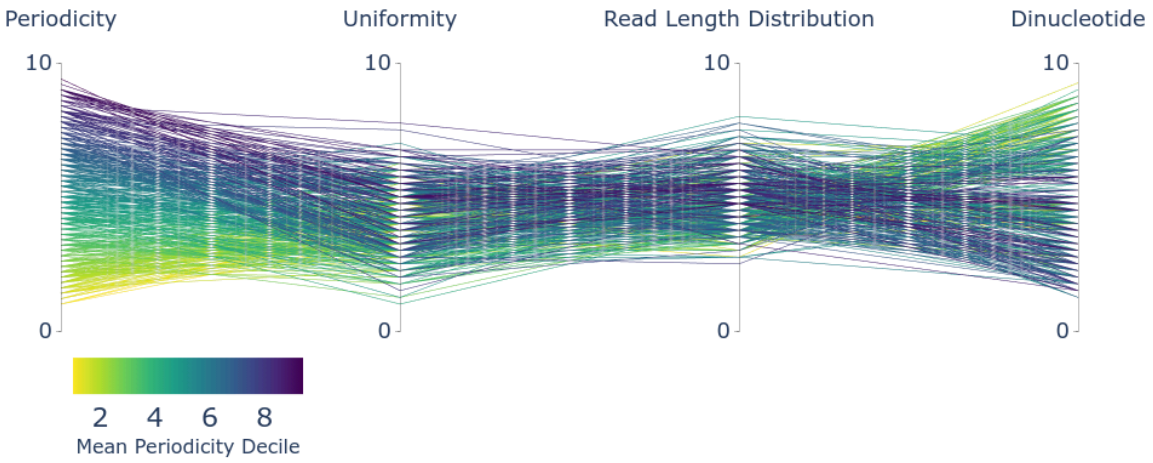


Figure 5.12: Parallel Coordinate plot of RiboMetric deciles averaged across metric-type categories. Lines are coloured according to their mean periodicity decile.

There are a number of caveats to this coarse approach to providing context to individual metrics and further work can be done to refine the metrics provided. For example, when assigning deciles, the distribution of metrics is often not near-uniform. Skewed distributions can result in large variability in the range of values that a decile will cover and variations within deciles may be overlooked.

5.6 Discussion

5.6.1 Overview of RiboMetric's Contributions

In this work, we introduce RiboMetric, the first metric-based property assessment tool for ribosome profiling data. The metrics implemented in this tool aim to characterise individual properties of Ribo-Seq datasets in numerical format in order to facilitate large-scale comparisons and further meta analysis. Although one of the primary applications of RiboMetric will be to carry out quality assessment on Ribo-Seq samples, RiboMetric does not assign good/bad quality classifications to datasets as dataset quality is not as simple as high/low scores on individual metrics. Some properties of Ribo-Seq datasets, such as triplet periodicity, vary considerably based on the library preparation protocol and depending on the intended application of a given dataset different priorities should be given to these properties. For example, a dataset with low triplet periodicity scores may be considered ‘bad quality’ for the prediction of translons (translated regions) but may be perfectly ‘good quality’ for the quantitative analysis of gene-level translation dynamics. As a result, RiboMetric is termed a property assessment tool and using the outputs of RiboMetric users can make their own assessments of whether datasets meet their ‘quality’ standards.

The metrics implemented in RiboMetric cover a variety of commonly discussed properties of Ribo-Seq samples that will aid the comparisons of individual datasets, such as those made available via the RiboSeq Data Portal (RDP), without the need for manual inspection of reports. RPFdb, like the RDP, is a web based interface for a database of ribosome profiling samples and studies. Both platforms make quality reports from FastQC available for manual inspection, however, cross-study comparisons of dataset properties are difficult and no Ribo-Seq specific reports are made available. Metric-based property assessment tools have a clear advantage in this regard as numerical properties of datasets can facilitate the sorting and filtering of these datasets based on metric properties.

Aside from the focus on metric-based assessment of dataset properties, RiboMetric also delineates itself from other Ribo-Seq quality assessment tools such as RiboSeQC, RibosomeProfilingQC and RiboStreamR in a number of ways. Firstly, it is written in Python

and is available as a command-line application without the requirement of additional scripting. This is significant as it facilitates the simple integration of RiboMetric into pipelines developed in workflow management systems such as Snakemake and Nextflow. R based applications required the development of custom R scripts to facilitate execution from the command-line adding an extra barrier to their inclusion in projects such as nf-core (45).

RiboMetric enables the quantitative and qualitative assessment of Ribo-Seq dataset properties via an easy to install command-line application which has the potential to greatly improve the standardisation and reproducibility of the comparative assessment of Ribo-Seq dataset properties. This has the potential to influence both the development of computational and wet-lab methods ultimately resulting in more reproducible and interpretable translatomic research which is explored in detail in the following sections as we discuss how to interpret RiboMetric results.

5.6.2 Interpretation of Metric Performance

When conducting the initial assessment of the performance novel metrics, such as we aimed to do in this work, simulated data is invaluable. With simulated data we can control the function inputs, easing the interpretation of the outputs. However, some dataset properties are easier to simulate than others and as a result, some metrics were difficult to assess and fine tune.

It is relatively easy to simulate ribosome profiles with varying degrees of 3nt periodicity which is fortunate as it is often touted as the most important property of Ribo-Seq data. These simulated profiles provided numerous insights into various global and local dynamics in periodicity metrics. The scatter matrix showing pairwise comparisons of each periodicity metric reinforced the preconceived notion that the proportion of reads in the coding frame and the true periodic nature of the signal are different but related properties. Signal based metrics, the Autocorrelation and Fourier transform based metrics, are the most closely related with a strong positive but non-linear relationship. These are the best metrics for assessing the periodic nature of the Ribo-Seq signal, however, there are some aspects that may make it appropriate to consider what to choose between these two metrics. The three nucleotide periodic signal originates from the codon-to-codon translocation of the ribosome and as a

result, is not perfectly sinusoidal as it occurs in discrete peaks of variable heights rather than a continuous sine wave. As this fits the assumptions of autocorrelation analysis better than fourier analysis it may be assumed that the autocorrelation metric would perform better on datasets with low local uniformity. However, this was not considered in the simulations and as a result should be investigated with real data.

The read frame dominance and Trips-Viz based metrics for the assessment of triplet periodicity are not signal based metrics and instead are based on the proportion of reads that has been assigned to each reading frame. These metrics also have a strong positive correlation but with a bit more variability. This is due to Trips-Viz based metric favouring profiles with clear single frame signal rather than strictly the proportion of reads in the coding frame. The dominance metric is a naive approach that is commonly used in filtering datasets to choose the subset with periodic read lengths from a Ribo-Seq dataset. The Trips-Viz metric was designed to tell a user of the transcriptome browser Trips-Viz which datasets will provide the clearest ribosome profiles when visualised. Each of these has its own value but are worse metrics for assessing true periodicity.

The information content based metric for the assessment of triplet periodicity is a novel introduction in this work. It is based on the degree of uncertainty as to which frame a read mapped to a given region originated from the translation of the coding frame. Although it is based on the proportion of reads per frame it correlated well with the signal based metrics. Due to this correlation, the Information Content metric may be a high-performing metric for the selection of periodic read lengths that is both nuanced and lightweight. This metric correlates weakest with the Trips-Viz metric which is of note as the motivation for the development of the information content based metric was to alleviate the shortcomings of the Trips-Viz metric.

The uniformity of Ribo-Seq signal is a complex data property to quantify as much of the unevenness observed in a ribosome profile may indeed be biological signal however some unevenness can be attributed to systematic biases in the library such as ligation biases or PCR duplicates. The differences between global and local uniformity were previously described. The variability in the origin of unevenness and the different types of uniformity makes it

difficult to simulate data to assess this data property and as a result, there are fewer concrete deductions to be made about uniformity assessment compared to triplet periodicity.

Local uniformity is the kind of uniformity more commonly associated with data quality and as a result, the uniformity metrics were primarily developed with this data property in mind and it was only post-application development that the concept of global uniformity was considered. As a result, we see that the implemented uniformity assessment metrics delineate between profiles with high and low local uniformity generally well. However, the lack of correlation between the Theil index based metric and the other three uniformity metrics raises questions about the suitability of the selected examples in this case, but the strong/moderate correlations observed between the Autocorrelation, Gini Index and Entropy based metrics imply these metrics are performing to some degree. Based on these simulations, without carrying out further investigations, the Gini index or Entropy based metrics would be the most appropriate metrics for assessing local uniformity.

From the select examples, we demonstrated that the autocorrelation based uniformity metric best delineates between the datasets with varying degrees of global uniformity. However, the Gini index and Entropy based metrics show a strong positive correlation indicating these two metrics are capturing some shared property of the data and it may be down the mode of assessment that results in the poor response of these metrics to changes in global uniformity. As this is just one aspect of one metric among many explored in this work there was not sufficient time to investigate this further, however this work does highlight a number of questions about both types of uniformity and how they can be characterised numerically.

Terminal nucleotide bias is a distinctly quality associated metric. In practice, there are no conditions where a canonical Ribo-Seq sample with high terminal nucleotide bias would be considered high quality. These biases can have implications for downstream analysis to the extent that a R package was developed specifically to handle this form of bias (47). Fortunately, both metrics developed to assess this property performed well on simulations when there is a bias towards one dinucleotide being found at the read terminus. The Kullbach Leibler divergence based metric however performs significantly better when there is more than one dinucleotide found at the read terminus than expected. As a result, this metric is the recommended way to detect this dataset property.

Finally, the read length distribution metrics aim to assess an important array of data properties as discussed in the Results section. In practice the best application of these specific metrics will be in the case of significant outliers in these metrics enabling the exclusion of datasets with confounding read populations.

5.6.3 Potential for Facilitating large-scale Analysis

Applied to large repositories of ribosome profiling datasets such as the RiboSeq Data Portal and RPFdb, RiboMetric has the potential to enhance the reusability of the data hosted on these platforms. Currently, only standalone reports are available on both of these platforms. Through this, users can investigate various properties of each dataset, however, the reports provided are generic read quality reports from FastQC and FastP and not specific to Ribo-Seq data. The application of RiboMetric to the data on these platforms would not only add Ribo-Seq specific quality reports but the metrics produced could also be used to filter datasets without manual interrogation of the individual reports.

A common application of ribosome profiling data that is gaining interest is the identification of translated regions (translons). The methods used for this purpose often rely on the identification of periodic Ribo-Seq signal spanning a candidate open reading frame. When sourcing data for translon calling, it would be advisable for researchers to identify appropriate datasets based on the RiboMetric periodicity metric that is most similar to the method by which they approximate whether an ORF is translated. As previously mentioned, periodicity metrics can disagree about whether the same profile is periodic or not and as a result, the metric used to subset datasets should match the one used to call translons.

In conjunction with the metadata associated with each sample, such as organism, inhibitor and biological condition, or study, including library preparation protocol, insights can be gleaned about the factors contributing to various dataset properties.

5.7 Conclusion

In this work we introduce RiboMetric, a command-line application that enables the assessment of ribosome profiling dataset properties through interactive reports and a comprehensive set of metrics that cover an array of dataset properties. These metrics enable researchers to evaluate and compare datasets based on key properties of Ribo-Seq data, including periodicity, uniformity, read length distribution and terminal nucleotide bias. Through the exploration of existing human data and a variety of simulated datasets, we explore a number of metrics across each of these categories informing the assumed preset parameters in the published package. RiboMetric is made readily available for installation from PyPI (https://pypi.org/project/RiboMetric/).

5.8 References

1. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009 Apr 10;324(5924):218–23.

2. Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. Nat Rev Mol Cell Biol. 2015 Nov;16(11):651–64.

3. Mudge JM, Ruiz-Orera J, Prensner JR, Brunet MA, Calvet F, Jungreis I, et al. Standardized annotation of translated open reading frames. Nat Biotechnol. 2022 Jul;40(7):994–9.

4. Calviello L, Ohler U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. Trends Genet. 2017 Oct 1;33(10):728–44.

5. Chothani S, Ho L, Schafer S, Rackham O. Discovering microproteins: making the most of ribosome profiling data. RNA Biol. 2023 Dec 31;20(1):943–54.

6. Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. Nat Rev Genet. 2014 Mar;15(3):205–13.

7. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012 Nov;22(11):2219–29.

8. Tierney JAS, Świrski M, Tjeldnes H, Mudge JM, Kufel J, Whiffin N, et al. Ribosome decision graphs for the representation of eukaryotic RNA translation complexity. Genome Res. 2024 Apr 1;34(4):530–8.

9. Ingolia NT, Hussmann JA, Weissman JS. Ribosome Profiling: Global Views of Translation. Cold Spring Harb Perspect Biol. 2019 May 1;11(5):a032698.

10. Gobet C, Naef F. Ribosome profiling and dynamic regulation of translation in mammals. Curr Opin Genet Dev. 2017 Apr 1;43:120–7.

11. Schneider-Poetsch T, Ju J, Eyler DE, Dang Y, Bhat S, Merrick WC, et al. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. Nat Chem Biol. 2010 Mar;6(3):209–17.

12. Obrig TG, Culp WJ, McKeehan WL, Hardesty B. The mechanism by which cycloheximide and related glutarimide antibiotics inhibit peptide synthesis on reticulocyte ribosomes. J Biol Chem. 1971 Jan 10;246(1):174–81.

13. Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2012 Sep 1;109(37):E2424-32.

14. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014 May 2;33(9):981–93.

15. Ingolia NT, Lareau LF, Weissman JS. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity of Mammalian Proteomes. Cell. 2011 Nov 11;147(4):789–802.

16. Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res. 2012 Nov 1;22(11):2208–18.

17. Gao X, Wan J, Liu B, Ma M, Shen B, Qian SB. Quantitative profiling of initiating ribosomes in vivo. Nat Methods. 2015 Feb;12(2):147–53.

18. Van Damme P, Gawron D, Van Criekinge W, Menschaert G. N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. Mol Cell Proteomics MCP. 2014 May;13(5):1245–61.

19. Zhang P, He D, Xu Y, Hou J, Pan BF, Wang Y, et al. Genome-wide identification and differential analysis of translational initiation. Nat Commun. 2017 Nov 23;8(1):1749.

20. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016

Feb;13(2):165–70.

21. Andreev DE, O'Connor PBF, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, et al. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. eLife. 2015 Jan 26;4:e03971.

22. Gerashchenko MV, Gladyshev VN. Translation inhibitors cause abnormalities in ribosome profiling experiments. Nucleic Acids Res. 2014;42(17):e134.

23. Sharma P, Wu J, Nilges BS, Leidel SA. Humans and other commonly used model organisms are resistant to cycloheximide-mediated biases in ribosome profiling experiments. Nat Commun. 2021 Aug 24;12(1):5094.

24. VanInsberghe M, van den Berg J, Andersson-Rolf A, Clevers H, van Oudenaarden A. Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. Nature. 2021 Sep;597(7877):561–5.

25. Ferguson L, Upton HE, Pimentel SC, Mok A, Lareau LF, Collins K, et al. Streamlined and sensitive mono- and di-ribosome profiling in yeast and human cells. Nat Methods. 2023 Nov;20(11):1704–15.

26. Archer SK, Shirokikh NE, Beilharz TH, Preiss T. Dynamics of ribosome scanning and recycling revealed by translation complex profiling. Nature. 2016 Jul;535(7613):570–4.

27. Giess A, Torres Cleuren YN, Tjeldnes H, Krause M, Bizuayehu TT, Hiensch S, et al. Profiling of Small Ribosomal Subunits Reveals Modes and Regulation of Translation Initiation. Cell Rep. 2020 Apr 21;31(3):107534.

28. Bohlen J, Fenzl K, Kramer G, Bukau B, Teleman AA. Selective 40S Footprinting Reveals Cap-Tethered Ribosome Scanning in Human Cells. Mol Cell. 2020 Aug 20;79(4):561-574.e5.

29. Wagner S, Herrmannová A, Hronová V, Gunišová S, Sen ND, Hannan RD, et al. Selective Translation Complex Profiling Reveals Staged Initiation and Co-translational Assembly of Initiation Factor Complexes. Mol Cell. 2020 Aug 20;79(4):546-560.e7.

30. Diament A, Feldman A, Schochet E, Kupiec M, Arava Y, Tuller T. The extent of ribosome queuing in budding yeast. PLoS Comput Biol. 2018 Jan;14(1):e1005951.

31. Verbruggen S, Menschaert G. mQC: A post-mapping data exploration tool for ribosome profiling. Comput Methods Programs Biomed. 2019 Nov;181:104806.

32. Perkins P, Mazzoni-Putman S, Stepanova A, Alonso J, Heber S. RiboStreamR: a web application for quality control, analysis, and visualization of Ribo-seq data. BMC Genomics. 2019 Jun 6;20(5):422.

33. Calviello L, Sydow D, Harnett D, Ohler U. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data [Internet]. bioRxiv; 2019 [cited 2024 Mar 10]. p. 601468. Available from: https://www.biorxiv.org/content/10.1101/601468v1

34. Ou J, Hoye M. ribosomeProfilingQC: Ribosome Profiling Quality Control. R Package Version. 2022;1(0).

35. Wang H, Yang L, Wang Y, Chen L, Li H, Xie Z. RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. Nucleic Acids Res. 2019 Jan 8;47(D1):D230–4.

36. Kiniry SJ, Judge CE, Michel AM, Baranov PV. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. Nucleic Acids Res. 2021 Jul 2;49(W1):W662–70.

37. O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB. GeneGini: Assessment via the Gini Coefficient of Reference "Housekeeping" Genes and Diverse Human Transporter Expression Profiles. Cell Syst. 2018 Feb 28;6(2):230-244.e1.

38. Wright Muelas M, Mughal F, O'Hagan S, Day PJ, Kell DB. The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. Sci Rep. 2019 Nov 29;9(1):17960.

39. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021 Feb 1;10(2):giab008.

40. Liu Q, Shvarts T, Sliz P, Gregory RI. RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. Nucleic Acids Res. 2020 Jul 2;48(W1):W218–29.

41. Birkeland Å, ChyŻyńska K, Valen E. Shoelaces: an interactive tool for ribosome profiling processing and visualization. BMC Genomics. 2018 Jul 18;19(1):543.

42. Ozadam H, Geng M, Cenik C. RiboFlow, RiboR and RiboPy: an ecosystem for analyzing ribosome profiling data at read length resolution. Bioinformatics. 2020 May 1;36(9):2929–31.

43. Chacko J, Ozadam H, Cenik C. RiboGraph: An interactive visualization system for ribosome profiling data at read length resolution. bioRxiv. 2024 Jan 12;2024.01.11.575228.

44. Tjeldnes H, Labun K, Torres Cleuren Y, Chyżyńska K, Świrski M, Valen E. ORFik: a comprehensive R toolkit for the analysis of translation. BMC Bioinformatics. 2021 Jun 19;22(1):336.

45. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020 Mar;38(3):276–8.

46. Kiniry SJ, O'Connor PBF, Michel AM, Baranov PV. Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. Nucleic Acids Res. 2019 Jan 8;47(D1):D847–52.

47. Mok A, Tunney R, Benegas G, Wallace EWJ, Lareau LF. choros: correction of sequence-based biases for accurate quantification of ribosome profiling data [Internet]. bioRxiv; 2023 [cited 2024 Mar 27]. p. 2023.02.21.529452. Available from: https://www.biorxiv.org/content/10.1101/2023.02.21.529452v1

6. Discussion

6.1 Advances in Translatomics

This thesis addresses two predominant challenges that impede the study of translatomic complexity in this modern era of high-throughput biological investigation. Namely, these are; the insufficiency of the current models for the representation of translational complexity and the inaccessibility of standardised translatomic data. By introducing novel concepts and bioinformatic applications we make significant strides towards advancing our ability to study the intricacies of RNA translation.

Across the first two chapters, we introduce a novel conceptual model, Ribosome Decision Graphs (RDGs), and describe the implementation of this concept in a Python package. These developments enable the accurate representation of translatomic complexity and through this data structure enable their application to translatomic analysis. The following two chapters look to provide accessible and standardised translatomic data with comprehensive metrics describing the dataset's properties. The RiboSeq Data Portal serves as an access point to thousands of preprocessed Ribo-Seq samples along with standardised sample metadata. While RiboMetric is a command-line application that enables the metric-based comparison of a comprehensive set of Ribo-Seq dataset properties.

The combination of RDGs and accessible, high-quality translatomic data has the potential to extensively capture and subsequently support the investigation of the complexity of gene expression and regulation at the level of translation. Collectively, these advancements lay the groundwork for more sophisticated and integrative approaches to translatomics research.

With the introduction of a framework for the accurate representation of translatomic complexity, in concert with accessible data, new dimensions of investigation are opened up across the field of translation.

RDGs have the potential to significantly change how the annotation of translation events is conducted. It is not uncommon to see reference to the 'correct start codon' in manuscripts that discuss the annotation of translons. This is a flawed concept and although errors in start codon

fidelity can come about through dysfunction or malfunction that is rarely what is referred to in this context. In truth, translation initiation sites (TISs) are stochastic rather than deterministic events and probabilities of initiation versus leaky scanning vary for a given TIS across tissues and conditions. Similarly, recording events such as stop codon readthrough and frameshifting are probabilistic events where the probabilities are altered under various conditions such as changes in polyamine levels. With RDGs we can construct an accurate representation of the occurrence and co-occurrence of known translation events (eg. initiation, readthrough, frameshifting etc) based solely on the transcript coordinates of each event provided we also have the transcript sequence. This shift from the annotation of deterministic ranges of coordinates, such as from open reading frame (ORF) start to stop, to the annotation of the locations of stochastic 'decisions' ensures accurate representation is possible.

Pantranscriptomics was described in the introduction as the combination of the representation of genomic variation encoded in the human pangenome with the representation of transcriptomic complexity introduced through splicing (1). This approach enabled researchers to explore haplotype-specific expression using these representations. These methods could first be applied to the data in the RiboSeq Data Portal (RDP) to study haplotype specific translation dynamics. However, this combination of representation of complexity can now be extended to known translatomic complexity by the integration of pangenomes, splice graphs and RDGs. This additional layer of translatomic complexity, added to both haplotype and alternative isoform analysis, using RDG representations has the potential to unveil the effects of splicing and genetic variation on gene expression dynamics. This would have the potential to revolutionise the analysis of gene expression and to be entirely multimodal while also accommodating genomic variants.

The detection of translation events has indirectly been the focus of many studies and bioinformatic methods to date. Oftentimes the methods developed to identify translated ORFs identify candidate open reading frames sequences for which there is clear 3nt periodicity and a uniform ribosome profile across the candidate sequence. As each candidate sequence is treated independently, signal originating from multiple overlapping translons is easily misattributed (2,3). This is one clear example of where RDGs have the potential to have significant impacts on the practical identification of translation events. As an RDG can be

constructed to contain relational information regarding all candidate open reading frames there is the potential to assign reads to candidate translons in a way not currently possible. This will significantly improve performance on overlapping translated regions including those originating from frameshifting assuming accurate candidate frameshift sites are obtainable.

Now that large amounts of standardised Ribo-Seq data are available in various cell lines and tissues via the RiboSeq Data Portal, we can combine RDGs with expression data to characterise how different translons are translated across tissues and conditions. Previous work has looked to assign probabilities of translation initiation to TISs from Ribo-Seq data (4). The principles of this work can be extended to make these approximations for all transcripts in all tissues for which we have sufficient data. These approaches however make the assumption that ribosomes that arrive at a given TIS got there via leaky scanning. When in reality, depending on the locus architecture, a significant number of ribosomes may arrive by reinitiation following the translation of an upstream translation. In related work not presented in this thesis we show that Translation Complex Profiling Sequencing (TCP-Seq) data can be used to qualitatively assess the leaky scanning/reinitiation dichotomy (See Supplementary Paper 1). There is the potential to combine RDGs with this data to ascertain the proportion that originates at given TISs via reinitiation/leaky scanning in a quantitative manner. This would represent a significant advance in the insights obtainable from TCP-Seq data.

Combined with the data provided by the RiboSeq data portal, RDGs can also be used to facilitate cross-species comparisons of the regulation of translation. Although the sequence of a regulatory upstream overlapping translon itself may not be conserved, its existence in orthologs or other members of the gene family may be. This conserved presence of regulatory translons is difficult to detect, however, it is expected that similar RDG architecture would originate from genes with similar regulatory mechanisms. As a result, the identification of shared or similar graph structure can serve as a proxy for the conservation of regulatory mechanisms. These investigations do not entirely rely on but would greatly benefit from the identification of translons from the multi-species Ribo-Seq data available on the RDP.

Ribosome decision graphs can also serve as a powerful hypothesis generation and result validation approach for wet-lab researchers studying complex translational events using western blots and luciferase assays. RDGs can be generated for each construct along with the

proteoforms that would be expressed if each translon was translated. Based on the size of these proteoforms and their molecular weights researchers can anticipate what they might observe on a western blot or explain an unexpected band in a blot they have already run. For luciferase assays, RDGs can guide construct design to eliminate potentially confounding translation events. RDG visualisation functionality can also be used to explain construct design in manuscript figures. RiboMetric can similarly help bridge the gap between wet and dry lab researchers as it can provide nuanced insights into Ribo-Seq data properties which could be of benefit in protocol optimisation.

Many of the concepts introduced in this thesis are quite novel and hopefully set solid foundations for many further research directions. However, there are already a number of foreseeable limitations that may impact which areas these tools are most applicable to.

Ribosome Decision Graphs enable the representation of many different translation events for any given locus. While the RDG python package does not yet support the addition of translational bypassing or StopGo recoding events it is possible to conceptualise how they may be included in an RDG. The key aspect to consider is what happens to the ribosome during these events.

The representation of translational bypassing in a ribosome decision graph can be thought of in two different ways. It could involve the creation of a branching point within a translon akin to that of a frameshifting event. However, in this case, the frame would be shifted forward by a much larger number of nucleotides than in canonical ribosomal frameshifting events. For example, in the case of the T4 bacteriophage's *gene 60,* this 'frameshifting' event would be a +50 shift (5). Alternatively, and potentially more accurately, this bypassing could be seen as two translational events: ribosomal take-off and ribosomal landing. Following the arrival of the elongating ribosome at the take-off site, a proportion, although reportedly near 0% for *gene 60*, of ribosomes will not initiate bypassing and as a result will terminate there. Subsequently, of the ribosomes that have taken off only a proportion of the ribosomes will land. The translation would be represented in two translons. One that ends at the take-off site that is translated by both those ribosomes that do not take off and those that do not land. The second translon then represents the ribosomes that successfully both take off and land. The case of StopGo is more straightforward as it is more akin to a high-efficiency stop codon

readthrough except that two peptide products are produced from the one translon. This does not necessarily require any further accommodations in RDGs as RDGs are not directly concerned with the peptide products. However, future work will look to best accommodate these events in a way that is both machine and human-interpretable.

There are further limitations in the general use of RDGs due to the uncertainty remaining in our understanding of the mechanisms of translation, stringency of start codon selection and effect of variants on translational event probabilities. These factors make it difficult to reliably characterise a locus's translation in an RDG. While it is possible to characterise large numbers of potential events, this increased graph complexity can dramatically reduce the usability of the graph as it limits interpretability. The increasing availability of high-quality translatomic data will aid in this characterisation across various organisms, tissues and conditions. However, further work is also required on how to best fit a RDG to the observed translatomic signal.

As the need for diverse and high-quality Ribo-Seq datasets increases to support accurate characterisation of translation, the importance of identifying these datasets on public repositories also increases. The RiboSeq Data Portal will continue to be updated so that the search terms and controlled vocabularies are continuously improved to maximise the usefulness of the resource. The current approach for metadata standardisation is basic from a natural language processing perspective. With the advent of low cost large language models it is becoming increasingly evident that more nuanced metadata standardisation is possible. These more complex models will not be limited to fetching metadata from specific columns of the metadata table but can be extended to freely accessible manuscripts on PMC or BioRxiv.

Addressing these limitations will be the focus of a number of, but not all of, the future directions stemming from the novel advancements introduced in this thesis. Two prominent avenues that we aim to address in the next phase of these projects are; the production of high confidence RDGs for various organisms, tissues and conditions for which we have sufficient supporting data and the investigation of the effects that common, supposedly non-coding, variants may have on gene expression using RDGs.

High-confidence RDGs are graphs representing the set of translons that we have sufficient evidence to say they are reproducibly detectable as translated from high-quality Ribo-Seq data across multiple experiments. High quality Ribo-Seq data can be obtained from the RiboSeq Data Portal using the scores from RiboMetric, which are already available internally for humans and can be extended to all organisms. Methods for the identification of high confidence translons from Ribo-Seq data using RDGs will need to be developed. There are varying degrees to which RDGs might be used. They may be used in a method as simple as implementing a scanning model informed open reading frame candidate selection step where only candidates that are likely to be reached by a scanning ribosome, either by leaky scanning or reinitiation, are considered. Alternatively, an entirely novel approach for translated region characterisation, many different potential RDGs are generated and subsequently the graph that best fits the observed data will be selected via maximum likelihood or Bayesian inference methods. The main barriers to achieving these goals are the determination of what is considered sufficient evidence for the translation of a translon and how to calculate the best fit of a graph to observed data. These high confidence RDGs can then subsequently be used as annotations of observed translation and will be particularly important for the interpretation of variants occurring in the 5` leaders.

Recent research has begun to investigate the effect of sequence variants located in the 5` leader of protein-coding genes (6,7). However, RDGs provide a unique framework for proposing the mechanisms by which these variants might affect CDS translation, either by the introduction of new upstream translons or by changes in the probability of different translation events due to alterations in sequence context. Current 5` UTR variant effect prediction methods focus on the creation or loss of upstream AUG codons, whereas, RDG-based methods can directly interrogate the relationship these variants may have with CDS translation. The integration of variant information with RDGs introduces a number of complications such as how to approximate the effect of probability altering variants and how to handle the combinatorial complexity introduced by the co-occurrence of multiple variants on the one transcript. Once these challenges are addressed RDG-based variant effect prediction methods have the potential to have substantial impacts on clinical genomics and personalised medicine. The more nuanced models of translatomic complexity provided by

RDGs could help hypothesise disease mechanisms and guide the development of targeted therapies.

In conclusion, this thesis introduces several concepts, methods and resources that collectively advance the study of translatomic complexity. Through the ribosome decision graph concept and associated implementation in a Python package, we introduce a novel framework and toolkit for working with this translational variation. While the RiboSeq Data Portal and RiboMetric resources introduce advances in the accessibility of high quality translatomic data that is not just important for the annotation of translation events, that can be subsequently represented in RDGs, but also for the characterisation of translation dynamics under various conditions. The RiboSeq Data Portal represents a significant advance in the standardisation of Ribo-Seq sample metadata and accessibility of processed datasets for reanalysis. While RiboMetric introduces the novel concept of metric based evaluation of Ribo-Seq data properties enabling large-scale dataset comparisons and selection for use in annotation and analysis projects. The developments detailed here form a foundation for many further improvements to these resources as well as many exciting avenues for future research. The development of a resource containing high-confidence ribosome decision graphs based on the available translatomic data followed by their subsequent application to variant effect prediction has the potential to advance these concepts as well as both basic research and clinical genomics. We anticipate that a number of the developments set out in this thesis will play a crucial role in representing and understanding translatomic complexity, ultimately leading to a more comprehensive understanding of the mechanisms underlying gene expression.

6.2 References

1. Sibbesen JA, Eizenga JM, Novak AM, Sirén J, Chang X, Garrison E, et al. Haplotype-aware pantranscriptome analyses using spliced pangenome graphs. Nat Methods. 2023 Feb;20(2):239–47.

2. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012 Nov;22(11):2219–29.

3. Erhard F, Halenius A, Zimmermann C, L'Hernault A, Kowalewski DJ, Weekes MP, et al. Improved Ribo-seq enables identification of cryptic translation events. Nat Methods. 2018 May;15(5):363–6.

4. Michel AM, Andreev DE, Baranov PV. Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. BMC Bioinformatics. 2014 Nov 21;15(1):380.

5. Samatova E, Konevega AL, Wills NM, Atkins JF, Rodnina MV. High-efficiency translational bypassing of non-coding nucleotides specified by mRNA structure and nascent peptide. Nat Commun. 2014 Jul 21;5(1):4459.

6. Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nat Commun. 2020 May 27;11:2523.

7. Rovelet-Lecrux A, Bonnevalle A, Quenez O, Delcroix W, Cassinari K, Richard AC, et al. Upstream open reading frame-introducing variants in patients with primary familial brain calcification. Eur J Hum Genet. 2024 Jul;32(7):779–85.

7. Supplementary Manuscript

This manuscript presented below originated over the course of the completion of this thesis and pertains in parts to the work previously described. At the time of submission it has been submitted to RNA Biology. This is primarily the ideas of Dr. Dmitri E. Andreev to which I contributed RDG figures, manuscript editing and the observations regarding sel-TCP-Seq in Figure 7.4

7.1 Translation complex profile sequencing allows discrimination of leaky scanning and reinitiation in upstream open reading frame-controlled translation

Authors: Dmitri E. Andreev [1,2 *], Jack A. S. Tierney [3,4], Pavel V. Baranov [3]

Affiliations:

1 Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS, 117997 Moscow, Russia

2 Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, 119992 Moscow, Russia

3 School of Biochemistry and Cell Biology, University College Cork, Cork T12 K8AF, Ireland

4 SFI Centre for Research Training in Genomics Data Science, University College Cork, Cork T12 K8AF, Ireland

* denotes corresponding authors

7.1.1 Abstract

Upstream open reading frames (uORFs) are a class of translated regions (translons) in mRNA 5’ leaders. uORFs are believed to be pervasive regulators of the translation of mammalian mRNAs. Some uORFs are highly repressive but others have little or no impact on downstream mRNA translation either due to inefficient recognition of their start codon(s) or/and due to efficient reinitiation after uORF translation. While experiments with uORF reporter constructs proved to be instrumental in the investigation of uORF-mediated mechanisms of translation control, they can have serious limitations as manipulations with uORF sequences can yield various artefacts. Here we propose a general approach for using translation complex profiling (TCP-seq) data for exploring uORF regulatory characteristics. Using several examples, we show how TCP-seq could be used to estimate both repressiveness and modes of action of individual uORFs. We demonstrate how this approach could be used to assess the mechanisms of uORF-mediated translation control in the mRNA of several human genes, including *EIF5*, *IFRD1*, *MDM2*, *MIEF1*, *PPP1R15B*, *TAF7,* and *UCP2*.

7.1.2 Introduction

Translation initiation in eukaryotes begins with 43S ribosome preinitiation complex (PIC) loading onto the 5’ end of mRNA, which is followed by scanning - a stepwise 5’-3’ PIC

movement, until a start codon, usually AUG, is recognized through successful establishment of codon-anticodon interaction in the ribosome P-site [1-5]. The scanning mode of translation initiation implies that the 5' proximal AUG codon of mRNA will most likely be the initiation site. Counterintuitively, for ~50% of mammalian mRNAs, the start codon of the protein-coding translon (CDS) is preceded by at least one upstream AUG-initiated uORF [6-9]. From the scanning mechanism perspective, any uORF can be an obstacle for scanning ribosomes, therefore uORFs are considered to be the most prevalent elements that repress mRNA translation [10-18]. Ribosome profiling, a technique which can detect mRNA footprints of translation 80S ribosomes at genome-wide level [19], allows the identification of thousands of mammalian translated uORFs, including those that use non-AUG initiation codons [20].

In theory, the inhibitory potential of a uORF depends on the efficiency of initiation at its start codon and the ability of post-terminating ribosomes to resume translation downstream. An inefficient uORF start codon allows leaky scanning and subsequent translation of CDS [2, 11]. In addition, the translation of a uORF can be followed by reinitiation of translation when the post-terminating 40S subunit gains initiation factors and resumes the scanning process. It is believed that the ribosome dwell time on uORF negatively correlates with reinitiation efficiency due to the loss of initiation factors bound to the 80S ribosome [11, 21, 22]. Substantial leaky scanning and reinitiation at uORFs can explain how mRNAs with dozens of uORFs are still able to affect CDS translation.

Our understanding of how uORFs regulate translation mostly originates from experiments utilising reporter constructs. The mutation of the start codon of a uORF to a non-initiating codon allows us to investigate how the uORF affects reporter gene translation, however, it does not allow us to resolve leaky scanning and reinitiation dichotomy. Several manipulations of the uORFs sequence can help to assess the relative contribution of these two processes to downstream translation (Figure 7.1A). Improvement of uORF start codon context or introduction of an additional in-frame start codon within the same uORF is expected to decrease leaky scanning. Persistent translation in the presence of such interventions suggests reinitiation. On the other hand, reinitiation can be ruled out by extending uORF beyond the

reporter stop codon. Reinitiation upstream of stop codons usually does not occur, thus the residual reporter activity can be attributed to the leaky scanning.

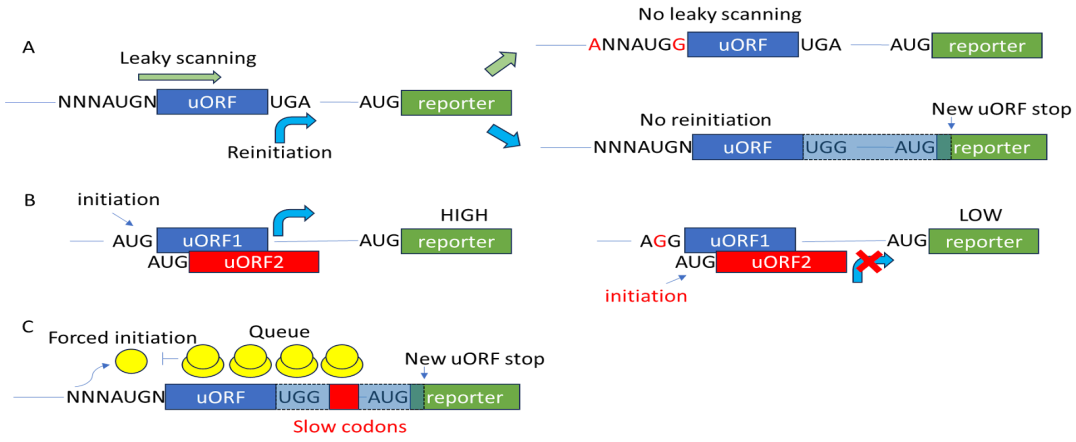


Figure 7.1. A) Workflow of uORF reporter analysis to investigate the impact of leaky scanning and reinitiation. B) Possible confounding results of mutational analysis in the case of a 5' leader with 2 uORFs. In the wild-type construct, uORF1 is preferentially translated and repressive so that uORF2 is skipped. Removal of uORF1's AUG codon will lead to uORF2 translation and subsequent reporter repression, C) Possible effect of extended uORF on uORF translation. uORF stop codon mutation creates a longer uORF sequence which contains non-optimal codons. Slow elongation within this artificial uORF sequence can lead to ribosome stalling and queuing, which in turn can enhance initiation at the uORF start codon.

However, experiments with reporter constructs have serious limitations – changing the uORF sequence can affect uORF-mediated mechanisms of translation control. Addressing the effect of an individual uORF on translation can be confounded by potential out-of-frame start codons within uORF (Figure 7.1B), and it may be difficult to identify them since some may be non-AUG. In this case, the removal of the uORF's start codon can activate the downstream uORF which is normally not translated. Furthermore, uORF extension by stop codon removal introduces a sequence that is not optimised for translation. This may artificially increase initiation at the uORF start codon due to ribosome queue formation. Such a phenomenon was previously reported for an AUU-initiated uORF in *AZIN1* mRNA, where a ribosome queue enhances initiation at an AUU codon and represses CDS translation [23]. It is worth noting that in most cases reporters lack native mRNA sequences that may contain cis-acting elements (e.g. those involved in long-range RNA-RNA interactions) that may be important for uORF-mediated translation control. Finally, the RNA levels of the reporters differ from native

mRNA levels and this could be important when RNA-binding proteins with limited availability are involved in the regulation.

Reporter constructs have been instrumental in the study of the relative contributions of leaky scanning and reinitiation following uORF translation. However, as previously described, they are not without their limitations that may confound reporter interpretation. Novel approaches for the assessment of the interplay between these mechanisms are required. Ideally, such approaches should enable genome-wide analysis and investigate uORF-mediated regulation in its native context.

7.1.3 Materials and Methods

TCP-Seq and Ribo-Seq data utilised in this work were from the Bohlen et al. 2023 dataset [24] and were accessed via the Trips-Viz transcriptome browser [25]. The Sequence Read Archive (SRA) run accessions of the samples used are available in supplementary information. The data made available on Trips-Viz are obtained from the SRA and processed in a standardised pipeline. The reads are trimmed of their adapters, and rRNA reads are removed prior to read alignment directly to the Gencode v25 human transcriptome. Ribosome profiling reads are offset on a per-read-length basis to predict the location of the A-site of each ribosome-protected fragment. This enables the assignment of each read to one of the three reading frames.

7.1.4 Results

The variation of ribosome profiling technique called translation complex profiling (TCP-seq) permits the measurement of scanning ribosomes on mRNA [26]. This approach is based on the fixation of ribosomal complexes on mRNA by chemical crosslinking followed by RNAse digestion and separation of PIC complexes from translating 80S ribosomes. Resulting mRNA footprints can be used to track the positions of PICs along mRNA 5' leaders. The information on PIC footprints can be used for both, the assessment of leaky scanning and on scanning of post-terminating ribosomal complexes downstream of uORFs, thus allowing for the assessment of uORF regulatory modality in its native context. Below we show how this could be done using the data obtained by Bohlen et.al. [24] and accessed via the Trips-Viz transcriptome browser [25].The Sequence Read Archive (SRA) run accessions of the samples used are available in supplementary information.

The principle of the approach is shown in Figure 7.2A. The signal upstream of uORF is a proxy of PIC loading efficiency onto mRNA 5' end and scanning towards the uORF start codon (red). PIC density within uORF corresponds to PICs that leaked through uORF start. If no reinitiation occurs downstream of the uORF, the density of PICs would be expected to stay the same, thus an increase in density is likely to occur due to reinitiation (Figure 7.2A).

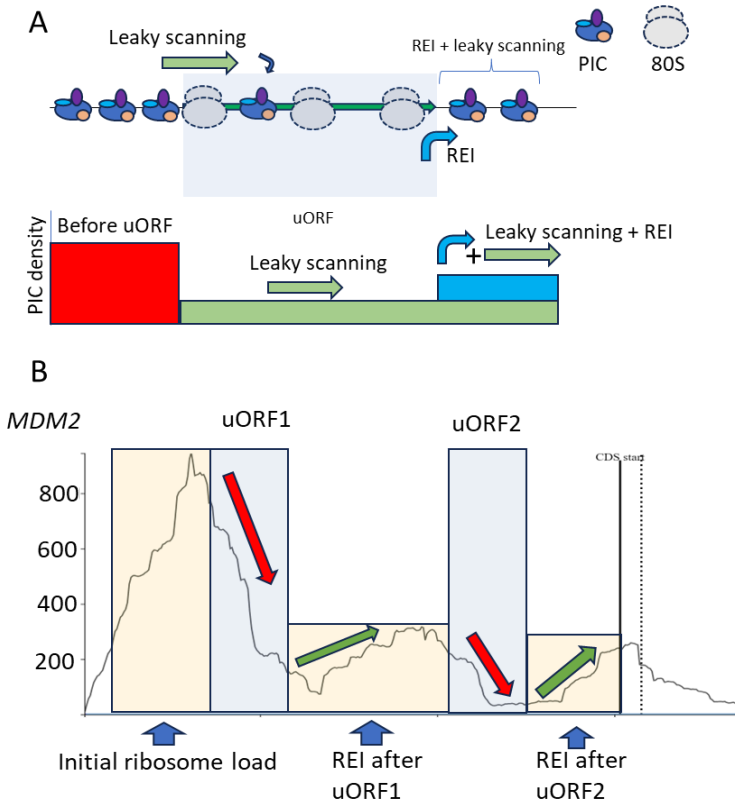


Figure 7.2. A) A schematic of expected changes in PIC distribution along 5' leader with a single uORF. B) TCP-seq profile of *MDM2* 5' leader. Red and green arrows show PIC density changes within regions of interest. uORFs are highlighted with blue boxes.

We chose *MDM2* as the first example, as *MDM2* contains two uORFs (Figure 7.2B). Previous experiments with reporter constructs demonstrated that mutation of the uORF1 start codon resulted in a ~4-fold increase in reporter activity while removal of the uORF2 start codon had a modest ~1.5-fold effect [27]. TCP-seq data show that high PIC signal significantly decreases within the uORF1 and partially restores in the spacer region between uORF1 and uORF2 (Figure 7.2B). This indicates that uORF1 is indeed repressive, and both leaky scanning and moderate levels of reinitiation contribute to PICs that arrive at the uORF2 start codon. However, a different pattern is observed for uORF2 – PIC density drops to baseline within uORF2 but is almost completely restored after it. Therefore, one can conclude that uORF2 promotes no leaky scanning but it supports very efficient reinitiation.

To further illustrate the high variability of uORF regulatory modalities we compare TCP-seq profiles of *TAF7* (Figure 7.3A) and *UCP2* (Figure 7.3B) mRNA 5' leaders. *TAF7* contains 70

codons long uORF. The PIC density upstream, downstream and within this uORF is roughly equal which suggests that this uORF is not inhibitory and its start is very leaky. The *UCP2* 5' leader contains a single uORF which contains three start codons in the optimal nucleotide context. Experiments with reporter constructs demonstrate more than 10-fold inhibitory activity of this uORF [28]. TCP-seq profile of *UCP2* 5' leader shows a dramatic gradual reduction of PIC density within the uORF and a slight increase downstream. This is consistent with the expectation that three start codons almost completely preclude leaky scanning and that the uORF does not allow for efficient reinitiation.

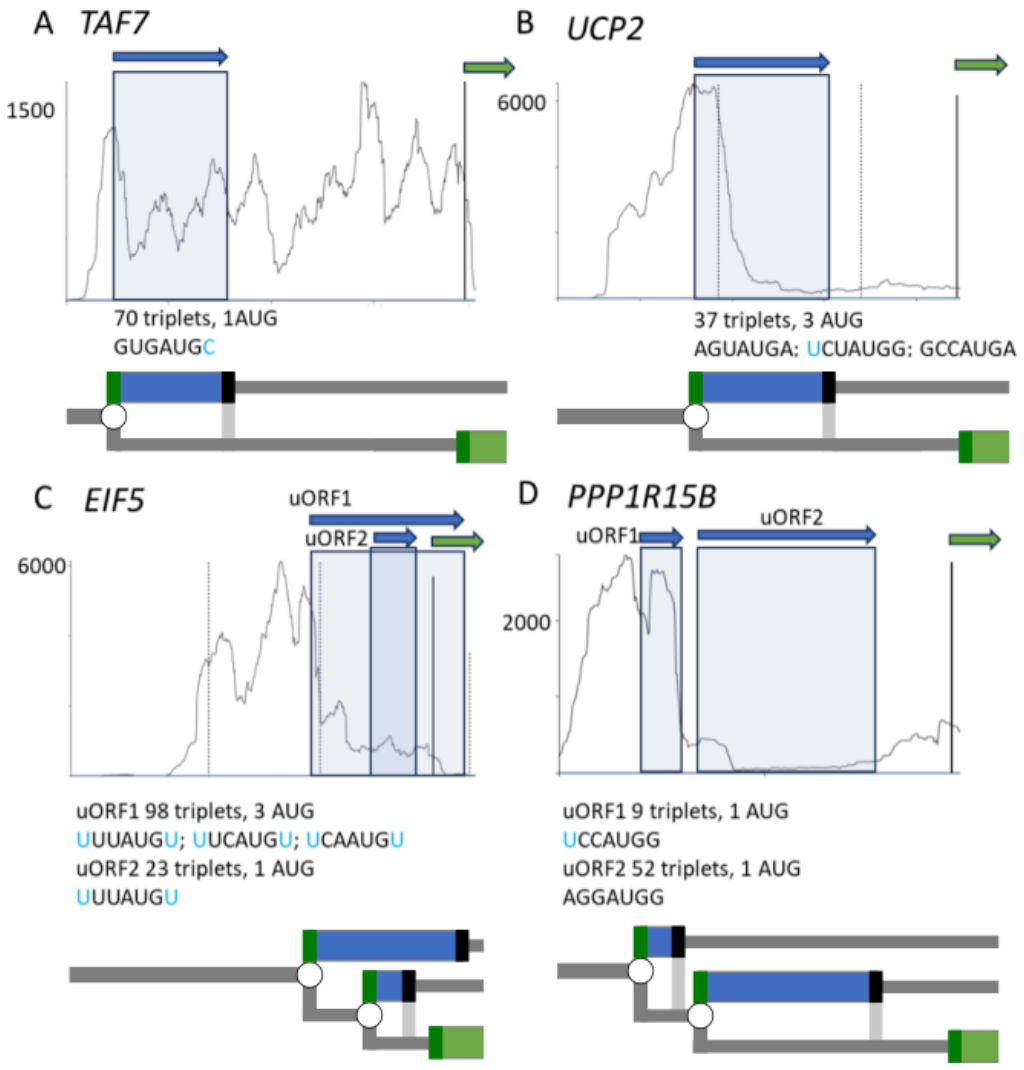


Figure 7.3. TCP-seq profiles of selected 5' leaders. uORFs are highlighted with blue boxes and blue arrows, green arrows show the beginning of CDS. The length of each uORF and its AUG start codons are provided below each graph. Nucleotides of non-optimal start codon contexts are shown in blue. Ribosome Decision Graphs [29] below each panel show the potential ribosome paths that may originate from the translation of the translons annotated with arrows.

Leaky scanning through uORF start codons can operate within regulatory circuits. It was shown that high levels of eIF5 initiation factor induces translation of inhibitory upstream open reading frames (uORFs) in its own mRNA that initiate at AUG codons in conserved poor contexts, and this, in turn, represses eIF5 production [30]. TCP-seq data suggest that PIC

density gradually drops ~5-fold inside long uORF1 overlapping CDS consistent with inefficient leaky scanning through uORFs AUG codons (Figure 7.3C).

A different pattern of PIC distribution is observed for *PPP1R15B* (Figure 7.3D). This 5' leader contains two uORFs. According to experiments with reporter constructs the first shorter uORF1 is not repressive and the second longer uORF2 downregulates translation ~ 8 fold [31]. TCP-seq profiling shows that PICs almost completely disappear within uORF2. This indicates that initiation at uORF2 start is very efficient and uORF2 start is not leaky. Instead, PICs reappear after uORF2 which is consistent with reinitiation.

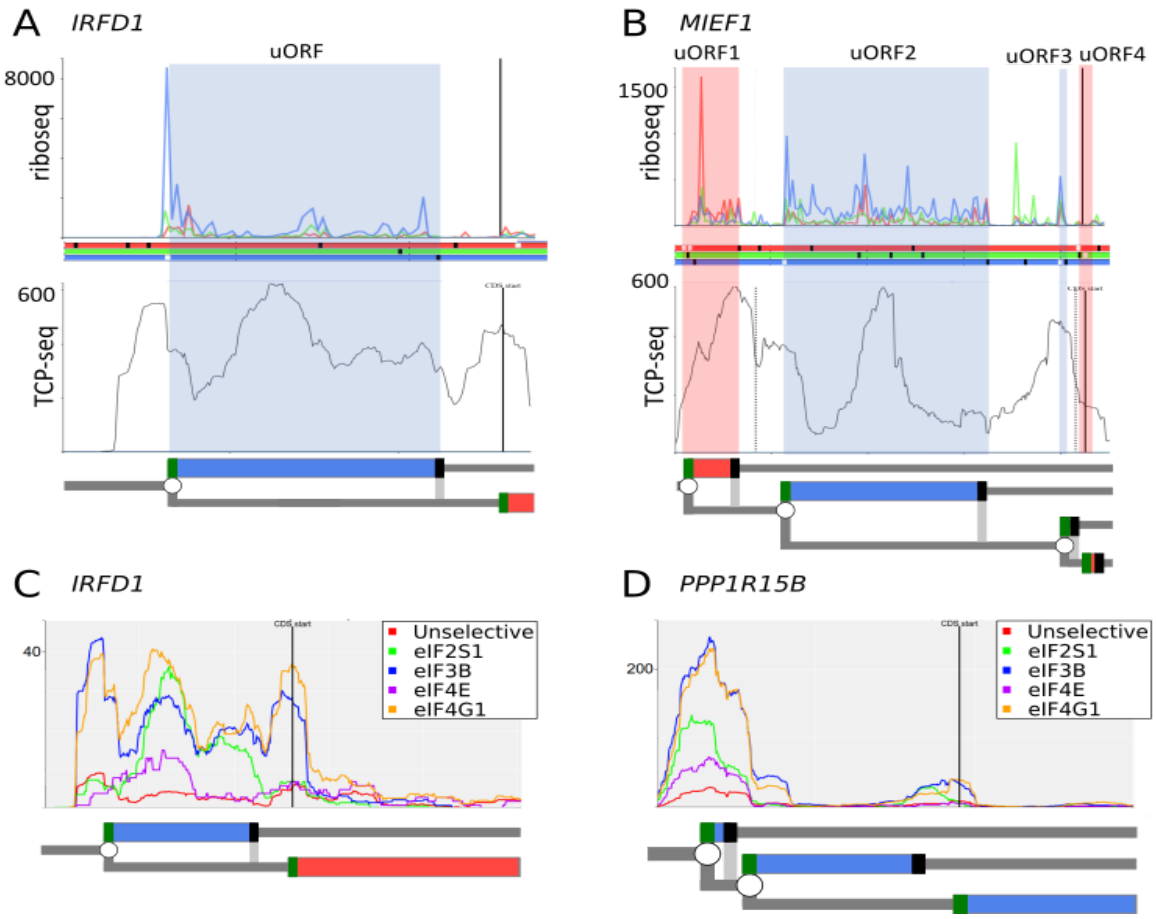


Figure 7.4. A and B: Riboseq and TCP-seq profiles of *IFRD1* and *MIEF1* 5' leaders. Top panels are aggregated ribosome profiling data from the Trips-Viz browser. uORFs are highlighted with coloured boxes which correspond to three reading frames (middle panel, white dashes correspond to AUG codons, black dashes – stop codons). Below are Ribosome Decision Graph representations of the depicted loci showing the potential ribosome paths that lead to the data shown in the profiles above. All plots share the same x-axis of coordinates of the transcript isoform. Branch points in the RDG correspond to the AUG codons shown in the ORF plot. C and D: Selective TCP-Seq profiles of *PPP1R15B* and *IFRD1* 5` leaders. Both plots show the same tracks for two different transcripts. Each trace is coloured according to the initiation factor that was selected. Samples aggregated in each trace are recorded in Supplementary Table 1. The read counts represented on the Y axis are normalised by the factor difference in the number of mapped reads between each track akin to the scaling factor used in transcripts per million (TPM) calculations.

In addition to reinitiation and leaky scanning, we also observed unusual patterns in PIC distribution that point to the existence of a novel mode of uORF-mediated translation control. An intriguing example is found for *IFRD1*. This 5' leader contains a single uORF which represses translation ~8-fold in reporter constructs [31]. Contrary to the expectation, we found

an accumulation of PIC density within the uORF (Figure 7.4A). Another example was found in the *MIEF1* 5' leader, where PICs form a peak in density within the long uORF2 (Figure 7.4B). We propose that PIC presence within inhibitory uORF is the result of the queuing of PICs and 80S ribosomes engaged in uORF translation. Probably, PICs do not always dissociate from mRNA upon collisions with 80S but the scanning rate within the uORF is decreased, and this in turn results in translation repression of the main ORF. We hypothesise that such stalled PICs contain additional translation initiation factors which stabilise them on uORF mRNA sequence, and mRNA nucleotides within ribosome mRNA binding cleft can also stabilise mRNA-ribosome interactions. Such a hypothetical mechanism can be called slow leaky scanning.

Re-emergence of scanning complexes in TCP-Seq after the translation of an upstream translon signifies the potential for reinitiation. However, it is the association of initiation factors that dictates the reinitiation ability. The dissociation of the 40S subunit upon termination precludes downstream translation but the continued association of the 40S with the mRNA is not sufficient for reinitiation on its own and the recruitment of initiation factors is still required.

Selective TCP-Seq (Sel-TCP-Seq), where 40S complexes bound to proteins of interest (initiation factors eIF3B, eIF4E, eIF2S1 and eIF4G1 in the Bohlen dataset) appear to increase the amplitude of the peaks and troughs in the profiles around translons. We posit that this is due to the additional immunoprecipitation step increasing the abundance of bonafide reads from PIC-associated 40S relative to non-PIC originating reads. All initiation factors selected for in the Bohlen study, excluding eIF4E, exhibit this increased amplitude downstream of the translon. However, the Sel-TCP-Seq profile for eIF2S1-associated 40S complexes differs considerably from the profiles of the structural subunits of the scanning complex that were also profiled (eIF3B and eIF4G1) in some cases. The eIF2 complex plays an essential role in the recruitment of ternary complexes to the P-site during translation initiation making its presence required for reinitiation to occur.

Figure 7.4B shows two transcripts where the relationship between eIF2S1 and eIF4G1/eIF3B differs. In *PPP1R15B*, eIF3B, eIF4G1 and eIF2S1 re-appear after translation of the upstream translon and this enables the translation of the CDS (as described above). Throughout the

translation of the upstream translon in IFRD1, eIF3B and eIF4G1 remain associated with scanning complexes at a relatively high rate indicating leaky scanning occurs. Following translation termination of this translon, the abundance of eIF3B and eIF4G1 associated 40S complexes increases while eIF2S1 association significantly decreases. Following the hypothesis that scanning complexes queue across this upstream translon it is plausible that eIF2S1 dissociates due to the increased time spent in the queue.

These examples demonstrate how TCP-Seq has the potential to elucidate the leaky scanning/reinitiation dichotomy that remains obfuscated in Ribo-Seq data and how further nuances can be obtained via the comparative examination of Sel-TCP-Seq profiles. However, the applicability of this approach to a given locus depends on several factors.

To identify reinitiation we look for the following: the presence of scanning complexes 5` of the translon start, a reduction in scanning complex abundance within the translon, followed by an increased abundance of 40S's 3` of the translon stop. As a result, the suitability of this approach for the assessment of reinitiation is dependent on the complexity of the potential signal origins and the translon architecture of the locus.

As previously mentioned, the population of reads within a TCP-Seq sample will consist of both PIC-associated 40S subunits and 40S subunits that were crosslinked to the mRNA while part of an elongating 80S ribosome and subsequently dissociated from the large subunit. As a result, there is the potential for reads from elongating ribosomes that were translating a translon to be misinterpreted as scanning complexes inflating the perceived degree to which leaky scanning occurs.

As with any transcriptomic high throughput signal, including Ribo-Seq and RNA-Seq, the exact transcript isoform from which a given read originated is often impossible to determine in isolation and difficult to infer from a sample. This introduces significant complexities when it comes to extrapolating translation dynamics from aligned ribosome-protected fragments from Ribo-Seq and TCP-Seq. In the instance where two transcript isoforms of the same gene are expressed to non-aberrant degrees, the signal observed in a profile of one transcript will contain a signal originating from the other as well as from itself on shared exons. Although we are yet to identify a definitive example of this, the re-emergence of the 40S signal

downstream of a translon stop may originate from an alternative isoform with a shorter leader sequence.

Short reads are notoriously difficult to map to large genomes/transcriptomes due to the ambiguity introduced by shared tracts of sequence across the reference. TCP-Seq datasets have a wider read length distribution than typical Ribo-Seq datasets do and as a result, enjoy a greater degree of mappability. However, there can still be issues mapping reads to certain genes and as a result, translons with no read coverage will be observed even in the event of low initiation probabilities. This has the potential to lead to the misattribution of signal to the reinitiation ribosome path rather than leaky scanning.

These aforementioned confounding factors, among others, have the potential to confound the interpretation of translation dynamics using TCP-Seq and Ribo-Seq. However, existing browsers made available via RiboSeq.org (namely, GWIPs-viz and Trips-Viz) contain much of the necessary data to enable a researcher to identify translation mechanisms despite these factors.

Additional data cannot however increase the number of 5` leaders with suitable translon architecture to enable the application of this approach. The organisation of translated regions is an important feature of many 5` leaders that impacts the regulation of downstream CDS translation. However, there exist instances where the organisation of active translons limits the use of this approach. The first instance is the distance between the stop of translon 1 and the start of translon 2 are too close together. In this case, reinitiation may well occur but not be detectable as no reemergence of TCP-Seq signal is observed in this region. Furthermore, in cases where there are a large number of active translation initiation sites in one 5` leader the relative changes in scanning complex abundances may be difficult to delineate obfuscating the relative contributions of each translon to the change in 40S densities.

7.1.5 Conclusion

We provide examples of how translation complex profiling (TCP-seq) can help uncover translation regulation at individual uORFs under normal growth conditions due to the availability of the data. Our study suggests that TCP-seq may be instrumental in understanding the mechanistic roles of uORF-mediated regulation under stress conditions or in response to certain metabolites. Furthermore, a number of non-canonical translation initiation factors such as MCTS/DENR heterodimer and eIF2D [32, 33], eIF4G2 [34-36], ribosome rescue factor PELOTA [37], RNA binding proteins PRRC2A, B, C proteins [38] and its ortholog Nocte in fruit fly [39] regulate leaky scanning or reinitiation at certain uORFs. Generation of TCP-seq under conditions of knockdown or overexpression of these factors may shed light on how they affect uORF translation globally.

7.1.6 References

1. Kozak, M. (1980). Evaluation of the "scanning model" for initiation of protein synthesis in eucaryotes. Cell. **22**, 7-8. https://doi.org/10.1016/0092-8674(80)90148-8.

2. Hinnebusch, A. G. (2017). Structural Insights into the Mechanism of Scanning and Start Codon Recognition in Eukaryotic Translation Initiation. Trends in biochemical sciences. **42**, 589-611. https://doi.org/10.1016/j.tibs.2017.03.004.

3. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. Science. **352**, 1413-1416. https://doi.org/10.1126/science.aad9868.

4. Shirokikh, N. E. & Preiss, T. (2018). Translation initiation by cap-dependent ribosome recruitment: Recent insights and open questions. Wiley interdisciplinary reviews. RNA. **9**, e1473. https://doi.org/10.1002/wrna.1473.

5. Hinnebusch, A. G. (2014). The scanning mechanism of eukaryotic translation initiation. Annual review of biochemistry. **83**, 779-812. https://doi.org/10.1146/annurev-biochem-060713-035802.

6. Rogozin, I. B., Kochetov, A. V., Kondrashov, F. A., Koonin, E. V. & Milanesi, L. (2001). Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. Bioinformatics. **17**, 890-900. https://doi.org/10.1093/bioinformatics/17.10.890.

7. Suzuki, Y., Ishihara, D., Sasaki, M., Nakagawa, H., Hata, H., Tsunoda, T., Watanabe, M., Komatsu, T., Ota, T., Isogai, T., Suyama, A. & Sugano, S. (2000). Statistical analysis of the 5' untranslated region of human mRNA using "Oligo-Capped" cDNA libraries. Genomics. **64**, 286-297. https://doi.org/10.1006/geno.2000.6076.

8. Pesole, G., Gissi, C., Grillo, G., Licciulli, F., Liuni, S. & Saccone, C. (2000). Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. Gene. **261**, 85-91. https://doi.org/10.1016/s0378-1119(00)00471-6.

9. Davuluri, R. V., Suzuki, Y., Sugano, S. & Zhang, M. Q. (2000). CART classification of human 5' UTR sequences. Genome research. **10**, 1807-1816. https://doi.org/10.1101/gr.gr-1460r.

10. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. (2009). Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proceedings of the National Academy of Sciences of the United States of America. **106**, 7507-7512. https://doi.org/10.1073/pnas.0810916106.

11. Dever, T. E., Ivanov, I. P. & Hinnebusch, A. G. (2023). Translational regulation by uORFs and start codon selection stringency. Genes & development. **37**, 474-489. https://doi.org/10.1101/gad.350752.123.

12. Renz, P. F., Valdivia-Francia, F. & Sendoel, A. (2020). Some like it translated: small ORFs in the 5'UTR. Experimental cell research. **396**, 112229. https://doi.org/10.1016/j.yexcr.2020.112229.

13. Silva, J., Fernandes, R. & Romao, L. (2019). Translational Regulation by Upstream Open Reading Frames and Human Diseases. Advances in experimental medicine and biology. **1157**, 99-116. https://doi.org/10.1007/978-3-030-19966-1_5.

14. Chen, H. H. & Tarn, W. Y. (2019). uORF-mediated translational control: recently elucidated mechanisms and implications in cancer. RNA biology. **16**, 1327-1338. https://doi.org/10.1080/15476286.2019.1632634.

15. Zhang, H., Wang, Y. & Lu, J. (2019). Function and Evolution of Upstream ORFs in Eukaryotes. Trends in biochemical sciences. **44**, 782-794. https://doi.org/10.1016/j.tibs.2019.03.002.

16. Wethmar, K. (2014). The regulatory potential of upstream open reading frames in eukaryotic gene expression. Wiley interdisciplinary reviews. RNA. **5**, 765-778. https://doi.org/10.1002/wrna.1245.

17. McGeachy, A. M. & Ingolia, N. T. (2016). Starting too soon: upstream reading frames repress downstream translation. The EMBO journal. **35**, 699-700. https://doi.org/10.15252/embj.201693946.

18. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. (2016). Upstream ORFs are prevalent translational repressors in vertebrates. The EMBO journal. **35**, 706-723. https://doi.org/10.15252/embj.201592759.

19. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. **324**, 218-223. https://doi.org/10.1126/science.1168978.

20. Mudge, J. M., Ruiz-Orera, J., Prensner, J. R., Brunet, M. A., Calvet, F., Jungreis, I., Gonzalez, J. M., Magrane, M., Martinez, T. F., Schulz, J. F., Yang, Y. T., Alba, M. M., Aspden, J. L., Baranov, P. V., Bazzini, A. A., Bruford, E., Martin, M. J., Calviello, L., Carvunis, A. R., Chen, J., Couso, J. P., Deutsch, E. W., Flicek, P., Frankish, A., Gerstein, M., Hubner, N., Ingolia, N. T., Kellis, M., Menschaert, G., Moritz, R. L., Ohler, U., Roucou, X., Saghatelian, A., Weissman, J. S. & van Heesch, S. (2022). Standardized annotation of translated open reading frames. Nature biotechnology. **40**, 994-999. https://doi.org/10.1038/s41587-022-01369-0.

21. Kozak, M. (2001). Constraints on reinitiation of translation in mammals. Nucleic acids research. **29**, 5226-5232. https://doi.org/10.1093/nar/29.24.5226.

22. Mohammad, M. P., Munzarova Pondelickova, V., Zeman, J., Gunisova, S. & Valasek, L. S. (2017). In vivo evidence that eIF3 stays bound to ribosomes elongating and terminating on short upstream ORFs to promote reinitiation. Nucleic acids research. **45**, 2658-2674. https://doi.org/10.1093/nar/gkx049.

23. Ivanov, I. P., Shin, B. S., Loughran, G., Tzani, I., Young-Baird, S. K., Cao, C., Atkins, J. F. & Dever, T. E. (2018). Polyamine Control of Translation Elongation Regulates Start Site Selection on Antizyme Inhibitor mRNA via Ribosome Queuing. Molecular cell. **70**, 254-264 e256. https://doi.org/10.1016/j.molcel.2018.03.015.

24. Bohlen, J., Fenzl, K., Kramer, G., Bukau, B. & Teleman, A. A. (2020). Selective 40S Footprinting Reveals Cap-Tethered Ribosome Scanning in Human Cells. Molecular cell. **79**, 561-574 e565. https://doi.org/10.1016/j.molcel.2020.06.005.

25. Kiniry, S. J., O'Connor, P. B. F., Michel, A. M. & Baranov, P. V. (2019). Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. Nucleic acids research. **47**, D847-D852. https://doi.org/10.1093/nar/gky842.

26. Archer, S. K., Shirokikh, N. E., Beilharz, T. H. & Preiss, T. (2016). Dynamics of ribosome scanning and recycling revealed by translation complex profiling. Nature. **535**, 570-574. https://doi.org/10.1038/nature18647.

27. Akulich, K. A., Sinitcyn, P. G., Makeeva, D. S., Andreev, D. E., Terenin, I. M., Anisimova, A. S., Shatsky, I. N. & Dmitriev, S. E. (2019). A novel uORF-based regulatory mechanism controls translation of the human MDM2 and eIF2D mRNAs during stress. Biochimie. **157**, 92-101. https://doi.org/10.1016/j.biochi.2018.11.005.

28. Pecqueur, C., Alves-Guerra, M. C., Gelly, C., Levi-Meyrueis, C., Couplan, E., Collins, S., Ricquier, D., Bouillaud, F. & Miroux, B. (2001). Uncoupling protein 2, in vivo distribution, induction upon oxidative stress, and evidence for translational regulation. The Journal of biological chemistry. **276**, 8705-8712. https://doi.org/10.1074/jbc.M006938200.

29. Tierney, J. A. S., Swirski, M., Tjeldnes, H., Mudge, J. M., Kufel, J., Whiffin, N., Valen, E. & Baranov, P. V. (2024). Ribosome decision graphs for the representation of eukaryotic RNA translation complexity. Genome research. **34**, 530-538. https://doi.org/10.1101/gr.278810.123.

30. Loughran, G., Sachs, M. S., Atkins, J. F. & Ivanov, I. P. (2012). Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. Nucleic acids research. **40**, 2898-2906. https://doi.org/10.1093/nar/gkr1192.

31. Andreev, D. E., O'Connor, P. B., Fahey, C., Kenny, E. M., Terenin, I. M., Dmitriev, S. E., Cormican, P., Morris, D. W., Shatsky, I. N. & Baranov, P. V. (2015). Translation of 5'

leaders is pervasive in genes resistant to eIF2 repression. eLife. **4**, e03971. https://doi.org/10.7554/eLife.03971.

32. Young, D. J., Makeeva, D. S., Zhang, F., Anisimova, A. S., Stolboushkina, E. A., Ghobakhlou, F., Shatsky, I. N., Dmitriev, S. E., Hinnebusch, A. G. & Guydosh, N. R. (2018). Tma64/eIF2D, Tma20/MCT-1, and Tma22/DENR Recycle Post-termination 40S Subunits In Vivo. Molecular cell. **71**, 761-774 e765. https://doi.org/10.1016/j.molcel.2018.07.028.

33. Schleich, S., Strassburger, K., Janiesch, P. C., Koledachkina, T., Miller, K. K., Haneke, K., Cheng, Y. S., Kuechler, K., Stoecklin, G., Duncan, K. E. & Teleman, A. A. (2014). DENR-MCT-1 promotes translation re-initiation downstream of uORFs to control tissue growth. Nature. **512**, 208-212. https://doi.org/10.1038/nature13401.

34. Weber, R., Kleemann, L., Hirschberg, I., Chung, M. Y., Valkov, E. & Igreja, C. (2022). DAP5 enables main ORF translation on mRNAs with structured and uORF-containing 5' leaders. Nature communications. **13**, 7510. https://doi.org/10.1038/s41467-022-35019-5.

35. Smirnova, V. V., Shestakova, E. D., Nogina, D. S., Mishchenko, P. A., Prikazchikova, T. A., Zatsepin, T. S., Kulakovskiy, I. V., Shatsky, I. N. & Terenin, I. M. (2022). Ribosomal leaky scanning through a translated uORF requires eIF4G2. Nucleic acids research. **50**, 1111-1127. https://doi.org/10.1093/nar/gkab1286.

36. She, R., Luo, J. & Weissman, J. S. (2023). Translational fidelity screens in mammalian cells reveal eIF3 and eIF4G2 as regulators of start codon selectivity. Nucleic acids research. **51**, 6355-6369. https://doi.org/10.1093/nar/gkad329.

37. Fernandez, S. G., Ferguson, L. & Ingolia, N. T. (2024). Ribosome rescue factor PELOTA modulates translation start site choice for C/EBPalpha protein isoforms. Life science alliance. **7**. https://doi.org/10.26508/lsa.202302501.

38. Bohlen, J., Roiuk, M., Neff, M. & Teleman, A. A. (2023). PRRC2 proteins impact translation initiation by promoting leaky scanning. Nucleic acids research. **51**, 3391-3409. https://doi.org/10.1093/nar/gkad135.

39. Zhang, T., Xue, Y., Su, S., Altouma, V., Ho, K., Martindale, J. L., Lee, S. K., Shen, W., Park, A., Zhang, Y., De, S., Gorospe, M. & Wang, W. (2024). RNA-binding protein Nocte regulates Drosophila development by promoting translation reinitiation on mRNAs with long upstream open reading frames. Nucleic acids research. **52**, 885-905. https://doi.org/10.1093/nar/gkad1122.