Ollscoil na hÉireann, Corcaigh

**National University of Ireland, Cork**

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

# Development of an online computational platform for the analysis of protein synthesis and detection of novel translated regions.

Thesis presented by

Stephen Kiniry

for the degree of

**PhD (Science)**

University College Cork

School of Biochemistry and Cell Biology

**Head of School:** Professor Justin McCarthy

**Supervisor:** Professor Pavel Baranov

2021

# Table of Contents

# 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.

Signed:_________________________

# Abbreviations

| **Abbreviation** | **Term** |
|---|---|
| Trips-Viz | Transcriptome-wide Information on Protein Synthesis Visualized |
| CDS | Coding Sequence |
| ISR | Integrated Stress Response |
| ORF | Open Reading Frame |
| iORF | Internal Open Reading Frame |
| RPF | Ribosome Protected Fragment |
| Ribo-seq | Ribosome profiling |
| mRNA-seq | messenger RNA sequencing |
| uORF | Upstream open reading frame |
| GUI | Graphical User Interface |

# Acknowledgements

I would first like to thank my supervisor, Professor Pavel Baranov both for providing me with the opportunity to join the lab as well as his expertise and limitless support which was invaluable in carrying out my work. I would also like to thank all of my collaborators as well as colleagues in the LAPTI lab, in particular Dr Audrey Michel, for guidance, support and valuable discussion that helped me immensely in completing my dissertation. Finally, I would like to express my sincere gratitude to my family, in particular my parents, for always encouraging me and providing support in many different ways.

# Research Goals

A number of publicly available resources exist that host processed ribosome profiling data. These resources are useful in allowing researchers to quickly investigate publicly available data and extract useful information such as evidence for the presence of a novel translated ORF. However, most of these resources have deficits in one or more areas, most notably in the use of static offsets and lack of reading frame colourisation for visualisation which makes interpretation difficult to impossible in some specific cases. The goal of this thesis was to create a platform that allows users to explore and analyse various aspects of ribosome profiling data  that would be easy to use and potentially more useful than existing resources, particularly when it comes to visualisation and detection of novel translated ORFs.

# Abstract

Ribosome profiling is a technique that allows us to capture and sequence mRNA fragments protected by ribosome complexes. Mapping these ribosome protected fragments or RPFs, back to a genome or transcriptome provides information on the precise location of elongating ribosomes. This data can then be used to detect novel translated regions, translational pausing and differentially translated genes.

Chapter 2 describes the development of Trips-Viz, an interactive online platform for the exploration and visualisation of RPFs mapped to the transcriptomes of various different organisms. This allows users to rapidly aggregate and visualise ribosome profiling data at a single transcript level allowing for visual detection of translated open reading frames. Trips-Viz also allows users to rapidly assess the quality of data through various meta-information plots as well as detect and visualise transcripts that are differentially expressed/translated between two conditions. These analyses can be carried out through a GUI, meaning users do not need any prior coding or command line experience to be able to use them.

Chapter 3 describes the major updates made to Trips-Viz since its original publication. This includes the addition of mass spectrometry data. Several thousand human mass spectrometry datasets have been processed and detected peptides mapped to the human transcriptome in the same manner as ribosome profiling data. This allows users to corroborate the evidence from the ribosome profiling data and provides information on whether a translated ORF is capable of producing a stable protein product. The differential expression/translation detection has also been improved with the inclusion of the Deseq2 and Anota2seq software. A method for the automatic detection of translated ORFs was also included which allows users to find translated uORFs, nested ORFs, downstream ORFs in a relatively timely manner. Other improvements include the addition of help videos to guide users through the

navigation and interacting with the users interface of Trips-Viz. Finally, incorporating the relevant scripts into RiboGalaxy made it easier for users to upload their own data and transcriptomes to Trips-Viz without any requirement for command line expertise.

# Chapter 1

## Computational methods for ribosome profiling data analysis.

*This chapter has been published in Wiley Interdiscip Rev RNA, Volume 11, Issue 3, e1577(Kiniry et al., 2020). All authors worked on writing the original draft as well as reviewing and editing of the manuscript.*

Since the introduction of the ribosome profiling technique in 2009 its popularity has greatly increased. It is widely used for the comprehensive assessment of gene expression and for studying the mechanisms of regulation at the translational level. As the number of ribosome profiling datasets being produced continues to grow, so too does the need for reliable software that can provide answers to the biological questions it can address. This review describes the computational methods and tools that have been developed to analyse ribosome profiling data at the different stages of the process. It starts with initial routine processing of raw data and follows with more specific tasks such as the identification of translated open reading frames, differential gene expression analysis, or evaluation of local or global codon decoding rates. The review pinpoints challenges associated with each step and explains the ways in which they are currently addressed. In addition it provides a comprehensive, albeit incomplete, list of publicly available software applicable to each step, which may be a beneficial starting point to those unexposed to ribosome profiling analysis. The outline of current challenges in ribosome profiling data analysis may inspire computational biologists to search for novel, potentially superior, solutions that will improve and expand the bioinformaticians' toolbox for ribosome profiling data analysis.

## 1.1 Introduction

Ribosome profiling or Ribo-Seq, involves the arrest of translating ribosomes (using translation inhibitors or other methods) as they traverse mRNA (Ingolia et al., 2009). A nuclease is then used to break down any section of mRNA not being protected by a ribosome (Figure 1.1a). The remaining protected fragments of mRNA (footprints) can then be isolated, sequenced, and mapped to a reference transcriptome or genome. These footprints are approximately 30 nucleotides in length and when mapped can provide both quantitative as well as qualitative information on translation, see (Andreev et al., 2017; Brar et al., 2015; Ingolia et al., 2014; Ingolia et al., 2018; Michel et al., 2013) for reviews. Common applications of Ribo-Seq data analysis include translated Open Reading Frame (ORF) detection, ribosome stalling/pause site detection, and differential gene expression analysis (Figure 1.1b).

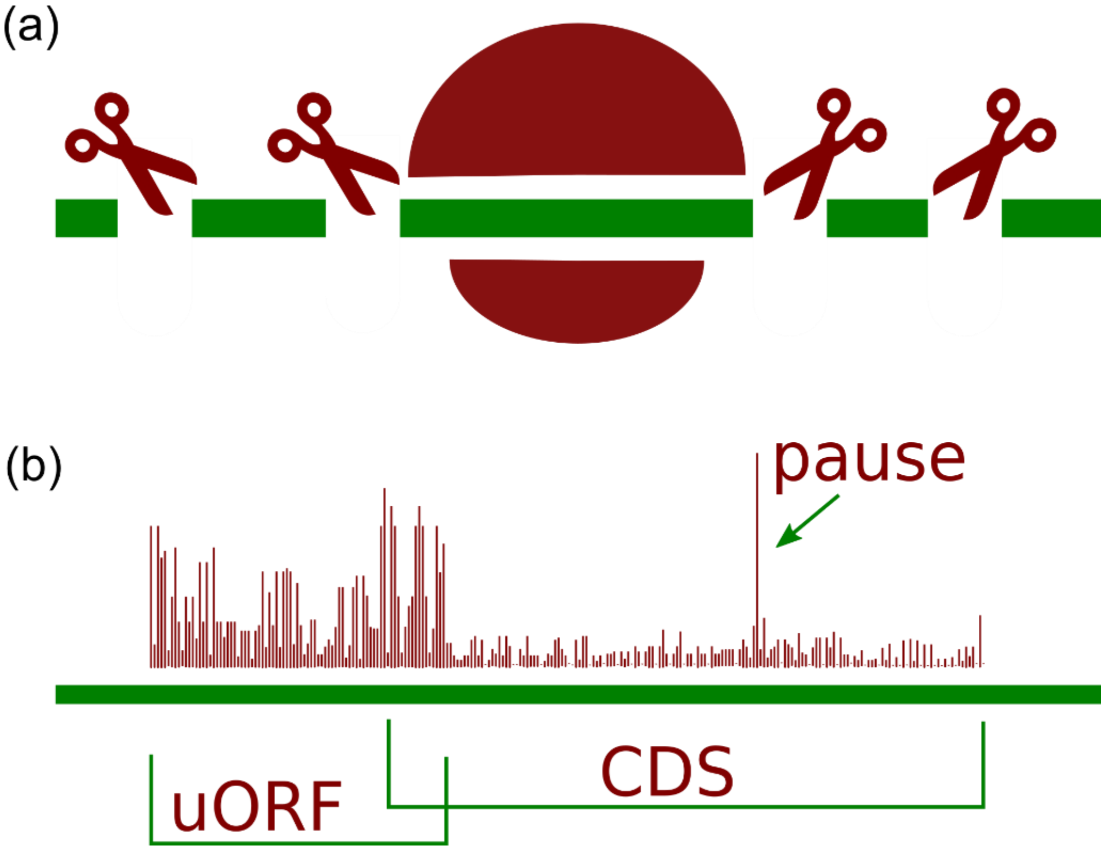

Figure 1.1: *The principle of ribosome profiling. (a) The ribosome protects mRNA from nuclease digestion. The sequences of the protected fragments (footprints) constitute ribosome profiling data. (b) A schematic example of a ribosome footprints density plot (ribosome profile). It shows positions of ribosome decoding centres (brown columns) inferred from sequences of ribosome footprints along an RNA transcript (green bar). The height of the columns reflects the number of footprints matching the corresponding mRNA position. The density suggests the efficient translation of an upstream Open Reading Frame (uORF) overlapping the annotated protein coding region (CDS) and the presence of a ribosome pause site in the CDS.*

The detection of translated regions of a genome is a task for which ribosome profiling is particularly well suited. Translation can be identified even at ORFs  consisting of only a start and a stop codon (Tanaka et al., 2016). Depending on the dataset this can be achieved at sub-codon resolution, meaning that even overlapping translated open reading frames (ORFs) can be detected (Michel et al., 2012). Even though the human genome has been sequenced a while ago, novel protein coding ORFs continue to be discovered, e.g. an upstream ORF (uORF) in the human *MIEF1* gene was predicted to code for a protein (Andreev et al., 2015) and was later found to be an assembly factor of mitochondrial ribosomes (Brown et al., 2017) and more recently characterized as the main product of *MIEF1* mRNA (Rathore et al., 2018).

Ribosome stalling/pause sites can also be characterized. A ribosome moving along an mRNA can pause or stall, blocking the path of other ribosomes, and thus regulate protein synthesis (Ivanov et al., 2018; Kurian et al., 2011; Yordanova et al., 2018) or trigger No-Go decay or Ribosome Quality Control pathways, see (Brandman et al., 2016; Buskirk et al., 2017; Inada, 2013) for reviews. Since ribosomes are more likely to occupy pause sites, more footprints are produced from these locations. Thus, the pause sites appear as local peaks of ribosome footprint density and can be detected computationally.

Another popular (though not unique) application of ribosome profiling is the quantitative characterization of differential gene expression, as it discriminates changes in mRNA translation from changes in mRNA levels. Translation regulation can also be assessed with polysome profiling where the levels of mRNA found in heavy polysome fractions are compared with total mRNA levels. The Ribo-Seq advantage over polysome profiling is that it provides information on the translation of a specific ORF (or ORFs) within an mRNA, however it has its own limitations, see (Gandin et al., 2016) for a comparison of the two approaches. Since ribosome profiling generates millions of sequencing reads the processing and analysis of the data requires intensive computation. The signal produced with ribosome profiling is far more complex and richer in potential applications than standard RNA-seq. Numerous computational approaches have been developed, see (Calviello et al., 2017) for review. We structured this review by detailing the steps carried out for ribosome profiling data analysis and specific goals and overview software that has been developed for these tasks. The accession information for the software tools and/or its sources are provided in tables that are separated into categories. Many tools are multifunctional and could be placed in more than one category while some tools are unique. The selection of software for this review is based on published literature rather than on usability, since testing and benchmarking all published software is an onerous task that should be carried out separately.

## 1.2 Technical considerations when processing raw sequencing reads

Raw ribosome profiling data are usually single-end unprocessed sequencing reads in FASTQ format that need to be processed and mapped to a reference genome or transcriptome. Processing of the reads typically involves removal of adapter/linker sequences as well as removal of any reads aligning to ribosomal RNA (rRNA) and/or transfer RNA (tRNA). There are many freely available tools for both removing adapters and aligning short reads. For example, cutadapt (Martin, 2011) is commonly used to remove adapters, while bowtie (Langmead et al., 2009) and STAR (Dobin et al., 2013) are commonly used for alignment. As these are not specific to ribosome profiling, they will not be discussed in detail here. However, there are several variable parameters involved in both processing and mapping which may significantly affect downstream analysis. Therefore, the initial read processing and alignment should be guided by how the data will be utilised downstream.

To reduce the mapping of non ribosome protected fragments, footprints whose lengths are below a certain threshold are usually discarded. This is done under the assumption that such shorter reads consist of RNA fragments other than those protected by the ribosome or of over-digested footprints. However, such length filtering needs to be applied with caution, because the length of footprints may depend on their sequence and location, e.g. in bacteria, footprints derived from ribosomes bound to Shine Dalgarno sequences are longer (O'Connor et al., 2013). Indeed, Allen Buskirk and colleagues have provided strong evidence suggesting that the earlier claim that Shine-Dalgarno sequences cause ribosome pauses in bacteria (Li et al., 2012) may be an artefact of the footprint length selection (Mohammad et al., 2016). It is also important to note that the length of footprints varies considerably across datasets. Most ribosome footprints in eukaryotes are approximately 28-30 nucleotides and this corresponds to the length of mRNA fragments protected by the ribosome in a specific conformation when its A-site is occupied with a tRNA. Such a conformation is stabilized by certain translation inhibitors that bind to the E-site which is empty in the pretranslocational ribosome conformation. This includes cycloheximide which is by far the most widely used inhibitor in ribosome profiling studies. However, in a posttranslocational conformation, when the A-site is unoccupied, eukaryotic ribosomes protect shorter (20-22 nucleotides) fragments and such fragments could become predominant if different inhibitors are used, such as anisomycin which inhibits the peptidyl transferase reaction (Lareau et al., 2014; Wu et al., 2019). Scanning ribosomes also leave footprints of varying length depending on their specific conformations (Archer et al., 2016). The heterogeneity of ribosome footprint lengths is further exacerbated by suboptimal nuclease digestion which may lead to over or under-digestion of footprints.

The mapping of ribosome footprints to genomic sequences poses yet another problem, namely the mapping across exon-exon junctions. While this is an issue for most techniques involving the sequencing of RNA, it is particularly acute for ribosome profiling due to the short length of Ribo-Seq reads. This results in a systematic bias manifested in reduced unambiguous mappings at exon-exon junctions. This can be clearly seen in the GWIPS-viz browser multiregion view (Kiniry, Michel, et al., 2018; Michel, Fox, et al., 2014). There are splice-aware aligners that are capable of mapping across exon-exon junctions, but the short length of ribosome footprints increases the chance of spurious mappings. One solution to this is to simply map ribosome footprints to transcriptome sequences, but this may not be desired when the accuracy of the transcriptome is in doubt or when its completeness is critical for downstream analysis, e.g. during the identification of novel translated regions.

PCR amplification of footprints during cDNA library generation is also a potential problem. While this bias is pertinent to many techniques requiring PCR amplification, it could be particularly acute for certain Ribo-Seq applications which rely on the accuracy of local footprint density measurements such as detection of ribosome pauses or estimation of codon decoding rates. Recent studies have started to solve this issue with the use of random barcodes introduced to cDNA during the first round of RT PCR reaction. Such barcodes are termed Unique Molecular Identifiers (UMI) and have been used in many applications (Islam et al., 2014; Kivioja et al., 2011). To our knowledge UMIs were first introduced to ribosome profiling by Miettinen and Björklund (Miettinen et al., 2015) and now are part of standard ribosome profiling protocol (McGlincy et al., 2017). During data processing, reads with the same sequence that also share the same UMI are considered to be PCR duplicates and counted as one. This can be done with specific software such as UMI tools (Smith et al., 2017). As the use of UMIs in Ribo-Seq studies is still relatively recent, it is difficult to assess how much of a problem is PCR duplication, though some studies suggest that with sufficient input material and low number of PCR cycles, PCR duplicates constitute only a small fraction of sequencing reads in Ribo-Seq data (Lecanda et al., 2016; McGlincy et al., 2017).

Accession information for software pipelines that can be used for data processing can be found in Table 1, however, certain software packages described later also contain pipelines for raw data processing and quality assessment.

Table 1.1: *Software environments for data processing, pipelines for quality assessment, offset detection and miscellaneous software.*

| Name | Notes | URL | Ref. |
|---|---|---|---|
| mQc | Quality assessment, part of PROTEOFORMER pipeline | https://github.com/Biobix/mQC | (Verbruggen et al., 2018) |
| Plastid | Python based library | https://plastid.readthedocs.io/en/latest/ | (Dunn et al., 2016) |
| Rfoot | Inference of RNA-binding protein sites | https://github.com/zhejilab/Rfoot | (Ji, 2018) |
| Ribodeblur | Offset determination | https://github.com/Kingsford-Group/ribodeblur | (Wang et al., 2017) |
| RiboGalaxy | Galaxy based environment | https://ribogalaxy.ucc.ie/ | (Michel et al., 2016) |
| Ribopip | Ruby based processing pipeline | https://github.com/stepf/RiboPip | (Stefan, 2016) |
| Riboprofiling | R based processing pipeline | http://bioconductor.org/packages/release/bioc/html/RiboProfiling.html | (Popa et al., 2016) |
| RiboProp | Offset determination | http://bioserv.mps.ohio-state.edu/RiboProP/ | (Zhao et al., 2018) |
| RiboseqR | R based processing pipeline | http://bioconductor.org/packages/release/bioc/html/riboSeqR.html | (Chung et al., 2015) |
| RibostreamR | Web based analysis of user generated Ribo-Seq data | https://github.com/pjperki2/riboStreamR | (Perkins et al., 2019) |
| RiboWaltz | Offset determination | https://github.com/LabTranslationalArchitectomics/riboWaltz | (Lauria et al., 2018) |
| Ribo-seQC | Quality assessment | https://github.com/ohlerlab/RiboseQC | (Calviello, Sydow, et al., 2019) |
| RRS | measures drop-off of ribosome footprint density at the end of ORFs | https://rdrr.io/github/JokingHero/ORFik/man/ribosomeReleaseScore.html | (Guttman et al., 2013) |
| SystemPipeR | R based processing pipeline | https://bioconductor.org/packages/release/bioc/html/systemPipeR.html | (Backman et al., 2016) |
| Trips-Viz | Web based analysis of public and user generated ribosome profiling data | https://trips.ucc.ie | (Kiniry, O'Connor, et al., 2018) |
| ShoeLaces | Offset determination and visualisation | https://bitbucket.org/valenlab/shoelaces | (Birkeland et al., 2018) |
| XPRESSyourself | Processing pipeline and visualisation | https://github.com/XPRESSyourself/ | (Berg et al., 2020) |
| RiboDoc | Docker based pipeline | https://github.com/equipeGST/RiboDoc | (Francois et al., 2021) |
| ORFik | Bioconductor package that can carry out differential expression and ORF detection as well as analyse other data types such as CAGE | http://bioconductor.org/packages/release/bioc/html/ORFik.html | (Tjeldnes et al., 2021) |

| | | | |
|---|---|---|---|
| RiboToolKit | A comprehensive web based platform that can carry out a number of different analyses including translated ORF detection, differential expression and codon occupancy | http://rnabioinfor.tch.harvard.edu/RiboToolkit | (Liu et al., 2020) |

## 1.3 Global assessment of the data quality

Assessing the quality of the data should be viewed as an obligatory requirement after initial pre-processing and mapping, as it saves wasted time trying to draw conclusions from poor quality data. Four relatively simple approaches are commonly utilised to achieve this; analysis of read length distributions, metagene profiles, a breakdown of regions to which Ribo-Seq reads align, and the triplet periodicity signal. Depending on the type of nuclease used for digestion, Ribo-Seq reads will also display a periodicity signal, with reads tending to map to every 3rd nucleotide. The majority of Ribo-Seq reads will also tend to map to annotated coding regions, this can be assessed with metagene profiles or looking at the amount of reads mapping to coding/non-coding regions. Deviations from these tendencies indicate that the Ribo-Seq data may be of poor quality. Other more general approaches include assessing the correlation among replicates and the number of useful mapped reads. The implementation of these features have become "*de facto*" best practice and while they are indicative of quality they should not be viewed as definitive.

A typical ribosome profiling dataset obtained from eukaryotic cells is characterized by a sharp distribution of lengths with a predominant length around 28-30 nucleotides. The variation depends on the nuclease digestion conditions and the inhibitors used (see above). The distribution is wider for ribosome profiling datasets obtained from bacterial cells due to the read length distribution associated with Shine-Dalgarno interactions (O'Connor et al., 2013). The read length distribution can be analysed with a number of tools, for example, FastQC (Andrews, 2010), a general tool for assessing the sequence quality of reads obtained with high throughput sequencing. FastQC also can be used to evaluate the accuracy of base calls and to quantify positional nucleotide frequencies, GC content and over-represented sequences. These analyses can often uncover problematic features such as the frequent addition of untemplated nucleotides during reverse transcription, untrimmed adapter sequences, etc.

Another important way to assess the quality of the datasets is with a *metagene profile*. The metagene profile provides the frequency of footprints relative to all annotated start and stop codons. There are several ways to generate metagene profiles. One is to simply count the frequency of all footprints (using a single footprint position, i.e. the 5’ or 3’ end) at a specific coordinate relative to the annotated start codons (or stops) of all transcripts. The procedure for building a metagene profile relative to start codons could be represented as $\boldsymbol{D(i)=\Sigma_K(d_k(i+s_k)),}$ where $\boldsymbol{D}$ is a metagene footprint density, i is the coordinate of metagene profile, $\boldsymbol{d_k}$ is a footprint density at a transcript $\boldsymbol{k}$ from transcriptome $\boldsymbol{K}$ with $\boldsymbol{s_k}$ being the coordinate of the annotated start. A potential issue with such a representation is that highly expressed mRNAs could dramatically skew the metagene profile. To mitigate this issue, the frequency of footprints could be normalized across individual mRNAs, so that they have equal influence on the overall picture. It is also possible to normalize the CDS length and analyse the frequency of footprints of different lengths, producing a very informative translatome representation as has been done by Thomas Preiss and colleagues (Archer et al., 2016). A metagene profile of a high-quality ribosome profiling dataset is expected to have a sharp difference in footprint density at the start and stop codons, so that the density is higher downstream of starts and upstream of stops (Figure 1.2). For the generation of metagene profiles in bacteria it is important to exclude overlapping CDS regions as well as closely located CDS regions to avoid signal interference. In a similar vein, the generation of metagene profiles in higher eukaryotes necessitates selecting a single transcript isoform where multiple isoforms exist to avoid an artificial amplification of footprints counts by the number of splice isoforms. Ideally the translated transcript isoform(s) at a gene locus should be used. However, metagene profile generation is typically carried out early in the Ribo-Seq data analysis process and isoform delineation, if required, performed further downstream. Hence heuristic approaches are often used such as selecting “principal isoforms” from the APPRIS database (Rodriguez et al., 2018). Other heuristic approaches of a single representative transcript selection and their limitations are discussed later in relation to differential expression analysis.

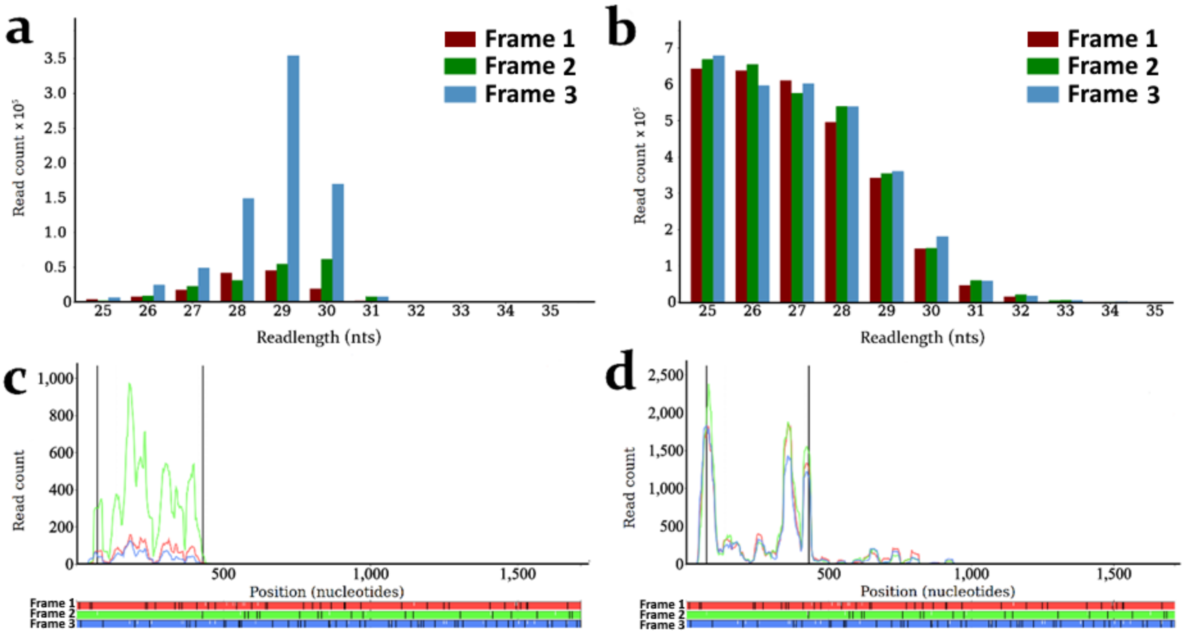

Figure 1.2: *Assessment of ribosome profiling data quality (a, b) Triplet periodicity plots that show the number of footprints aligning to one of the three subcodon positions (differentially colored) for each subcodon position. (a) An example of good quality data showing strong periodicity and desirable read length distribution. (b) An example of data showing no triplet periodicity and an unexpected read length distribution. (c, d) Sub-codon ribosome profile of an ENSEMBL transcript expressed from the human B2M locus visualized with Trips-Viz. The ORF plot at the bottom shows three reading frames (differentially colored) with white dashes for AUG codons and black dashes for stops. The annotated CDS is demarked by the vertical black lines in the main plot and corresponds to the second reading frame. The footprint density is shown separately depending on the sub-codon phase of the aligned reads as curves that are colored to match the color of the supported reading frames. The reading frame detection is possible in (c), but not in (d) which correspond to (a) and (b) respectively. In addition in (c) the vast majority of reads map entirely within the CDS, while in (d) there are reads which map to the 3' trailer region that are unlikely to be derived from translating ribosomes. For the source of the data see text.*

Triplet periodicity refers to the unequal distribution of read mappings relative to subcodon positions due to the triplet nature of the genetic code: elongating ribosomes move along mRNA in discrete steps of three nucleotides. The strength of the triplet periodicity can be assessed using the frequency with which a single footprint coordinate (e.g. 5' or 3' end) aligns to one of the three subcodon positions. See (Figure 1.3 (a,c)) for an example of a dataset with strong periodicity derived from (Calviello et al., 2016) and (Figure 1.3 (b,d)) for an example of a dataset with poor periodicity (Kirchner et al., 2017). Strong periodicity is a good indicator that the data is genuinely Ribo-Seq data and it can be used for the detection of translated reading frames (Michel et al., 2012), although it is not definitive as even RNA-seq data may indicate some periodicity due to crossover of sequencing biases and GC3 skew. However, the periodicity is dependent on the uniformity of the digestion position relative to the ribosome's decoding centre and thus varies depending on digestion conditions (Gerashchenko et al., 2017) and specifics of the translation apparatus (see above). Thus, the absence of strong periodicity does not necessarily mean that the other useful features of the data are also poor. One way to express the periodicity quantitatively is to calculate the proportion of reads at the predominant subcodon position. Another is to assess the divergence from an equiprobable distribution using Shannon Entropy ($-\Sigma p_i log_2(p_i)$ where $p_i$ is the relative frequency of footprints at the $i$ subcodon position). Shannon Entropy is a metric used in information theory that is used to assess the periodicity of signals. Random signals will tend to have lower entropy, and more ordered signals such as those seen in Ribo-Seq data with strong periodicity, tend to have higher entropy. The periodicity can also be detected with Fourier (Calviello et al., 2016; Chun et al., 2016) and wavelets transformations (Xu et al., 2018). Both metagene profiles and triplet periodicity visualization plots can be produced by many different tools such as RibostreamR (Perkins et al., 2019), Ribo-SeQC (Calviello, Sydow, et al., 2019), RiboGalaxy (Michel et al., 2016), Plastid (Dunn et al., 2016), riboseqR (Chung et al., 2015) and mQC (Verbruggen et al., 2018).

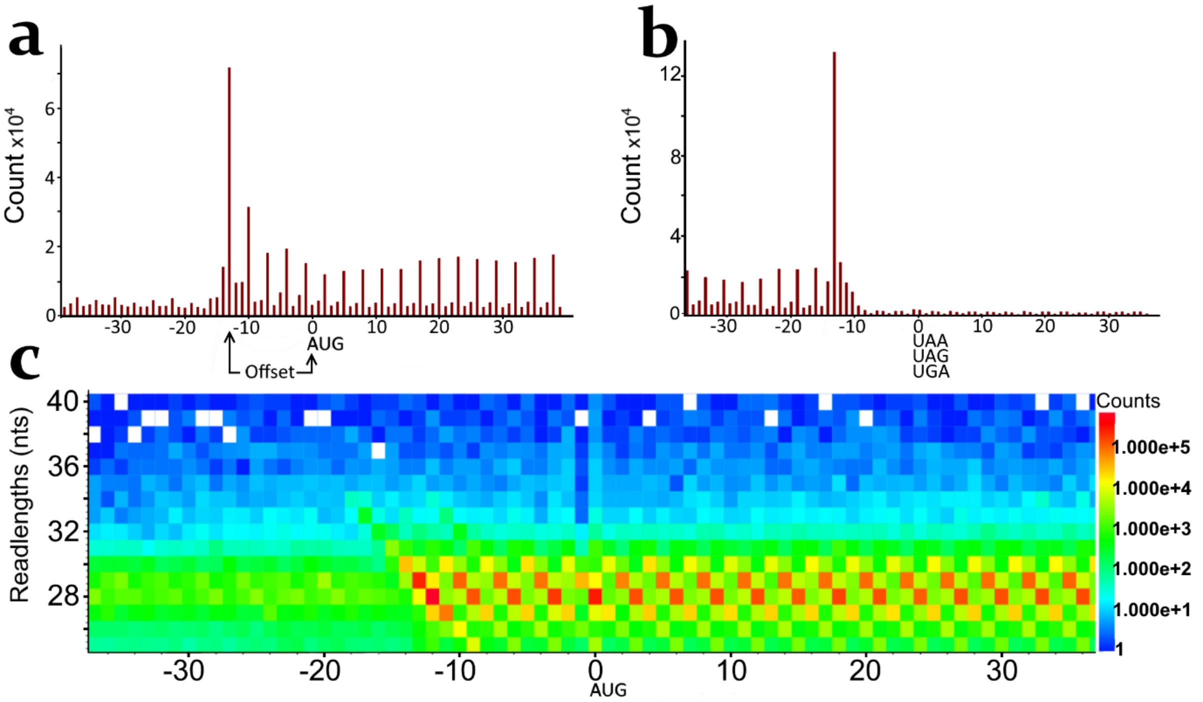

Figure 1.3: *Examples of metagene profiles. (a) The profile was created by aggregating Ribo-Seq counts from a region surrounding the annotated start codon (zero coordinate) of every gene for a single read length. This example shows the positions of footprint 5' ends, but 3' ends may also be used. Since initiation is slower than elongation, a peak of footprint density is expected at the start codon. Thus the location of the 5' end peak density indicates the distance between footprints 5' ends and ribosome P-site codon where tRNA-Met$_i$ is being incorporated (offset). (b) Same as (a) but relative to annotated stop codons (zero coordinate). A drop of footprint density is observed upstream of the stop. (c) A start codon metagene profile constructed as a heatmap has the advantage of displaying multiple read lengths simultaneously. It can be seen that the distance between 5' ends and P-site codons vary depending on read lengths suggesting that different offsets should be applied to the reads depending on their length.*

## 1.4 Determining the position of the decoding center

An offset is typically applied to the sequence of ribosome footprints to infer the position of the A- or P-site of the ribosome that produced it. This is an integer which is added to the coordinate of the 5' end of a mapped read or, alternatively subtracted from the coordinate of its 3' end. The metagene profiles are often used to determine the offset, assuming that the first sharp increase in footprint density corresponds to the footprints of the ribosomes at the start codons. Since the start codons are recognised at the P-site, the distance between this increase and the first nucleotide of the start codon is used as the offset for determining positions of the P-sites, see (Figure 1.2(a)) for a metagene profile made using data from (Calviello et al., 2016). To determine the positions of the A-sites, 3 nucleotides are added if the metagene profile is based on the 5’ ends or subtracted if it is based on the 3’ ends. Typically, when reads are not stratified by read lengths, 5’ end mappings produce a greater triplet periodicity in eukaryotic organisms, while 3’ ends produce greater periodicity in bacteria (Woolstenhulme et al., 2015). This is most likely due to the asymmetric variability of read lengths relative to the decoding centre which in case of bacteria could be attributed to Shine-Dalgarno interactions with anti-Shine-Dalgarno (O'Connor et al., 2013). Applying a 'static' offset regardless of read length is often sufficient to determine positions of A- or P-sites with an accuracy that is satisfactory for numerous Ribo-Seq applications. However, the accurate determination of A or P-site positions is critical for certain applications such as the measurement of ribosome dwell times at specific codons (e.g. estimating codon decoding rates). The accuracy can be further improved with setting specific offsets for each read length, i.e. using separate metagene profiles made for each read length, see (Figure 1.2(c)) for a heatmap made using data from (Albert et al., 2014).

This approach can provide more accurate inferred A- or P-site locations than a static offset and thus improve the periodicity signal. However, in any given dataset there may be read lengths that are not abundant compared to the predominant read length. These low-abundance read lengths are difficult to correctly assign an offset to.

RiboWaltz (Lauria et al., 2018) aims to correct this by using offset values from abundant read lengths to infer the optimal offsets for less abundant read lengths. More sophisticated methods of offset determination have also been developed, (O'Connor et al., 2016) proposed the determination of the offset that maximises the difference of the estimated dwell time between codons. This assumes that the A-site has a predominant role in influencing the decoding rate. Ribodeblur (Wang et al., 2017), uses an expectation maximization-like procedure to obtain a more accurate estimate of A-sites. RiboproP (Zhao et al., 2018) is specifically designed to mitigate the sequence bias introduced from Ribo-Seq data generated with MNase , thus improving offsetting. See Table 1 for accession information to these tools.

## 1.5 Translated ORF detection

The detection of translated ORFs is an application for which ribosome profiling is uniquely well suited, particularly of short ORFs, whose products cannot be easily detected with proteomics techniques. Detecting translation using Ribo-Seq data is not straight forward as the presence of a footprint in a given genomic region does not necessarily mean that that region is being translated. In addition to the artefacts of mapping mentioned previously, not all sequences found in a ribosome profiling cDNA library derive from genuine ribosome protected fragments within the ribosome mRNA channel. In fact, most of the cDNA reads in any ribosome profiling library come from the ribosome itself as its rRNA gets digested during the procedure. Similarly fragments of other RNAs bound to the ribosome could contaminate the sample (fragments of tRNAs are also very abundant). Additional sources of contamination are fragments of RNAs from nucleoprotein complexes that could be coisolated with ribosomal complexes. Thus, the difficult aspect of translated ORF detection is the discrimination of the signal obtained with genuine ribosome footprints from other RNA fragments.

Nonetheless potentially translated regions can often be easily recognized upon manual visual inspection of the corresponding sequence region. Several existing resources provide such functionality such as Svist4get (Egorov et al., 2019) SmProt (Hao et al., 2018), GWIPS-Viz (Michel et al., 2018), Trips-Viz (Kiniry, O'Connor, et al., 2018)  HRPDViewer (Wu et al., 2018) and RiboViz (Carja et al., 2017). Many allow for viewing Ribo-Seq datasets from multiple studies simultaneously which can significantly boost the signal to noise ratio making translated regions easier to detect. Manual visual detection is a simple and straightforward method of translated ORF detection, particularly when the translated ORF is highly translated and does not overlap with others. However, when several ORFs overlap or are nested within each other, their detection based purely on the density of footprints is difficult due to the heterogeneity of the signal within an ORF. Manual visual detection in these cases can be improved when footprints are discriminated based on the phase of their triplet periodicity. This could be done either by generating separate subcodon profiles or using differential colors for the reads depending on their phase relative to subcodon positions as in RiboSeqR (Chung et al., 2015), RiboGalaxy (Michel et al., 2016) or Trips-Viz (Kiniry, O'Connor, et al., 2018), see Table 1.2. The main disadvantage of manual identification of translated ORFs is the low throughput. Manual inspection of even a bacterial genome is impractical. Thus, numerous tools have been developed to enable automatic high throughput detection of translated ORFs using Ribo-Seq data.

Table 1.2: *Data resources and visualization environments.*

| | | | |
|---|---|---|---|
| GWIPS-viz | Genome browser for visualization of Ribo-Seq data aligned to genomes | https://gwips.ucc.ie | (Michel, Fox, et al., 2014) |
| HRPDViewer | A resource for visualization of Ribo-Seq data aligned to transcriptomes | http://cosbi4.ee.ncku.edu.tw/HRPDviewer/ | (Wu et al., 2018) |
| Openprot | Database and viewer for exploration of Ribo-Seq and mass-spec data supporting translation of non-annotated ORFs | https://openprot.org/ | (Brunet, Brunelle, et al., 2018) |
| RiboSeqDB | Repository of human and mouse ribosome profiling data | https://micro.biouml.org/bioumlweb/ | (Liu et al., 2018) |
| RiboViz | Online tool for visualization of publicly available Ribo-Seq data | https://riboviz.org/ | (Carja et al., 2017) |
| RPFdb | Database of ribosome profiling datasets rich in metainformation and their genomic alignments | http://sysbio.gzzoc.com/rpfdb/ | (Xie et al., 2016) |
| sORFs.org | Database of short ORFs whose translation is supported with Ribo-Seq data | http://sorfs.org | (Olexiouk et al., 2018) |
| svist4get | Command-line visualization tool | https://bitbucket.org/artegorov/svist4get/ | (Egorov et al., 2019) |
| TranslatomeDB | On-line resource for visualization of public and user generated data | http://translatomedb.net/ | (Liu et al., 2018) |
| Trips-Viz | On-line environment for graphical exploration of public and user generated ribosome profiling data aligned to transcriptomes. | https://trips.ucc.ie | (Kiniry, O'Connor, et al., 2018) |

They utilize different computational concepts including statistical tests as in Ribo-TISH (Zhang et al., 2017), linear regression as in ORF-RATER (Fields et al., 2015), robustness of triplet periodicity as in RiboTaper (Calviello et al., 2016) and RiboWave (Xu et al., 2018), Hidden Markov models as in RiboHMM (Raj et al., 2016) as well as machine learning techniques, e. g. in REPARATION for bacterial genome reannotations (Ndah et al., 2017).

An in-depth analysis of these approaches requires a separate dedicated review. Further we will summarize the most common features of translated ORFs that are often used by these tools to predict their translation.

The similarity between the patterns of footprints in mRNA 5’ leaders and lincRNAs observed in early mammalian datasets provoked a suggestion that translation takes place in RNA transcripts (and their parts) that were normally considered non-coding (Chew et al., 2013; Ingolia et al., 2011). In response to this claim Gutman et al (Guttman et al., 2013) developed Ribosome Release Score (RRS) which measures the drop of ribosome footprint density downstream of ORF stop codons and have shown that a high RRS score is a signature of annotated protein coding ORFs, but not of ORFs found in 5’ leaders and lincRNAs. While RRS provides a useful metric for estimating the accuracy of translation termination, its use as a sole signature of translation is peculiar since it assumes that no re-initiation or leaky scanning takes place, while both phenomena are well documented in eukaryotic cells, see (Hinnebusch, 2014; Hinnebusch et al., 2016; Shirokikh et al., 2018) for reviews. Re-initiation often takes place after termination at short ORFs and a large fraction of ribosome scanning complexes bypass start codons in a poor initiation context. This leads to a complex organization of short overlapping translated ORFs in the beginning of RNA transcripts where high ribosome density is observed both upstream and downstream of stop codons leading to low RRS scores. While RRS can indeed be used as a signature of ORF translation, since isolated ORFs are expected to exhibit high RRS scores, it is important to be aware of RRS limitations in detecting overlapping or closely located translated ORFs.

Indeed, a follow up study (Ingolia et al., 2014) developed another metric, fragment length organization similarity score (FLOSS) that is based on the similarity of length distributions of footprints across different transcripts and have shown that they can successfully discriminate RNA fragments mapped to genuine non-coding RNAs from those observed at translated ORFs providing further support to the initial claim that ribosomes do translate many short ORFs in 5’ leaders and RNA transcripts previously annotated as non-coding. Rfoot (Ji, 2018) uses the same principle of analysing read length distributions to identify non-ribosomal RNA footprints.

In addition to a characteristic distribution of read lengths in translated ORFs another feature that is strongly associated with translation is triplet periodicity, however, the detection of triplet periodicity is difficult when the ORF length is short due to the high heterogeneity of the signal. To mitigate this issue Calviello et al. (Calviello et al., 2016) designed RiboTaper which is based on the multitaper approach (Thomson, 1982) developed for signal processing that performs a spectral analysis on a signal that has been transformed in a number of different ways (tapers). SPECtre (Chun et al., 2016) is another tool for detecting periodicity based on spectral analysis of aligned Ribo-Seq data developed around the same time. More recently RiboWave (Xu et al., 2018) was developed, which makes use of wavelet transformation to denoise ribosome profiling signal, and claims to outperform previously developed tools. Changes in ribosome footprint density can also be used as signature of translation. In addition to a drop of ribosome density at the ends of ORFs, many datasets exhibit characteristic patterns with elevated ribosome density at the beginning and the end of ORFs and this information can be taken into account when scoring potentially translated ORFs as in RiboHMM  (Raj et al., 2016). The problem of this approach is that such changes in footprint density are often data specific and HMM emission probabilities obtained from the analysis of one dataset may not suit another dataset.

There are certain variations of ribosome profiling methods that enrich ribosomes at the starts of translation initiation using specific translation inhibitors or their combinations (Gao et al., 2015; Ingolia et al., 2011). This information can also be utilized for the detection of translated ORFs as in Ribo-TISH (Zhang et al., 2017) and is especially useful for localisation of start codons at which ORF translation is initiated, as it is often more difficult than detection of translation itself since translation initiation often takes place at non-AUG codons (Ivanov et al., 2011) especially when close to the 5’ ends  (Michel, Andreev, et al., 2014) and sometimes multiple start codons are being used to initiate the same ORF as in *PTEN* (Tzani et al., 2016). Some tools such as Ribo-TISH (Zhang et al., 2017) and the recently developed DeepRibo (Clauwaert et al., 2019) which uses neural networks to annotate bacterial genomes, are capable of utilising both elongating and initiating Ribo-Seq data.

While many tools for predicting translated ORFs exist (see Table 3), their predictions differ considerably. Moreover, it is difficult to make specific recommendations on what software to use in the absence of independent benchmarking studies. Such benchmarking is very difficult to carry out due to a lack of gold standard sets of translated ORFs and adequate methodology orthogonal to ribosome profiling. A set of annotated protein coding genes cannot be used as a gold standard dataset since it is biased towards long ORFs coding for functional proteins. Although mass spectrometry analysis (Van Damme et al., 2014; Vanderperre et al., 2013) and phylogenetic analysis  (Andreev et al., 2015; Bazzini et al., 2014) are being used as orthogonal methodology, neither is truly adequate. Many of the short translated ORFs are unlikely to produce stable peptides that can be detected with mass spectrometry, though efforts have been made to combine proteomics and ribosome profiling evidence such as with Proteoformer (Crappe et al., 2015), Proteoformer 2 (Verbruggen et al., 2019) and OpenProt (Brunet, Brunelle, et al., 2018). Similarly, the signal obtained from phylogenetic conservation depends on the length of the ORF and the depth of its conservation. Translation of some ORFs may not affect fitness and would evolve neutrally. Also, a functional ORF was recently reported for which no evidence of evolutionary selection was found (Xie et al., 2019). Therefore, in the absence of benchmarking standards and appropriate orthogonal methodology, the software described in this section can be used for exploratory analysis only. Despite these limitations ribosome profiling has been used to successfully confirm novel translated regions (Castelo-Szekely et al., 2019; Chugunova et al., 2019; Hardy et al., 2019) and even to discover a novel mechanism of translation regulation (Yordanova et al., 2018).

Table 1.3: *Software tools for automatic detection of translated ORFs.*

| Name | Notes | URL | Ref. |
|---|---|---|---|
| DeepRibo | Detection of translated ORFs in bacterial genomes | https://github.com/Biobix/DeepRibo | (Clauwaert et al., 2019) |
| orfRater | Detection of translated ORFs based on linear regression | https://github.com/alexfields/ORF-RATER | (Fields et al., 2015) |
| ORFScore | Scoring translated ORFs based on triplet periodicity | https://rdrr.io/bioc/ORFik/man/orfScore.html | (Bazzini et al., 2014) |
| PreTis | Detection of translation initiation starts based on linear regression | http://service.bioinformatik.uni-saarland.de/pretis/ | (Reuter et al., 2016) |
| PRICE | Detection of translated ORFs using EM algorithm | https://github.com/erhard-lab/price | (Erhard et al., 2018) |
| Proteoformer | Detection of translated ORFs with support from mass-spec data | https://github.com/Biobix/proteoformer | (Crappe et al., 2015; Verbruggen et al., 2019) |
| REPARATION | Detection of translated ORFs in bacterial genomes | https://github.com/Biobix/REPARATION | (Ndah et al., 2017) |
| Ribocode | Detection of translated ORFs based on triplet periodicity | https://github.com/xryanglab/RiboCode | (Xiao et al., 2018) |
| riboHMM | HMM based detection of translated ORFs | https://github.com/rajanil/riboHMM | (Raj et al., 2016) |
| RibORF | SVM based identification of translated ORFs | https://github.com/zhejilab/RibORF | (Ji et al., 2015) |
| Ribosome profiling analysis framework | Detection of translated ORFs based on triplet periodicity | https://github.com/LUMC/ribosome-profiling-analysis-framework | (de Klerk et al., 2015) |
| RiboTaper | Detection of translated ORFs based on spectral analysis of Ribo-Seq signal using multitaper | https://ohlerlab.mdc-berlin.de/software/RiboTaper_126/ | (Calviello et al., 2016) |
| Ribo-TISH | Is able to use Ribo-Seq data enriched at starts of initiation in addition to regular Ribo-Seq. | https://github.com/zhpn1024/ribotish | (Zhang et al., 2017) |
| RiboWave | Detection of translated ORFs based on spectral analysis of Ribo-Seq signal with Wavelet transformation | https://github.com/lulab/Ribowave | (Xu et al., 2018b) |
| Rp-Bp | Bayesian approach for detecting translated ORFs. | https://github.com/dieterich-lab/rp-bp | (Malone et al., 2017) |
| SPECtre | Detection of translated ORFs based on spectral analysis of Ribo-Seq signal | https://github.com/mills-lab/spectre | (Chun et al., 2016) |
| uORF-seqr | Regression based detection of translated ORFs. | https://github.com/pspealman/uorfseqr | (Spealman et al., 2018) |
| ORFLine | Uses filters based on previously defined features, e.g ORFScore/RRS. | https://github.com/boboppie/ORFLine | (Turner et al., 2021) |

## 1.6 Differential gene expression

Ribosome profiling analysis is probably most frequently used for the characterization of differential gene expression as part of a time series or control/treatment group. It is assumed that slowly and rapidly decoded codons are distributed somewhat equally and therefore the relative frequency with which footprints are mapped to a specific ORF should be proportional to the levels of RNA bearing this ORF and efficiency of translation initiation at this ORF. In other words, the ribosome profiling signal is reflective of the total protein synthesis which accounts for the RNA levels (synthesis and degradation) and the rate of RNA translation. Ribosome profiling experiments usually are carried out in parallel with RNA-seq experiments that allow determination of RNA levels. When RNA levels do not change, but the ribosome profiling signal changes, it is reasonable to attribute these changes to changes in translation efficiencies. Note, however, that changes in local densities could be also caused by ribosome pausing. In this case, the induction of a ribosome pause at a specific location may be misinterpreted as increased translation. For example, Lobanov et al 2017 have noticed that Euplotes mRNAs containing sites of ribosomal frameshifting have a higher ratio of Ribo-Seq to RNA-seq reads than mRNAs translated without frameshifting. They attributed this difference to ribosome pauses rather than higher translation rates. To mitigate the influence of ribosome pauses on the assessment of differential translation, coordinates with the highest peaks of density could be excluded from the analysis as has been done in Andreev et al 2015. Yet another alternative would be to do a bootstrap sampling of densities from random CDS coordinates. Inconsistencies in differential gene expression analysis revealed by such a bootstrapping procedure would indicate a potential problem associated with ribosome pausing.

Often attempts are made to measure differential translation even when RNA levels do change simply by dividing the number of ribosome footprints aligning to an ORF by the number of RNA-seq reads. Such a procedure has several problems. First, it results in ratios that, unlike countable data (footprints and RNA-seq reads), do not carry information on the statistical significance, e.g. the ratio of 2/4 equals the ratio 200/400. Second, the best fit for the distribution of such ratio values is believed to follow the Cauchy distribution that is hard to model since both its mean and variance are undefinable. Finally, Ola Larsson and colleagues pointed out that spurious correlation between such ratios and their components (e.g. RNA levels) is necessitated mathematically (Larsson et al., 2010).

In principle, differential translation could be defined as a miscorrelation between the RNA-seq and ribosome profiling signal and it can be detected with the tools designed for RNA-seq analysis such as DESeq2 (Love et al., 2014) and EdgeR (Robinson et al., 2010) . Nonetheless, several standalone tools designed specifically for the characterization of differential translation efficiency from ribosome profiling data have been developed recently. Examples include babel (Olshen et al., 2013) , RiboDiff (Zhong et al., 2017), Riborex (Li et al., 2017), Xtail (Xiao et al., 2016), RIVET (Ernlund et al., 2018) and Anota2Seq (Oertlin et al., 2019), see Table 4. Online databases such as Trips-Viz (Kiniry, O'Connor, et al., 2018) and TranslatomeDB (Liu et al., 2018) also provide functionalities for differential gene expression characterization with the former applying a simple Z-score transformation for this purpose (Andreev et al., 2015; Quackenbush, 2002). As the statistical frameworks of these tools differ, not surprisingly, the sets and the number of genes predicted by them as differentially regulated differ. The field is seemingly in need of objective and independent benchmarking. It is important to note that irrespective of the specific approaches used for the assessment of differential translation, the differences are relative and not absolute.

Measurements of absolute changes in differential translation are not possible without spike-in controls allowing for the normalisation of the number of reads relative to the number of cells. Although attempts to introduce spike-in controls in ribosome profiling experiments have been made, e.g. Ingolia et al 2014, Andreev et al 2015, Iwasaki et al 2016, Popa et al 2016, and Gorochowski et al 2019, their suitability have not yet been rigorously assessed.

Table 1.4: *Software for the analysis of differential translation.*

| **Name** | **URL** | **Ref.** |
|---|---|---|
| Orqas | http://www.cs.cmu.edu/~ckingsf/software/ribomap/ | (Reixachs-Solé et al., 2019) |
| Ribomap | https://github.com/lcalviell/SaTAnn | (Wang et al., 2016) |
| SaTann | https://github.com/comprna/ORQAS | (Calviello, Hirsekorn, et al., 2019) |
| RPiso | http://cosbi7.ee.ncku.edu.tw/RPiso/ | (Wu et al., 2021) |

Table 1.5: *Software for the analysis of specific isoforms.*

| **Name** | **URL** | **Ref.** |
|---|---|---|
| Anota2Seq | https://bioconductor.org/packages/release/bioc/html/anota2seq.html | (Oertlin et al., 2019) |
| Babel | https://cran.r-project.org/web/packages/babel/index.html | (Olshen et al., 2013) |
| Ribodiff | https://github.com/ratschlab/RiboDiff | (Zhang et al., 2017) |
| Riborex | https://github.com/smithlabcode/riborex | (Li et al., 2017) |
| Rivet | https://ruggleslab.github.io/rivet/ | (Ernlund et al., 2018) |
| Xtail | https://github.com/xryanglab/xtail | (Xiao et al., 2016) |

Besides difficulties in evaluating differential expression based on two countable signals, the task is exacerbated by the existence and translation of multiple RNA isoforms due to alternative splicing and transcription initiation in complex eukaryotes, such as mammals (Blencowe, 2006). By mapping ribosome footprints across exon-exon junctions of alternatively spliced isoforms it has been shown that alternative isoforms could indeed be simultaneously translated (Weatheritt et al., 2016). However, when more than one RNA isoform is translated in the same sample, it is extremely difficult to compare their relative translation. Even when a certain cell type expresses only one predominant isoform, it is not apparent how to choose the one that will be used as a reference. In practice several heuristics are commonly applied to deal with this problem, each of which could lead to specific artefacts. One method is to use the longest isoform or the isoform with the longest annotated coding region. The rationale is that even if such an isoform differs from what is present in the cell, reads derived from the shorter isoform would align to the longer one allowing for measurement of expression differences. However, this can be problematic in cases where the shorter isoforms have coding exons that are missing in longer isoforms. This problem could be solved with creating a “union” of all transcripts by collapsing the genomic co-ordinates of all possible exons. While this is a sensible approach for the analysis of differential gene expression at the “gene level”, it may not be appropriate for the analysis of translated features within mRNA, e.g. uORFs, because such a union may disrupt such uORFs. When it is necessary to choose only a single transcript, so called “principal isoforms” could be used which are curated in the APPRIS database (Rodriguez et al., 2018). However, 5’ leaders/3’trailers are not taken into account here, meaning multiple isoforms that differ only in their noncoding regions would all be annotated as the principal isoform. In an attempt to move away from these heuristic approaches and their shortcomings, software has been developed which takes Ribo-Seq data into account to do transcript isoform level quantification, i.e. Ribomap (Wang, McManus, & Kingsford, 2016), ORQAS (Reixachs-Solé et al., 2019) and SaTann (Calviello, Hirsekorn, et al., 2019), see Table 5. However, these tools assign footprints to different isoforms under the premise that their protein synthesis input is directly proportional to their RNA levels, i.e. they are translated with the same efficiency. This, however, may not always be the case, especially when different start codons are used in different isoforms. The information that can be used within the Ribo-Seq data itself is reads that uniquely align to a specific isoform (unique exons and exon-exon junctions), however, because this is often within short regions of mRNAs, the number of footprints mapped to them could be sensitive to differences in ribosome dwell times at these locations. On top of that Ribo-Seq data could not be used to discriminate between alternative isoforms that differ in exons that are not translated.

## 1.7 Pause detection

Elongation rates varies as the ribosome traverses an mRNA and ribosomes could pause or stall at certain locations. Ribosome stalling can be caused by factors such as the secondary structure (Pop et al., 2014; Somogyi et al., 1993; Tholstrup et al., 2012), the interaction of the nascent peptide with the ribosome peptide channel (Becker et al., 2013; Tenson et al., 2002) and certain combinations of codons (Woolstenhulme et al., 2015). Pause sites have been shown to play important roles in translation in areas such as protein folding (Fluman et al., 2014; Tsai et al., 2008), and regulation of protein synthesis (Ivanov et al., 2018; Kurian et al., 2011; Yordanova et al., 2018). Pause sites are reflected in the Ribo-Seq data by high peaks relative to the surrounding region (Figure 1.1b). Like with translated ORF detection, pauses in Ribo-Seq data can be identified with manual visual inspection of ribosome footprint density profiles of individual mRNAs, but genome or transcriptome scale detection of pauses requires dedicated software.

PausePred (Kumari et al., 2018) is one such tool, available in both browser based and standalone versions that allows users to upload Ribo-Seq data and an optional annotation file. It then uses a sliding window approach to search for regions with high Ribo-Seq peaks relative to the background density. An accompanying tool, Rfeet then allows visualisation of the Ribo-Seq and (optionally) corresponding RNA-Seq data. Taking RNA-Seq into consideration is an important step when detecting pauses in Ribo-Seq data since it controls for the peaks caused by alignment artefacts. For example, when ambiguous mapping is allowed, a short region in a lowly expressed gene that shares sequence similarity with a highly expressed gene will appear as a peak in Ribo-Seq data. Similarly, if allowing only unambiguous alignments a short unique sequence surrounded by non-unique sequence will appear as a peak. Finally, a region with low sequence complexity that has many reads mapped just by chance can also appear as a peak in Ribo-Seq data. In all of these cases RNA-Seq data will also exhibit a pause at the same location, but not so in the case of a genuine ribosomal pause.

## 1.8 Prediction of footprint density

Ribo-Seq profiles are noticeably non-uniform, arising in part from differences in ribosome decoding rates in addition to the presence of sequencing biases occurring due to substrate sequence specificity of the enzymes used in generation and sequencing of cDNA libraries. Global assessment of footprint density allows for the magnitude of these biases to be estimated. A number of tools (see Table 6) have been developed to assess footprint density, including RUST (Ribo-Seq Unit Step Transformation) which allows for the measurement of how much various sequence features consistently influence the density of footprints at specific positions relative to the decoding center of the ribosome, .i.e. RUST would not detect features that only influence a small subset of unique locations (O'Connor et al., 2016). Metafootprint plots generated with RUST, (Figure 1.4), visualize these dependencies, Figure 4a is an example of a dataset with low sequencing bias from (Eichhorn et al., 2014) and

Figure 1.4b is an example of a dataset with high sequencing bias from (Reid et al., 2017). It is expected that the influence of the sequence at the decoding centre (i.e. A- and P-sites) should exceed that at the regions corresponding to read ends due to sequencing biases. Using the parameters of theses dependencies RUST can be used to predict ribosome profiling densities at the sequences for which no data exist with high accuracy.

Riboshape (Liu et al., 2016) is another tool which aims to understand the sequence features responsible for Ribo-Seqs non-uniformity. It does this using kernel smoothing to predict sequence features and then predicts the "shape" of ribosome profiles. The authors find that footprint density in *Saccharomyces cerevisiae* can be predicted with high accuracy. More recently developed tools utilise the power of deep learning to predict footprint density, such as ROSE (RibosOme Stalling Estimator) (Zhang et al., 2017) which is trained on transcripts with high ribosome profiling density to predict locations of ribosome pauses on transcripts with little to no signal. Finally iXnos (Tunney et al., 2018), which also uses neural networks, also aims to predict footprint densities. The authors compared the performance of iXnos to RUST and Riboshape and have shown that it outperforms them both on a single test dataset. They also demonstrated the utility of iXnos for optimizing the coding sequence to increase the translation efficiency.

Table 1.6: *Software for the analysis of local footprint densities.*

| **Name** | **Notes** | **URL** | **Ref.** |
|---|---|---|---|
| iXnos | Neural network based model of local densities. Can be used to predict local densities and for sequence optimization for increased expression | https://github.com/lareaulab/iXnos | (Tunney et al., 2018) |
| Pausepred | Detection of local peaks | https://pausepred.ucc.ie/ | (Kumari et al., 2018) |
| Riboshape | A kernel-smoothing model enabling prediction of local densities. | https://sourceforge.net/projects/riboshape/ | (Liu et al., 2016) |
| Rose | An approach for predicting ribosome stalling sites using deep convolutional network. | https://github.com/mlcb-thu/rose | (Zhang et al., 2017) |
| RUST | Unit-step based normalization of footprint densities for the analysis of sequence features effecting footprint densities, can be used to predict local densities. | https://lapti.ucc.ie/rust/ | (O'Connor et al., 2016) |

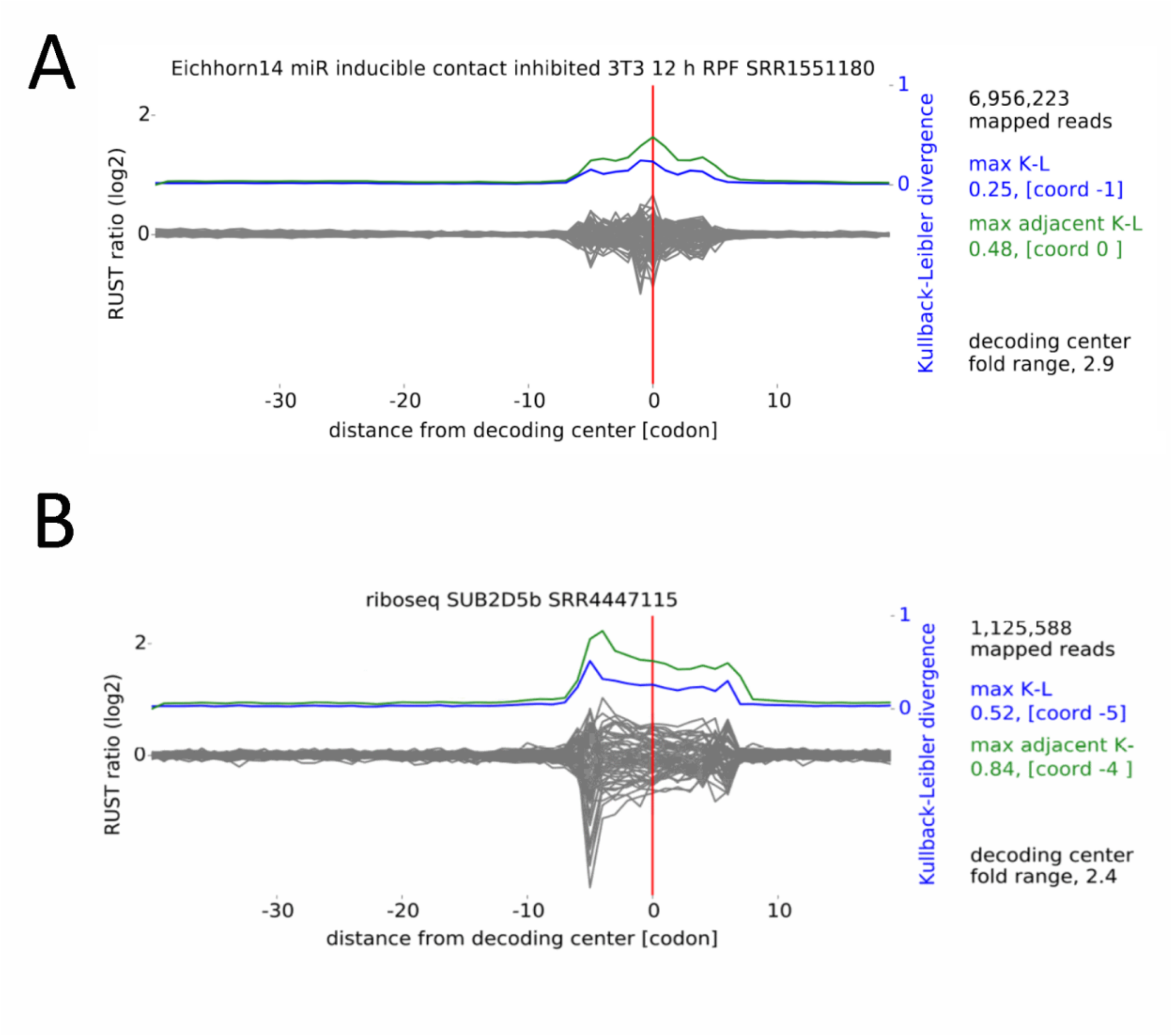

Figure 1.4: *RUST metafootprint profiles that can be used for the assessment of sequencing biases that are manifested by high relative entropy (measured as Kullback-Leibler divergence) at the ends of footprints. The decoding center of the ribosome (A-site) is denoted by the vertical red line. The blue line represents Kullback-Leibler divergence at an individual codon level. The green line represents Kullback-Leibler divergence for adjacent codons. In the absence of sequencing biases the Kullback-Leibler divergence is expected to be the highest at the decoding center. (a) A dataset with low sequencing bias. (b) A dataset with high sequencing bias at the 5' ends of footprints. For the data sources see the text.*

## 1.9 Pipelines, libraries, environments and data resources

In the absence of dedicated software, the early ribosome profiling analysis was carried out with tools developed for other high throughput sequencing applications and with *ad hoc* computer scripts. Over the past decade a number of different pipelines, computational environments and data resources have been developed. Researchers now have a considerable choice of existing freely available software to suit their needs, platform preferences and style. There is a large number of pipelines written in different languages for processing raw ribosome profiling data (see Table 1), e. g. the ruby based pipeline RiboPip (Stefan, 2016) and the R package systemPipeR (Backman et al., 2016) that provide full workflows for Ribo-Seq and RNA-Seq data analysis as well as other techniques such as CHIP-Seq in the latter package. Since raw data processing is not specific to ribosome profiling (outside of the above considerations) many packages and pipelines developed specifically for ribosome profiling take processed aligned reads in BAM file as input and provide only additional functionality. An example is a rich and extensible python library Plastid (Dunn et al., 2016). Likewise R packages Riboprofiling (Popa et al., 2016) and RiboseqR (Chung et al., 2015) also take alignment files as input and enable multifunctional downstream analysis.

The above software packages are operational through a command line and expect a certain familiarity with the Linux operating system. Moreover, setting up such software may require a certain effort and additional expertise for installing the software to a specific environment. The required skills and available time are often understandably lacking among wet lab researchers. The Galaxy Project (Afgan et al., 2018) offers a solution to this problem by providing a graphical web-based interface for data analysis, where workflows can be saved and rerun making the analysis reproducible. Software packages that do not have their own graphical interface could easily be integrated into Galaxy. Numerous specialized Galaxy servers have been created that provide the tools needed for a specific type of data. RiboGalaxy (Michel et al., 2016) is such an instance of Galaxy that provides several pipelines for the analysis of ribosome profiling data. RiboGalaxy is a part of the RiboSeq.Org collection (Figure 1.5).

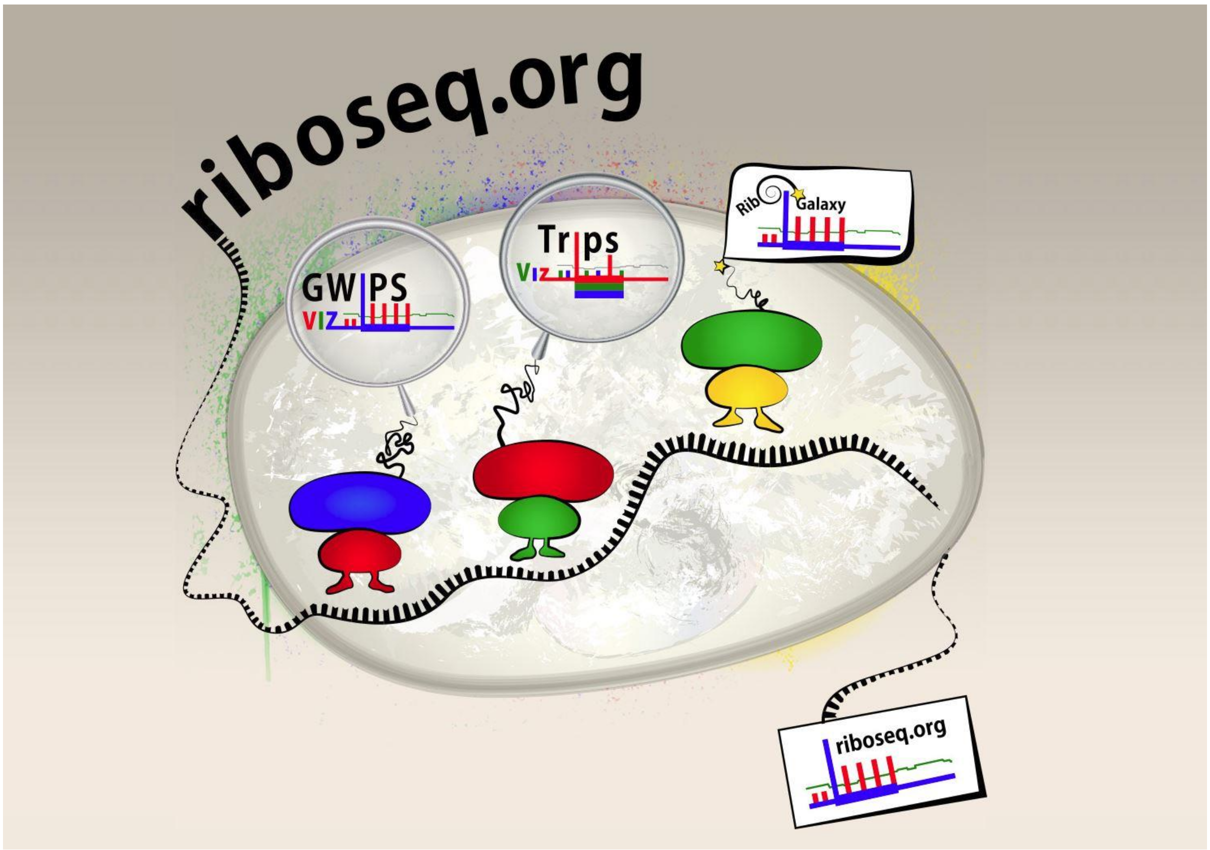

Figure 1.5: *The RiboSeq.Org web portal serves as an entry point to GWIPS-viz, Trips-Viz and RiboGalaxy. GWIPS-viz provides visualizations of publicly available ribosome footprints mapped to several genomes. Trips-Viz offers rich functionality for the analysis of public and user generated data aligned to transcriptomes. RiboGalaxy provides cross-platform graphical interface for the tools initially written as command line software.*

Like other sequencing data, ribosome profiling data can be found in public databases and many journals make deposition of the data to public archives a prerequisite for publication. The availability of the raw data, however, does not mean that these data can be easily utilised. Processing the raw data requires software, computational power, time and most importantly a certain level of familiarity with ribosome profiling data and technical issues described in this review. To democratize the data and to make it available to a large biomedical community that could benefit from it, it is important to provide access not only to raw data, but also to processed alignments. The first such database was GWIPS-Viz (for Genome Wide Information on Protein Synthesis Visualized) (Michel, Fox, et al., 2014). It provides genomic alignments of uniformly processed ribosome footprints and corresponding RNA-seq fragments. The alignments can be visualized either individually for specific datasets or as aggregates. To date GWIPS-Viz hosts Ribo-Seq data from 23 organisms (Michel et al., 2018). See (Kiniry, Michel, et al., 2018; Michel et al., 2015) for tutorials on how to use GWIPS-viz. Another large database of processed and aligned ribosome profiling data is RPFdb (Wang et al., 2018; Xie et al., 2016). While both GWIPS-viz and RPFdb are databases of genomic alignments of ribosome footprints, their functionality is markedly different, and their abilities overlap minimally. For example, RPFdb provides rich information on specific datasets such as raw counts and RPKM values for specific loci. Other databases such as the newly developed resource, Trips-Viz (for Transcriptome Information on Protein Synthesis Visualized) (Kiniry, O'Connor, et al., 2018) align data to a transcriptome as this has the advantage of eliminating the problem of mapping across exon-exon junctions. Trips-Viz is a web-based collaborative interactive environment for graphical computational analysis of publicly available Ribo-Seq data (although user generated data can also be uploaded). Several other databases have been developed that provide information derived from ribosome profiling analysis, see Table 1.2.

## 1.10 Conclusion

Over the past decade, many software modules, pipelines, visualization tools and data resources have been developed for Ribo-Seq analysis. As outlined in this review, more than one solution is now available for many tasks, from raw data processing to high-end applications such as the detection of translated ORFs. Subsequently researchers can choose tools to fit their specific computational backgrounds and styles. Nonetheless, despite the ample availability of resources, the field is far from saturation. We predict that it will continue to develop, perhaps, at an accelerated pace due to the following reasons.

The ribosome profiling protocol itself continues to develop. Specific modifications of experimental procedures sometimes require development of new tools. Even more importantly, a number of issues in ribosome profiling data analysis remain unsolved, e.g. differential expression analysis of allelic variants. While parallel approaches exist for the same tasks (e.g. translated ORF detection, differential gene expression), the results obtained with these approaches often poorly converge. This is largely due to the lack of gold standards and reliable criteria for evaluating the performance of these tools. Development of benchmarking approaches is expected to lead to improvement of these tools by providing the means for their comparison and optimization.

Recent developments in the characterization of mRNA translation, largely fuelled by ribosome profiling, further revealed the complexity of the translational landscapes of individual mRNAs, especially of high eukaryotes and specifically human mRNAs. Translation initiation could take place on many codons in the same mRNA (Fritsch et al., 2012; Lee et al., 2012) , leading to the production of proteoforms with different N-termini (Ivanov et al., 2011; Menschaert et al., 2013). At the same time ribosomes reading through the stop codons lead to the generation of proteoforms with different C-termini (Jungreis et al., 2011; Loughran et al., 2018; Rajput et al., 2019; Schueren et al., 2014). On top of that a large proportion of mRNAs contain short translated ORFs (Andreev et al., 2015; Ji et al., 2015; Johnstone et al., 2016),  some of which encoding functional proteins as in the human *MIEF1* mRNA (Andreev et al., 2015; Brown et al., 2017; Delcourt et al., 2018).  The currently used data structures for representation of RNA transcripts based on a single reference transcript with a single CDS are not suited for the representation of this complexity (Brunet, Levesque, et al., 2018). Thus, we envision the development of new, more adequate, data structures. The computational tools will need to adapt subsequently to these data structures.

Ribosome profiling allows for the quantitative assessment of only a single aspect of cellular activity, translation of its mRNAs. Often taking advantage of these data requires integration with other types of data (transcription initiation sites mapping, epitranscriptomics, mass spectrometry, etc.), and hence the tools for ribosome profiling data analysis need to provide such functionality either directly or through interoperability with the computational tools developed for the analysis of the data obtained with other techniques.

Yet another challenge is posed by the changes occurring in the analysis of big biodata in general. The volume of data in the Sequencing Data Archive doubles every 10-20 months (Langmead et al., 2018) which is faster than the growth of computational power. While ribosome profiling data currently represents only a microscopic fraction of these data, it is unlikely that the volume of ribosome profiling data will be growing at a slower pace. Thus, the computational efficiency of the algorithms will become critical. As the data volumes increase their physical transfer between servers is becoming increasingly less practical. This necessitates a paradigm shift from data-to-tools to the tools-to-data which requires the development of dedicated cloud infrastructure (Langmead et al., 2018). In the future tools will need to be adapted for these new environments.

# Chapter 2

## Trips-Viz: a transcriptome browser for exploring Ribo-Seq data.

*This chapter has been published in Nucleic Acids Research, Volume 47, D847-D852 (Kiniry, O'Connor, et al., 2018). For this work I wrote the majority of the code for Trips-Viz, POC worked on the z-score implementation in Trips-Viz. All authors worked on writing, reviewing and editing the manuscript.*

Ribosome profiling (Ribo-Seq) is a technique that allows for the isolation and sequencing of mRNA fragments protected from nuclease digestion by actively translating ribosomes. Mapping these ribosome footprints to a genome or transcriptome generates quantitative information on translated regions. To provide access to publicly available ribosome profiling data in the context of transcriptomes we developed Trips-Viz (**Tr**anscriptome-wide **i**nformation on **p**rotein **s**ynthesis- **Vi**suali**z**ed). Trips-Viz provides a large range of graphical tools for exploring global properties of translatomes and of individual transcripts. It enables analysis of aligned footprints to evaluate datasets quality, differential gene expression detection, visual identification of upstream ORFs and alternative proteoforms. Trips-Viz is available at https://trips.ucc.ie

## 2.1 Introduction

Ribosome profiling (Ingolia et al., 2009), also known as Ribo-Seq, is a technique that allows for large scale isolation of mRNA fragments that are being protected by actively translating ribosomes, see reviews (Andreev et al., 2017; Brar et al., 2015; Calviello et al., 2017; Ingolia, 2014; McGlincy et al., 2017; Michel et al., 2013; Stern-Ginossar et al., 2015). Sequencing these fragments, mapping them to a genome or transcriptome, and visualising these mappings can produce a global snapshot of which regions are being translated. There are a number of existing web based browsers which allow users to explore the alignments of publicly available ribosome profiling data. GWIPS-Viz (Michel, Fox, et al., 2014) which provides both ribosome profiling and mRNA-seq data aligned to the genome was the first such browser developed for this purpose. To date, GWIPS-Viz hosts data from 23 organisms (Michel et al., 2018). SmProt (Hao et al., 2018) is another web based tool that aligns ribosome profiling data to the genomes of 8 different organisms, combined with literature mining and mass spectrometry data it aims to find short translated ORFs (open reading frames) and allows users to explore each of these data types extensively.  RPF-db (Xie et al., 2016) also permits visualisation of ribosome profiling data aligned to 8 different organisms at a genomic level, as well as providing in depth information such as count tables, and meta-information such as the number of reads mapping to exonic/intronic/intergenic regions. Unlike these genome based tools, RiboViz (Carja et al., 2017) provides data aligned to the *Saccharomyces cerevisiae* transcriptome. It processes the data to analyse useful characteristics of the datasets, e.g. readlength distribution, triplet periodicity, as well as translation efficiencies. TranslatomeDb also aligns Ribo-Seq data to the transcriptomes of 13 different organisms, along with RNA-Seq and RNC-Seq data *(Liu et al., 2018).*

Mapping data to the transcriptome has certain advantages over mapping to the genome. Ribo-Seq reads are typically short (~30 nucleotides in length) and so the difficulty of mapping these short reads across splice junctions is relieved. The absence of long or numerous intronic regions makes the interpretation of the mapped reads easier from a user perspective when mapping to a transcriptome. However, it should be noted that aligning to the transcriptome is not inherently superior to genomic alignments, transcriptomic alignments for example are annotation dependent, meaning alignments would have to be re-done for each different version of the transcriptome. Transcriptome aligned data cannot be used for the analysis of translation outside of exons, e.g. translation of retained introns (Zafrir et al., 2016). As both methods have their advantages/disadvantages it would be best to make use of both transcriptomic and genomic alignments when analysing sequencing data.

Trips-Viz presents transcriptomic alignments of Ribo-Seq and mRNA-seq data. Currently the number of organisms available in Trips-Viz stands at 7 (*Homo sapiens, Rattus norvegicus, Saccharomyces cerevisiae, Mus musculus*, *Drosophila melanogaster*, *Escherichia coli*, and *Caenorhabditis elegans*). At the time of writing there are 1460 Ribo-Seq datasets and 335 mRNA-seq datasets available.

Trips-Viz utilizes a number of visualization solutions, not implemented by other tools. For instance, reads are coloured depending on matching subcodon position, to visualize triplet periodicity of Ribo-Seq data. Colour coding the reads can give a clear picture of which reading frames of a transcript are likely being translated, particularly if using an aggregate of data from many studies. This is particularly useful when multiple ORFs of the same transcript are being translated, e.g. CDS (Coding Sequence) and overlapping upstream ORFs (Michel et al., 2012).

Trips-Viz provides a versatile set of graphical analysis tools including the readlength distribution, triplet periodicity, metagene profiles and more. Trips-Viz also provides the ability to plot multiple datasets on the same graph for the same transcript. This allows for comparison of translated features across different samples, e.g. cell lines/tissues as well as across conditions and in response to drug treatments. Lastly, Trips-Viz allows the user to detect differentially expressed genes (at the level of RNA and protein synthesis).

## 2.2 Materials and methods

The Trips-Viz pipeline for processing Ribo-Seq data is as follows: publicly available ribosome profiling and corresponding RNA-seq datasets are downloaded from the gene expression omnibus https://www.ncbi.nlm.nih.gov/sra/ in SRA format. These are converted to FASTQ format and then the adapter sequence is clipped using cutadapt (Martin, 2011), reads below 25 nucleotides are removed. Bowtie (Langmead et al., 2009) is then used to remove any reads mapping to ribosomal RNA. Bowtie is again used to map the remaining reads to a reference transcriptome. Samtools (Li et al., 2009) is used to convert the resulting SAM file to BAM file format. Finally, the BAM file is parsed using a custom python script to pull out the necessary information for Trips-Viz, this includes determination of offsets for Ribo-Seq reads. This is a numerical value added to the position of the 5’ end of reads (or subtracted from the 3’ end) to approximate the A-site. This is done by creating a metagene profile, an aggregation of reads from all coding transcripts centred around annotated start codons. The distance in nucleotides between the highest peak upstream of the start codon (or downstream if determining a 3’ end offset) and the start codon itself (located at the P-site) is determined. This value is modified by adding 3 to set the 5’ end offset (or subtracting 3 to set the 3’ offset). Both 5’ end offsets and 3’ end offsets are determined separately for every read length. Offsets and other information extracted from the BAM file are stored in SQLite format.

The web framework for Trips-Viz is handled using the python package Flask (http://flask.pocoo.org/). All plots are generated using either mpld3 (http://mpld3.github.io) or bokeh (https://bokeh.pydata.org/en/latest) python packages. Currently we intend to include all publicly available Ribo-Seq data, however this may change as the number of ribosome profiling studies increases.

## 2.3 Discussion

The primary use of Trips-Viz is the interactive visualization of an aggregate of ribosome profiling data at subcodon resolution in the context of single transcripts, a feature not provided by other existing databases. To do this the user selects an organism and transcriptome assembly and then selects ***Single transcript plot***. Settings such as the gene of interest, minimum and maximum readlengths, ambiguous mapping filters and other settings can be changed at the top of the page. Ribo-Seq and mRNA-seq data files can be chosen at the centre of the page by selecting a sequence type, a study name and then clicking checkboxes next to file names. Clicking the *View Plot* button at the end of the page will produce a plot of the transcript in question. More detailed instructions on how to select data files and what each setting does can be found on the help pages or by clicking the link next to any of the settings labelled “*What's this*”.

There are three horizontal bars below the plot coloured in red, green and blue. These represent the three reading frames of the transcript, with short vertical white lines representing start codons and longer vertical grey lines representing stop codons. The main window shows densities of mapped footprints as line graphs of either red, green, or blue colours depending on the reading frame whose translation is the best supported by the reads based on their alignments relative to subcodon positions. The coloured boxes on the right of the graph represent the control panel with coloured buttons that allow the user to hide/display corresponding items in the main window. There are 4 icons below the plot, the first three when clicked, allows users to reset/move/zoom the view in the main window. The fourth icon allows the user to download the nucleotides sequence and read counts from the current transcript in csv format.

To demonstrate the utility of this plot an example is shown in Figure 2.1. Here a plot from the ***Single transcript plot*** page of Trips-Viz has been generated for the human *KIAA0100* gene using an aggregate of Ribosome Profiling datasets. The annotated coding region of this transcript starts at position 75 in the second frame (green). As can be seen in the figure most of the Ribo-Seq reads after position 75 are represented predominantly by green line graphs (up until the annotated stop codon at position 6780 where the read density decreases drastically) indicating translation in the second frame, as expected. Translation of a short upstream ORF at the coordinates 33-72 is also evident.  Within the CDS one notable exception to the predominantly green reads lies between positions 235 and 454 where the reads are predominantly blue. This corresponds to an ORF within the third line (blue) of the ORF architecture, which likely means this ORF is also translated.  Detection of such nested ORFs in particular highlights the currently unique utility of Trips-Viz that is enabled by differential read density colouring.

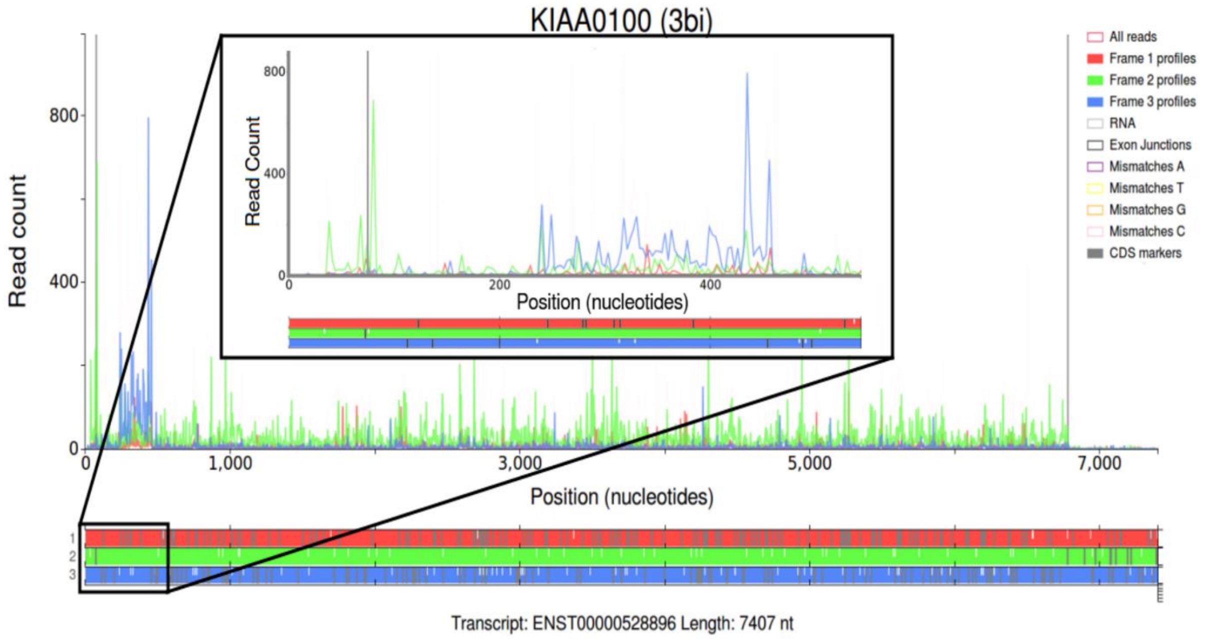

Figure 2.1: *Modified screenshots of the Trips-Viz single transcript plots for a Gencode transcript of the human KIAA0100 gene (large plot) and its 5' area (small plot). Ribo-Seq read densities are displayed in the main window, colour coded according to their mapping phase relative to the reading frame subcodon positions. Transcript coordinates are shown on the x-axis, while read counts are shown on the y-axis. The ORF architecture is shown below with three different reading frames differentially coloured, stop codons indicated as vertical grey dashes and AUGs as white dashes.*

Another useful feature of Trips-Viz is the ability to plot data obtained from multiple different samples on the same transcript simultaneously to allow comparative analysis. This can be achieved using the ***Single transcript comparison plot***. Here users can specify the transcript at the top of the page and choose whether to normalise the data over the number of mapped reads per sample, which is useful when comparing datasets with large differences in coverage. Users can set up groups of data using study names at the centre of the page. This is done by selecting a colour (by clicking on the coloured button), selecting a file and then clicking the *Add* button. The data between the groups are differentially coloured enabling comparison *via* visual inspection.

An example is shown in Figure *2.2* for the human *CSDE1* that illustrates how its translation is changed during Integrated Stress Response (ISR) using data from the Andreev *et al.* study (Andreev et al., 2015). For the samples treated with sodium arsenite (a trigger of ISR), Ribo-Seq and RNA-Seq read densities are displayed using line graphs of light red and dark red colours respectively. Read densities from untreated control samples are displayed in light green (Ribo-Seq) and dark green (RNA-Seq). It can be seen that both mRNA-Seq datasets have very similar densities, indicating that there is little or no RNA level changes in response to the arsenite treatment. In contrast, the Ribo-Seq density from arsenite treated cells is lower than that for the Ribo-Seq data obtained from the untreated cells, indicating that translation of this gene is reduced substantially during ISR in comparison with translation of other genes.

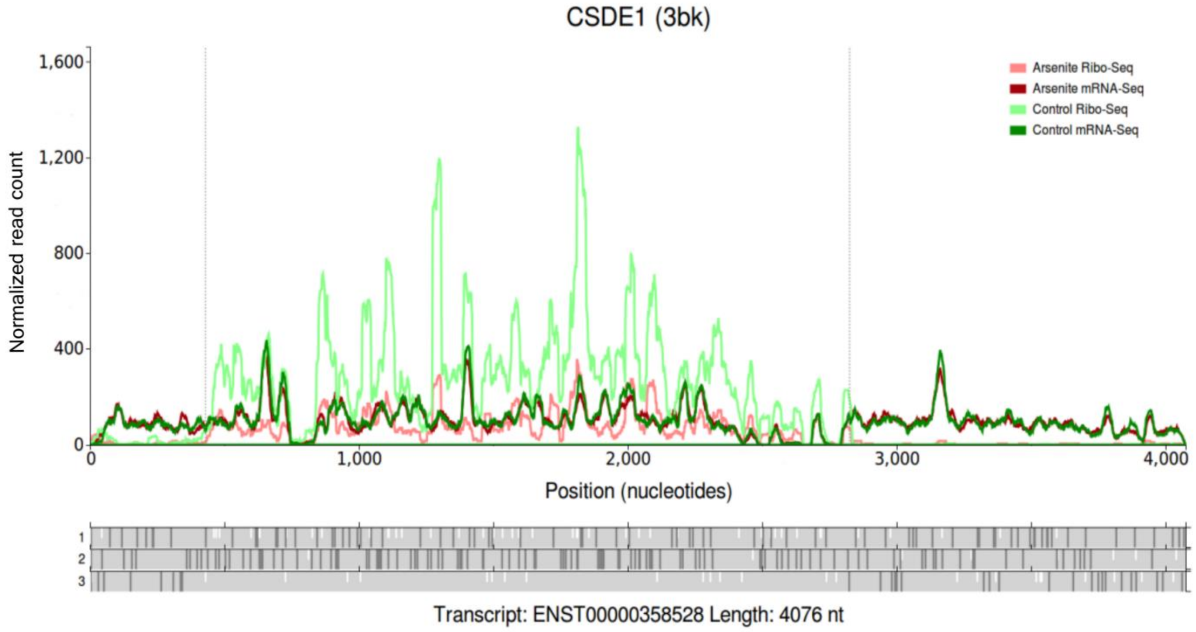

Figure 2.2: *A modified screenshot of a single transcript comparison plot for CSDE1 gene. The read densities from four datasets are shown as line graphs highlighted differentially as indicated by the legend in the top right corner. The other features are similar to Figure 2.1*

Unlike the two previous plot types the ***Meta-Information*** page gets its information from an entire dataset, aggregating information from multiple transcripts, for example, the triplet periodicity plot displays information from all annotated coding transcripts. This page allows the user to create a number of different plots which can be selected at the top left of the page. File selection is handled at the centre of the page in the same manner as the ***Single transcript plot*** page. In general, this page can be used to assess the quality of datasets as these plots provide general characteristics of the datasets that could reveal dataset defects. Examples are shown in Figure 2.3. A detailed description of each plot type can be found on the help pages, **https://trips.ucc.ie/help**.

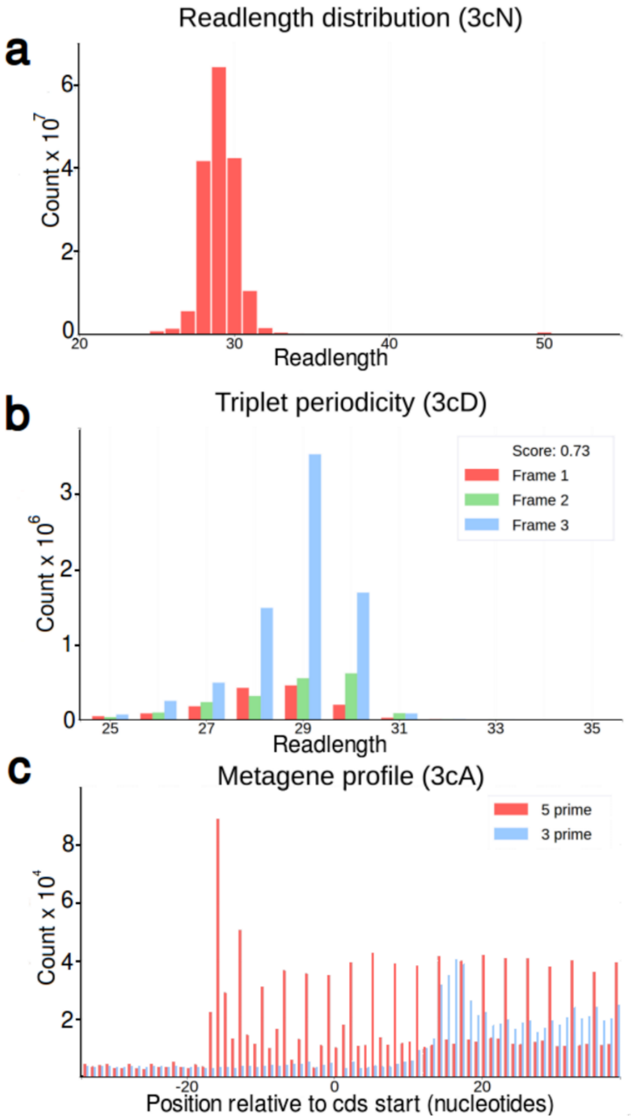

Figure 2.3: *Dataset characterizations.* ***a****. Distribution of read lengths from Matsuo et al. dataset (Matsuo et al., 2017).* ***b****. Triplet periodicity plot for a Ribo-Seq dataset from Loayza-Puch et al. (Loayza-Puch et al., 2016). Here each readlength is displayed using 3 bars depending on their phase to the first subcodon position of three different reading frames. Only reads aligned to annotated coding regions are used in this plot. The difference between bars indicates the strength of triplet periodicity. The datasets with stronger periodicity has a greater power for detecting translated reading frames as in the example shown in Figure 2.1.* ***c.*** *A metagene profile of a Ribo-Seq dataset from Neri et al. (Neri et al., 2017). Here the frequency of Ribo-Seq reads is shown relative to start codons (0 coordinate) across all protein coding transcripts and displayed either for reads 5' (red) or 3' (blue) ends. Since most ribosome footprints are expected to be found inside CDS regions, an increase in ribosome density is expected upstream of CDS. Metagene plots can be used for inferring an offsets between the decoding centre of the ribosome (A or P-sites) and the ends of ribosome footprints. The plot also indicates the strength and consistency of triplet periodicity.*

Lastly there is the ***Differential plot*** page, where users can find genes whose expression is significantly up/down-regulated relative to others. Users can organize the data into groups and compare relative RNA levels or protein synthesis levels between the groups and set minimum/maximum Z-scores at the top of the page. Up/down-regulated transcripts will then be detected using the Z-score transformation approach (Andreev et al., 2015). An example of the resulting plot can be seen in Figure 2.4. Here transcripts are represented as points on a scatter plot, with yellow lines specifying the upper and lower thresholds to indicate the z-score cut-off (as chosen by the user). Points above the upper threshold are coloured green (up-regulated) while points below the lower threshold are coloured red (down-regulated). Hovering the mouse cursor over a specific point will display the transcript ID and the number of reads mapped to it, while clicking on the point will open up a separate tab where the read densities for that gene will be displayed on the ***Single transcript comparison plot*** page.

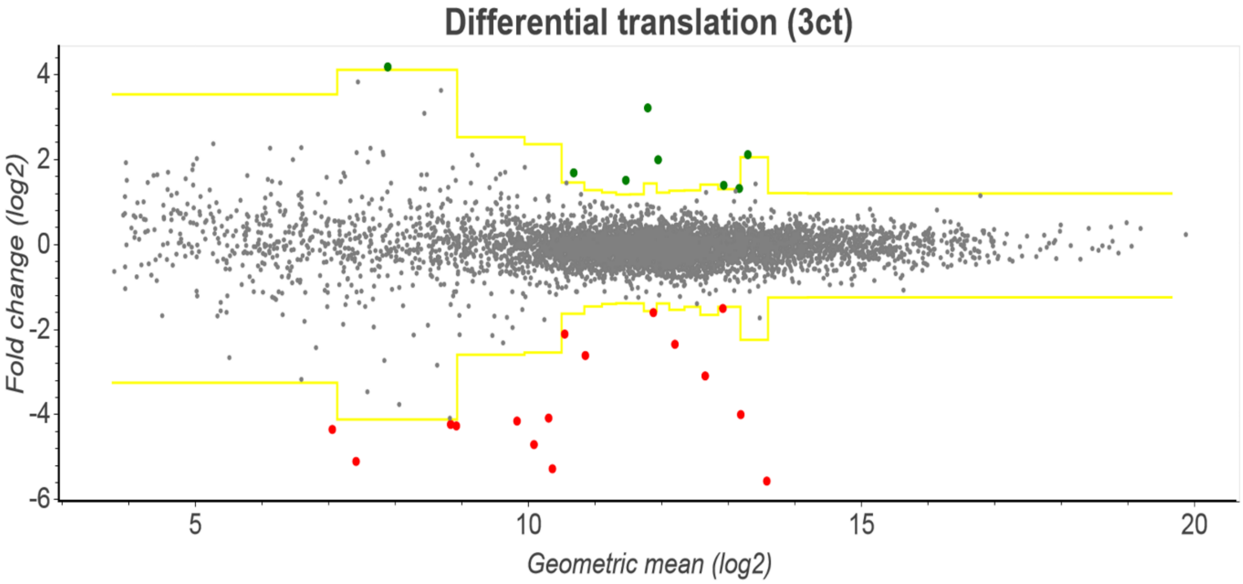

Figure 2.4: *A modified screenshot of Trips-Viz showing a plot from the Differential plot page for the datasets obtained in the Albert et al. dataset (Albert et al., 2014). Here fold change log ratios are shown on the y-axis while the geometric mean of the read counts in each condition is shown on the x-axis. Transcripts are grouped into bins of size 300 based on the geometric mean. Based on parameters of log ratios within each bin, a z-score is calculated for each transcript. The yellow lines on this graph represent the positive and negative z-score threshold (as chosen by the user), and transcripts that fall above/below that threshold are coloured green/red.*

In addition to data visualizations Trips-Viz provides a platform for collaborative research and data sharing. For every plot created on Trips-Viz a URL is created which contains information such as the files and settings used to create the plot. This URL can then be sent to another user, where Trips-Viz will use the information in the URL to recreate the plot in their browser. For convenience, rather than displaying the URL directly to the user, the URL is given a unique short code which is visible between parentheses in the title of every plot on Trips-Viz, including the plots presented in this manuscript. The URL can then be sent in the following form https//trips.ucc.ie/short/short_code. For example to recreate the plot shown in

Figure 2.1 users can follow the link https://trips.ucc.ie/short/3bi and explore the plot interactively in a browser. These links will last for the lifetime of Trips-Viz, with the exception of links associated with private data.

Private data can be uploaded by any user with an account on Trips-Viz, an account can be created using the *Sign up* link at the top of any page. Uploaded data must be in a specific format which can be created by running a python script and passing it a BAM file. Users can download this script from the Trips-Viz *downloads* page, a link to which is given at the top of every page and instructions on how to use it are included in the script itself. The *downloads* page also provides the relevant transcriptome fasta file and gtf file for each organism/assembly in Trips-Viz. Files can be uploaded using the *uploads* link at the top of every page. The user's data will be securely hidden from all other users by default but the uploader can share the data with other users of their choosing *via* the *uploads* page. Signing up also allows users to customize the graphic display of Trips-Viz, e.g. the background colour of plots. This can be accessed by visiting the *settings* link at the top of any page while signed in.

We plan to continually expand the number of organisms and Ribo-Seq/mRNA-Seq datasets available in Trips-Viz by including data as they become publicly available. However, it is conceivable that our computational capacities will not match the rapid pace of data growth. In this case we aim to develop a policy for data selection/prioritization based on data quality and their general scientific interest. We plan to streamline uploading of private data by providing a data processing workflow on Ribogalaxy (Michel et al., 2016). We also plan to generate a docker image of the site for users who may want to run their own instance of Trips-Viz. Also, we intend to explore the possibility of providing other types of publicly available sequencing data that are relevant to mRNA translation, e.g. epitranscriptomics data (Jantsch et al., 2018). We encourage users to contact us *via* the contact page https://trips.ucc.ie/contactus to provide feedback or suggestions, Trips-Viz related comments are also welcomed at the GWIPS-viz forum https://gwips.ucc.ie/Forum/ . The current version of Trips-Viz was optimized and tested with Chrome and Firefox browsers. Its full functionality with other Internet browsers is not guaranteed at present.

# Chapter 3

## Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data

*This chapter has been published in Nucleic Acids Research, Volume 49, Issue W1, Pages W662–W670 (Kiniry et al., 2021). For this work I wrote the majority of the code for Trips-Viz, CEJ worked on the pipeline for processing proteomics data. All authors worked on writing, reviewing and editing the manuscript.*

Trips-Viz (**https://trips.ucc.ie/**) is an interactive platform for the analysis and visualization of ribosome profiling (Ribo-Seq) and RNA sequencing (RNA-seq) data. This includes publicly available and user generated data, hence Trips-Viz can be classified as a database and as a server. As a database it provides access to many processed Ribo-Seq and RNA-seq data aligned to reference transcriptomes which has been expanded considerably since its inception. Here we focus on the server functionality of Trips-viz which also has been greatly improved. Trips-viz now enables visualisation of proteomics data from a large number of processed mass spectrometry datasets. It can be used to support translation inferred from Ribo-Seq data. Users are now able to upload a custom reference transcriptome as well as data types other than Ribo-Seq/RNA-Seq. Incorporating custom data has been streamlined with RiboGalaxy (**https://ribogalaxy.ucc.ie/**) integration. The other new functionality is the rapid detection of translated open reading frames (ORFs) through a simple easy to use interface. The analysis of differential expression has been also improved via integration of DESeq2 and Anota2seq in addition to a number of other improvements of existing Trips-viz features.

## 3.1 Introduction

Ribosome profiling (Ribo-Seq) is a technique that allows for large scale isolation of mRNA fragments which are being protected by actively translating ribosomes (Ingolia et al., 2009). These fragments can then be mapped to a genome or transcriptome and utilized in a number of different ways. This includes detection of novel translated open reading frames and pause sites, as well as identification of differentially translated genes, for reviews see (Andreev et al., 2017; Brar et al., 2015; Ingolia, 2014). To date there has been a number of different software packages created to explore each of these aspects of ribosome profiling (Kiniry et al., 2020). Many of these require some computational expertise and familiarity with command line usage. In addition, specific expertise and time are required to process and map the raw ribosome profiling reads. This too has been addressed by many packages (Dunn et al., 2016; H. Backman TW et al., 2016; Q. Liu et al., 2020; Michel et al., 2016; Ozadam et al., 2020) which aim to simplify the task of processing ribosome profiling data. Furthermore, many databases now exist which provide pre-processed publicly available ribosome profiling data (Brunet et al., 2019; Liu et al., 2018; Michel, Fox, et al., 2014; Olexiouk et al., 2018; Wang et al., 2018) , allowing users to carry out analysis either explicitly or implicitly through visualization of the data.

Among these is Trips-Viz, a transcriptome analysis platform with a focus on visualization and analysis of processed Ribo-Seq and RNA-Seq data. The triplet periodicity of ribosome profiling data allows for the detection of the translated reading frame (Michel et al., 2012). Trips-Viz takes advantage of this by colour coding Ribo-Seq reads according to the supported reading frame. This facilitates users to rapidly view Ribo-Seq profiles aggregated from numerous studies, providing the functionality to visually decipher not just the location but also the reading frame where translation is most likely occurring for a given mRNA transcript. Other capabilities include the option to directly compare multiple datasets on a single mRNA transcript, the functionality to carry out differential expression/translation analysis and calculate and visualize simple meta data statistics for individual datasets such as the distribution of read lengths, strength of triplet periodicity and metagene profiles. These statistics are useful for assessing data quality. Thus, Trips-Viz provides users with a large amount of relevant information which they can obtain very quickly and without the need for computational expertise and resources. Here we will discuss the major updates to Trips-Viz since its original publication (Kiniry, O'Connor, et al., 2018) focusing on its server functionality. For a full list of updates see **https://trips.ucc.ie/stats/**.

## 3.2 New and enhanced features

### 3.2.1 Improved ease of use

Since the launch of Trips-Viz users were able to process and upload their own files to be viewed privately and shared with collaborators. However, this required some familiarity with the command line and was a complex process. As one of the goals of Trips-Viz is to reduce users' computational workload, this was not an ideal solution. To address this, the relevant scripts were streamlined and incorporated into RiboGalaxy (Michel et al., 2016). RiboGalaxy is a GUI based platform made for processing Ribo-Seq and RNA-Seq data, based on the Galaxy platform (Afgan et al., 2018), designed to streamline and standardize analysis of biological data while making the process transparent and reproducible. Users can now easily carry out all the steps necessary to process a raw fastq file for uploading to Trips-Viz, all within RiboGalaxy. This includes upload of custom transcriptomes to Trips-Viz, a feature that was absent at the time of the Trips-Viz launch. Thus, Trips-Viz can now be used for data obtained from any species irrespective of whether a corresponding transcriptome is already available. The custom transcriptomes can also be processed using RiboGalaxy if users upload the relevant transcriptome fasta and GTF files.

While comprehensive help pages have been available from the beginning (**https://trips.ucc.ie/help/**), Trips-Viz is a GUI based tool which makes it difficult to easily explain the steps needed to carry out certain analysis when compared to a command line tool. This, coupled with the growing functionality and diversity of Trips-Viz visualizations makes using it more daunting for new users. To address this, videos have been embedded in the help pages, one for each plot type. Videos walk new users through the use of each plot, explaining the meanings of various settings and parameters and effects that they make on a specific visualization. This makes it easier for new users to quickly become familiar with the Trips-Viz interface and use it to its full potential. Users can now also download most plots on Trips-Viz in high resolution in .png format in addition to having more control over the size and colour of different plot elements on the settings page.

### 3.2.2 Mass spectrometry data

Trips-viz was originally designed solely for the analysis and visualization of Ribo-Seq and RNA-Seq data. Since then, we have expanded it to incorporate other data types, primarily mass spectrometry data. A popular application of Ribo-Seq data is to look for evidence of translation outside of regions annotated as protein-coding (Calviello et al., 2016; Ingolia et al., 2014). As mass spectrometry data also provide information on translation, it is reasonable to conclude that interrogating both types of data simultaneously can be greatly beneficial

(Brunet et al., 2019; Cao et al., 2020; Martinez et al., 2020; Verbruggen et al., 2019) . While there are numerous useful resources to explore publicly available mass spectrometry data (Desiere et al., 2006; Schmidt et al., 2018), many look only for support from existing annotated CDS's, diminishing their usefulness in terms of providing supporting information to Ribo-Seq findings. In Trips-Viz we do not limit the peptide search to CDS ORFs, opting instead to search all 3 reading frames across the entire transcript. This is done for all principal (Rodriguez et al., 2018) transcript isoforms in the transcriptome. This enables us to find proteomics support for translated regions regardless of location within the transcript and leverage the same graphs and colour scheme used to display Ribo-Seq data allowing users to easily see the frame and location of detected peptides.

To date there are 3152 processed mass spectrometry datasets available on Trips-Viz. The pipeline for Trips-Viz proteomics data integration involves searching for peptides in all 3 reading frames using MSFragger (Kong et al., 2017), then removing peptides with an FDR > 1% using Philosopher (da Veiga Leprevost et al., 2020). The output is then parsed and results are uploaded to Trips-Viz, where they are coloured according to the matching reading frame, in a similar manner to Ribo-Seq data. Their visualization can then be used to find novel translated ORFs, or to corroborate results observed from Ribo-seq data. See an example in Figure 3.1 where peptides from a uORF can be seen for the human gene *MIEF1,* which has previously been shown to be translated and was predicted to code for a functional protein (Andreev et al., 2015). Subsequently, its product was identified as a part of protein complex involved in assembly of mitochondrial ribosome (Brown et al., 2017) and further evidence supported its function in mitochondrial translation (Rathore et al., 2018), the proteomics data also suggest that the product encoded by *MIEF1* uORF is the main product of its mRNA(Andreev et al., 2015; Delcourt et al., 2018), while the synthesis of the MIEF1 protein is activated by stress conditions. Additionally, users can now also upload custom mappings of mass spectrometry data to Trips-Viz.

### 3.2.3 Detecting non-canonical ribo-seq signals

Detecting translated open reading frames (ORFs) using ribosome profiling data has been a subject of much interest in recent years. While many different programs now exist that can detect translated ORFs (Bazzini et al., 2014; Calviello et al., 2016; Clauwaert et al., 2019; Crappe et al., 2015; Erhard et al., 2018; Fields et al., 2015; Ndah et al., 2017; Reuter et al., 2016; Xiao et al., 2018), the majority of them require bioinformatic expertise as well as processed Ribo-Seq data, both of which may be expensive to acquire in terms of time and computational power. Trips-viz is now capable of automatically detecting Ribo-Seq signals outside annotated CDS regions in a simple but effective manner using previously processed Ribo-Seq data. This allows users to quickly and easily use an aggregate of data from multiple studies with good periodicity which can dramatically improve detection.

Trips-viz differs somewhat in its approach from most existing translated ORF detection approaches. It does not use machine learning methods as these rely on the availability of a “gold-standard” set of translated ORFs, which can be difficult to achieve even in well annotated organisms. Instead, at present, Trips-Viz first discards all read-lengths with weak triplet periodicity and then extracts 3 to 4 Ribo-Seq features (depending on the region of interest) from ORFs and ranks these features individually from strongest translational signal to weakest. These features include the increase of Ribo-Seq density at the start codon, the drop in Ribo-Seq density at the stop codon, the difference in in-frame and out-of-frame Ribo-Seq reads and the number of codons in the region of interest where the in-frame reads are higher than the out-of-frame reads. These individual ranks are then aggregated to determine a global rank for every ORF. It will then display a list of ORFs from strongest to weakest Ribo-Seq signal. This simplistic method does not allow for binary classification of ORFs as translated/untranslated as many other programs do. However, the goal of Trips-Viz differs in that it aims to allow users to rapidly find *individual* examples of high confidence non-canonical translation via manual inspection that warrant deeper investigation.

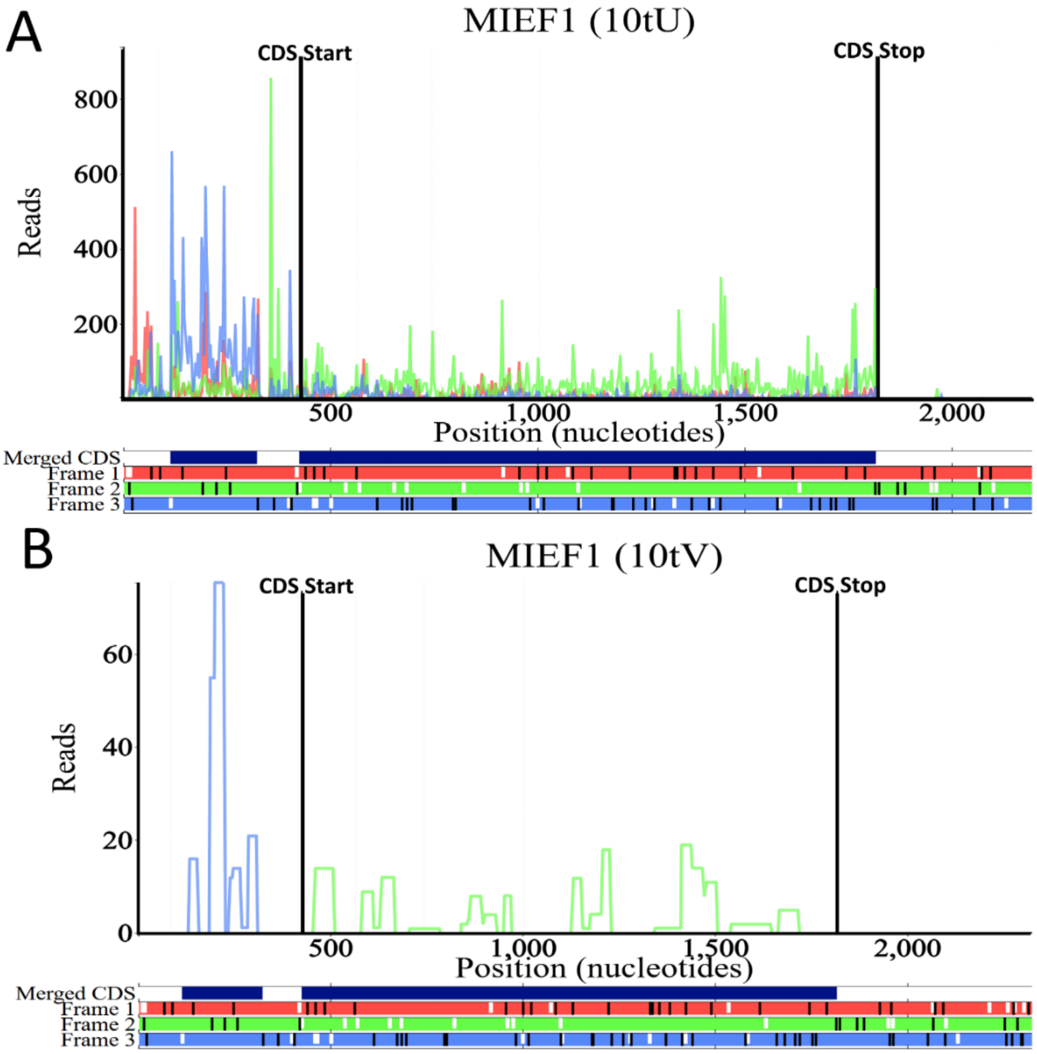

Figure 3.1: *The Trips-Viz single transcript plot for the human gene MIEF1 principal isoform (Transcript ENST00000325301) (zoomed in to show the 5' Leader/CDS).* ***(A)*** *An aggregate of Ribo-Seq reads from multiple studies. The triplet periodicity of the Ribo-Seq reads clearly shows a bias towards the second reading frame (green) which matches the reading frame of the annotated CDS. Ribo-Seq reads are mostly but not wholly confined to the CDS. Atypically there are many reads present in the 5' leader which are biased towards the third reading frame (blue) matching the location of an ORF in the third reading frame. This is corroborated by the proteomics data in panel* ***(B)****. Locations encoding peptides from an aggregate of mass spectrometry datasets are displayed. All peptides bar one in the 3' trailer (not shown) are found either within the CDS (frame 2) or in the third reading frame matching the position of the uORF in the 5' leader. The code in brackets in the title of the plot can be used to generate the profile in a browser, e.g following this link https://trips.ucc.ie/short/10tU will load the plot shown in panel A, for more information on short codes see the Trips-Viz help pages.*

To aid in this manual inspection, results are displayed in the form of a table showing the top 1000 ranked ORFs, with the option of downloading the entire table. Each ORF will have a link allowing the user to view the ORF in question in the corresponding transcript with the selected data, allowing users to rapidly visualize each ORF using only the datasets they selected, see Figure 3.2. Translation of ORFs that belong to noncoding RNAs can be detected in addition to ORFs from annotated coding transcripts which are broken down into the following categories depending on their location relative to the CDS: upstream ORFs, overlapping upstream ORFs, nested ORFs, downstream ORFs and n-terminal extensions.

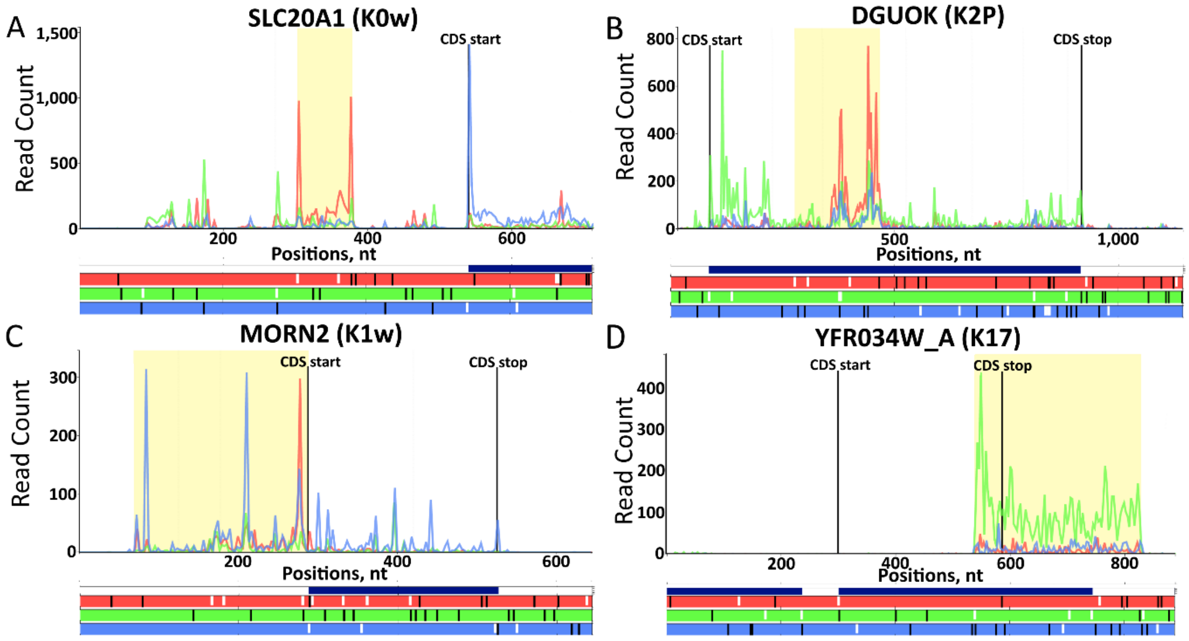

Figure 3.2: *Examples of highly ranked ORFs detected by Trips-Viz. At the bottom of each plot the open reading frame architecture is represented with three horizontal bars coloured red, green and blue to display each of the three reading frames. AUG codons are denoted by short white lines while stop codons are denoted by longer black lines. The CDS start and CDS stop positions are shown with the vertical black lines on the main plot along with counts of Ribo-Seq reads displayed in red, green and blue matching each of the three reading frames beneath. The "merged CDS" bar above reading frames bars displays a union of all annotated CDS regions in the corresponding locus. (**A**) An AUG-initiated uORF of the human gene SLC20A1 (Transcript ENST00000272542) in frame 1 (the annotated CDS is in frame 3). (**B**) A nested ORF in frame 1 of the human gene DGUOK (Transcript ENST00000264093) where the annotated CDS is in frame 2. (**C**) An N-terminal extension of the mouse gene Morn2 (Transcript ENSMUST00000061703). (**D**) An overlapping downstream ORF of the yeast gene YFR034W_A. Here the merged CDS bar extends into the 3' trailer of YFR034W_A indicating the presence of another gene at the same locus. In this case the gene in question is YFR035C which is transcribed from the opposite strand, so it cannot explain the Ribo-Seq reads in the 3' trailer of YFR034W_A. In addition, the reads extend beyond the merged CDS bar.*

### 3.2.4 Differential expression/translation

Since its launch, Trips-Viz provided a single option for carrying out differential expression analysis on principal transcript isoforms using the z-score transformation (Andreev et al., 2015). While this performs adequately, there are more accurate and powerful approaches for this purpose (Zhong et al., 2017). To this end two new options were incorporated into Trips-Viz, DESeq2 (Love et al., 2014) and anota2seq (Oertlin et al., 2019), which will allow users to quickly compare the results across the 3 methods. An example plot can be seen in Figure 3.3, showing the Ribo-Seq fold change versus the RNA-Seq fold change. It allows users to quickly see expression of which genes are affected at the RNA and/or translation levels. Similarly, to the z-score plot, users can click on any point in the plot to invoke a comparison plot where footprint densities are compared for two condition for the corresponding transcript. Users can also download the inputs and outputs of DESeq2 and anota2seq for further exploration. It is recommended that DESeq2 and anota2seq be used over the z-score method, however these require a minimum of 2 and 3 replicates respectively, thus the z-score transformation approach remains the only option for exploring datasets lacking replicates which could be useful during preliminary data generation and pilot experiments.

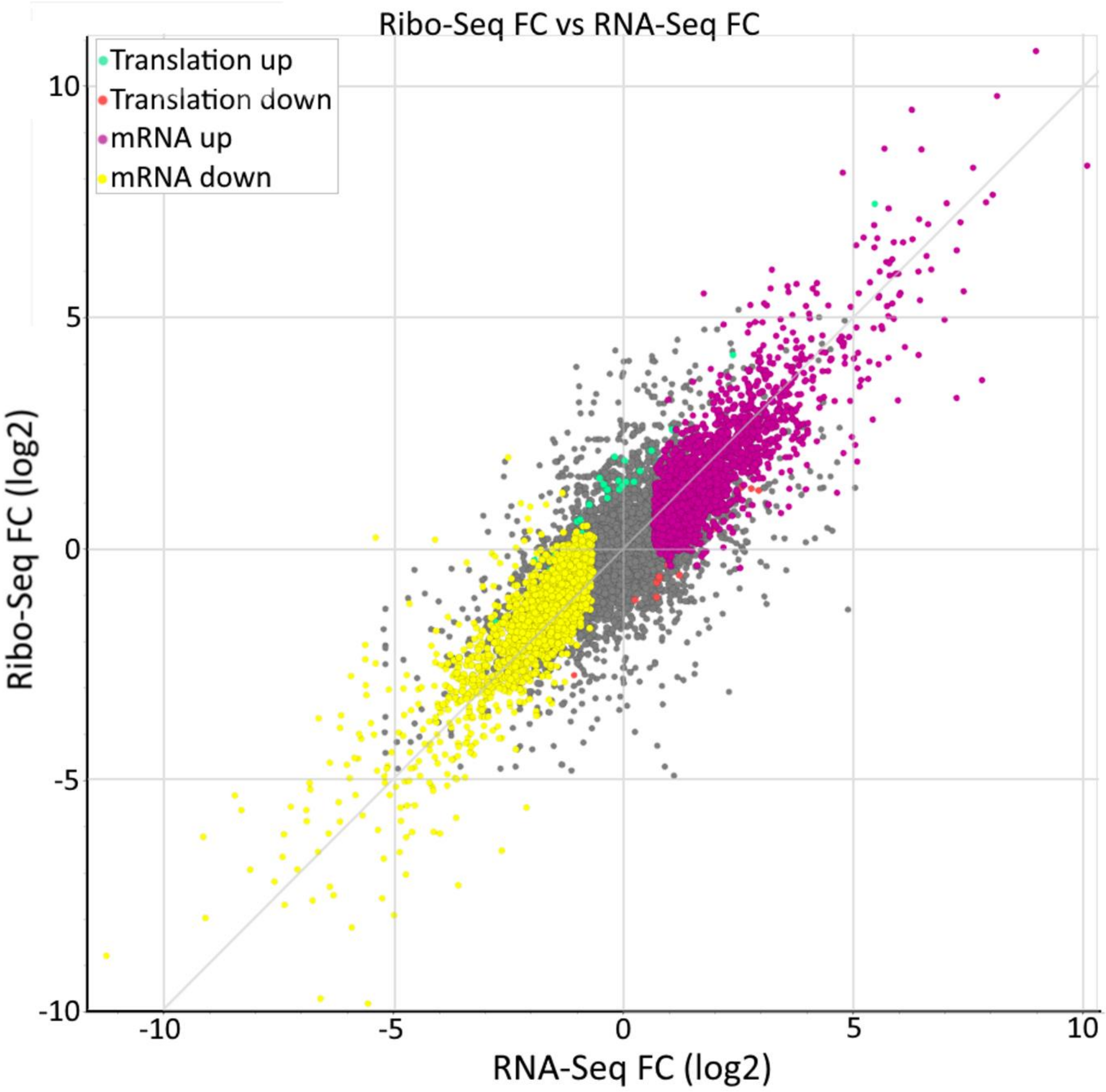

Figure 3.3: *An example of the differential gene expression analysis using DESeq2, using data from Iwasaki et al. (Iwasaki et al., 2016). Genes whose expression did not change significantly are coloured grey, translationally upregulated/downregulated genes are in green/red, while changes in mRNA levels are in purple/yellow. Hovering over any of the points on the plot triggers a pop-up window with information specific to the corresponding gene/transcript and fold changes. Clicking on any of the points invokes a separate tab showing the comparison plot for ribosome footprints mapped to the corresponding gene.*

### 3.2.5 Transcriptome metainformation

A new section has been added to Trips-Viz to address all queries not directly related to Ribo-seq/RNA-Seq or other data types. This can be used to address simple questions about a transcriptome such as how many genes/transcripts are annotated and how many are coding/non-coding, what is the codon usage in CDS regions or what is the difference in GC content in 5’ leaders (commonly known as 5’ UTR’s)  versus 3’ trailers (commonly known as 3’ UTR’s). It can also be used to retrieve nucleotide sequences of some or all transcripts, either in their entirety or for specific subsections (5’ leader, CDS, 3’ trailer).  However, most plots on this page can be generated using subsets of transcripts. This can be used to gain a deeper understanding of differential expression/translation results, for example, by comparing these features between groups of upregulated and downregulated genes. An example is presented in Figure 3.4.

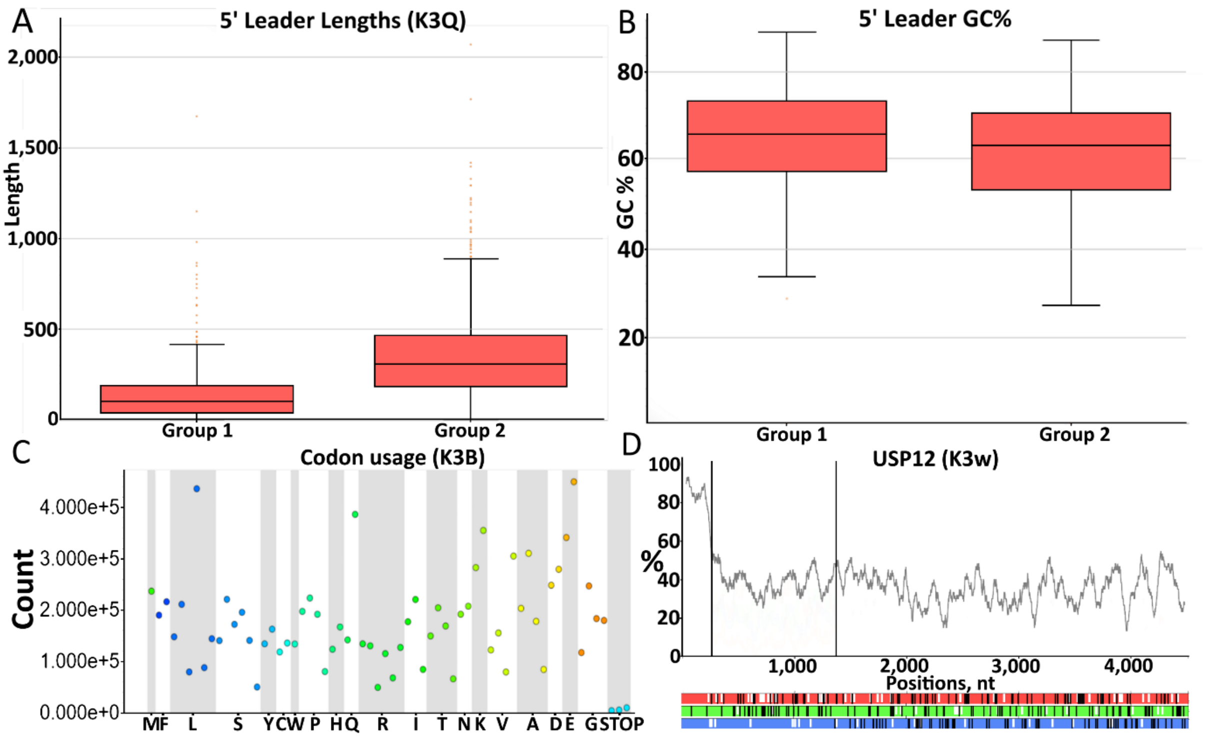

Figure 3.4: *Examples of the types of plots generated from the transcriptome info page.* ***(A)*** *A comparison of 5' leader lengths and (****B****) 5' leader GC% between upregulated/downregulated genes in response to RocA treatment (Iwasaki et al., 2016). (****C****) The codon usage occurrence within the CDS of all principal transcript isoforms in the human transcriptome using Gencode v25 (Frankish et al., 2019). (****D****) The GC content of the human gene USP12. The same visualization can be used for individual nucleotide frequencies within sliding windows or for minimum free energy of potential RNA secondary structures as calculated by ViennaRNA (Lorenz et al., 2011).*

## 3.3 Comparison with other tools

It has now been over a decade since the introduction of the ribosome profiling technique and in that time a plethora of different tools have been developed that cover almost every aspect of ribosome profiling data analysis (Kiniry et al., 2020). Carrying out a detailed analysis against all available tools would be difficult due to the sheer number of them. Instead, these have been broadly split into two categories. There now exists many offline tools such as

Plastid (Dunn et al., 2016), RiboProfiling (Popa et al., 2016), riboflow (Ozadam et al., 2020), and , ribotaper (Calviello et al., 2016) to name just a few, which are designed to be downloaded and installed locally for users to process and analyze their own data. These tools have considerable overlap with Trips-Viz in terms of the type of analysis that they provide but as these tools typically require some computational expertise the target audience differs from Trips-Viz which aims to provide a solution to those without such expertise. Instead, a more detailed comparison was made to other online databases which either provide pre-processed data or provide an easy way to process Ribo-Seq data which does not require computational expertise. These tools include RiboToolKit (Liu et al., 2020), SmProt (Hao et al., 2018), HRPDViewer (Wu et al., 2018), TranslatomeDB (Liu et al., 2018), RiboViz (Carja et al., 2017), RPFdb (Wang et al., 2018), OpenProt (Brunet et al., 2019), GWIPS-Viz (Michel et al., 2015), RiboGalaxy (Michel et al., 2016), and RiboStreamR (Perkins et al., 2019). The features of these tools are listed in Table 1.

Table 3.1: *Comparison table showing the presence (tick) or absence (x) of various features (rows) available in Trips-Viz and similar tools (columns)*

| | *Trips-Viz* | *RiboToolKit* | *SmProt* | *HRPDViewer* | *TranslatomeDB* | *RiboViz* | *RPFdb* | *OpenProt* | *GWIPS-VIZ* | *RiboGalaxy* | *RibostreamR* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Web-Upload | ✔ | ✔ | ✘ | ✘ | ✔ | ✘ | ✘ | ✔ | ✔ | ✔ | ✔ |
| Batch-Uploading | ✔ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✔ | ✘ | ✘ | ✔ |
| Local Install | ✔ | ✔ | ✘ | ✘ | ✘ | ✔ | ✘ | ✘ | ✘ | ✔ | ✘ |
| Pre-processed data | ✔ | ✘ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✔ | ✘ | ✘ |
| Contamination checking | ✔ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✔ | ✔ |
| Quality checking | ✔ | ✔ | ✘ | ✘ | ✘ | ✔ | ✔ | ✘ | ✔ | ✔ | ✔ |
| RPF visualisation | ✔ | ✔ | ✘ | ✔ | ✔ | ✔ | ✔ | ✘ | ✔ | ✔ | ✔ |
| RNA-Seq visualisation | ✔ | ✘ | ✘ | ✘ | ✔ | ✔ | ✘ | ✘ | ✔ | ✔ | ✔ |
| Mismatch detection | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ |
| Visualization of subcodon profiles | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✔ | ✔ |
| Nucleotide sequence retrieval | ✔ | ✘ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✔ | ✔ | ✘ |
| Codon Occupancy | ✘ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✔ | ✔ | ✘ |
| Pause detection | ✘ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ |
| Codon Frequency | ✔ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ |
| Meta-Codon plots | ✘ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ |
| mRNA expression | ✔ | ✔ | ✘ | ✘ | ✔ | ✔ | ✘ | ✘ | ✘ | ✔ | ✔ |
| RPF expression | ✔ | ✔ | ✘ | ✔ | ✔ | ✔ | ✔ | ✘ | ✘ | ✔ | ✔ |
| Translation efficiency analysis | ✔ | ✔ | ✘ | ✘ | ✔ | ✔ | ✘ | ✘ | ✘ | ✔ | ✔ |
| Differential translation analysis | ✔ | ✔ | ✘ | ✘ | ✔ | ✘ | ✘ | ✘ | ✘ | ✔ | ✔ |
| Translated ORF detection | ✔ | ✔ | ✔ | ✘ | ✘ | ✘ | ✔ | ✔ | ✘ | ✘ | ✘ |
| Proteomics Analysis | ✔ | ✘ | ✔ | ✘ | ✘ | ✘ | ✘ | ✔ | ✘ | ✘ | ✘ |
| GO/Pathway analysis | ✘ | ✔ | ✘ | ✘ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ |
| MetaGene Plots | ✔ | ✔ | ✘ | ✘ | ✘ | ✔ | ✔ | ✘ | ✘ | ✔ | ✔ |
| Reproducibility between replicates | ✔ | ✘ | ✘ | ✘ | ✔ | ✘ | ✘ | ✘ | ✘ | ✘ | ✘ |

While this table attempts to capture the main differences between Trips-Viz and similar tools it is difficult to simplify all differences into simple binary categories. To that end we discuss a specific example which shows various features of Trips-Viz which can be used in concert to investigate the translation of specific RNA transcripts and quickly make interesting biological observations.

The human gene *POLG* has been recently shown to encode an additional protein in an overlapping upstream open reading frame (ouORF) (Khan et al., 2020; Loughran et al., 2020). Visualizing the translation of *POLG* mRNA using currently available public Ribo-Seq data in Trips-Viz makes the translation of the ouORF clear due to a number of features (Figure 3.5). Most important is the ability to visualize subcodon profiles by colouring reads according to the reading frame in which they are found (as determined by the inferred A-site). This is what makes it clear that the read density in the first reading frame (red) is much higher within the ouORF which then decreases at the ouORF stop codon. The majority of tools used for visualizing Ribo-Seq data do not employ this technique making it much more difficult to visually identify dual coding regions. Trips-Viz also has the functionality to set a periodicity score, which filters out all reads with poor periodicity making the signal from the ouORF evident, saving users from having to identify and manually select studies with strong periodicity.

The ORF architecture beneath the subcodon profile in Figure 3.5 (horizontal red, green and blue bars) displays the positions of the AUG’s (short white lines) and stops (longer black lines). No AUG is visible in the first reading frame (red) that could act as a potential start for the ouORF. Trips-Viz, however, allows users to optionally enter any nucleotide sequence to be highlighted in the ORF architecture. In Figure 3.5 CUG codons are shown as short black lines in Frame 1 making it easier to see the exact position where the ouORF initiates. The merged CDS bar (dark blue bar just above the ORF architecture), shows all the regions of the transcript which overlap with other annotated CDS regions. As there is no dark blue bar in the non-overlapping region of the ouORF, it is possible to tell from this plot that the ouORF is not a part of the annotated (in the current annotation version) CDS in an alternative transcript without having to explore the exon architecture and annotation of the corresponding genomic locus.

The short code in brackets above the plot (14HD) can be used to recreate the plot in the browser using the same settings and files used to create the plot initially. Navigating to https://trips.ucc.ie /short/14HD in a browser will recreate the plot shown in Figure 3.5 and show which settings and files were used to generate the plot, as well as allowing users to make full use of the interactive features of the plot, like the ability to pan, zoom, turn on/off different plot elements, download the image as a high quality .png, and download the raw counts. Every plot created in Trips-Viz is linked to one of these short codes allowing them to easily be reproduced and shared. For example, navigating to https://trips.ucc.ie/short/14SJ will load a plot showing the same transcript isoform for *POLG* but with proteomics data, where sequences encoding peptides that match mass spectra are found. Their presence within the ouORF further strengthening the confidence that this ORF is translated and likely encodes a stable protein product. While performing a similar analysis on other platforms is certainly achievable, this example highlights the power and flexibility of Trips-Viz visualization.  The combination of a large number of publicly available data with a versatile set of computational tools and visualizations makes such analyses both quick and easy.

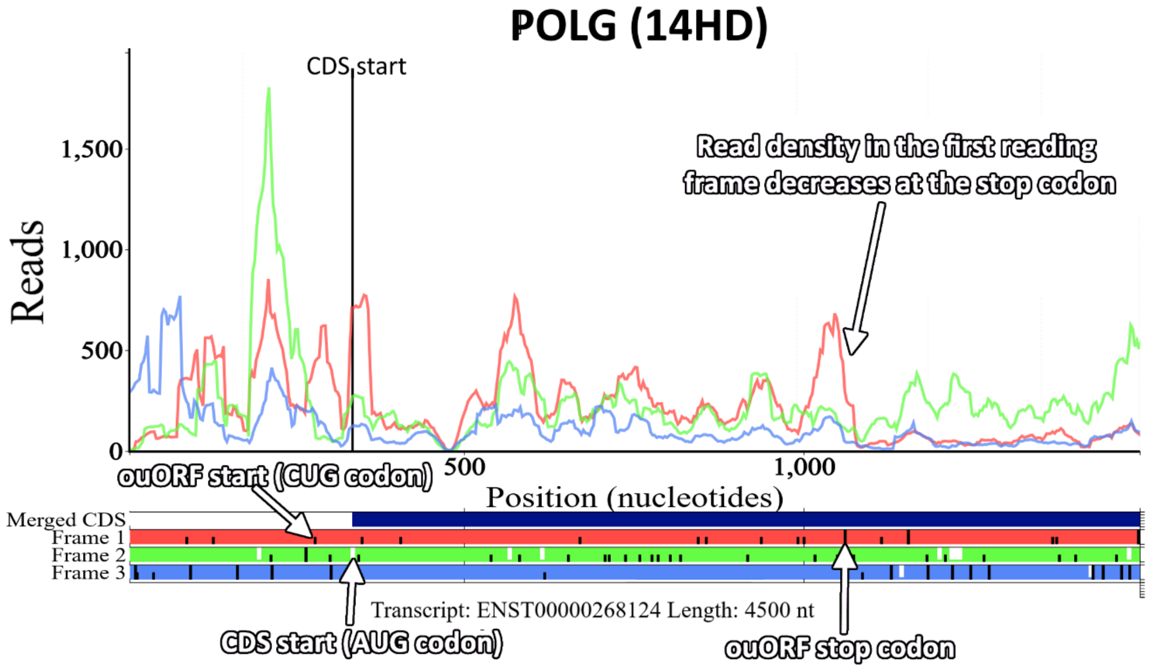

Figure 3.5: *The Trips-Viz single transcript plot for the human gene POLG (transcript ENST00000268124) using an aggregate of Ribo-Seq reads from multiple studies. The view has been zoomed in around the annotated CDS start. There is a Ribo-seq bias within the first reading frame (red) across the entirety of the ouORF supporting the translation of the correct frame. The ORF architecture beneath the plot (horizontal red, green and blue bars), shows the positions of AUG codons (short white lines), stop codons (longer black lines) and CUG codons (short black lines), making it easier to see the start and stop positions of the ouORF. The third CUG codon in the first reading frame is the translation initiation site of the ouORF (Khan et al., 2020; Loughran et al., 2020), which has been highlighted in the figure.*

## General impact and future perspectives

Visualisation is an important tool in the analysis of ribosome profiling data. It can be used to find novel translated ORFs as well as to identify both pause sites and translational recoding. When I began my research there existed several databases which were capable of visualising Ribo-Seq data. One such site, GWIPS-Viz (Michel et al., 2018), aggregates the signal from many different Ribo-Seq studies aligned to the genome. Aggregation of Ribo-Seq data can make translated ORFs on even lowly expressed genes easily visible/detectable. A screenshot showing a translated uORF as seen in GWIPS-Viz can be seen in Figure 4.1.

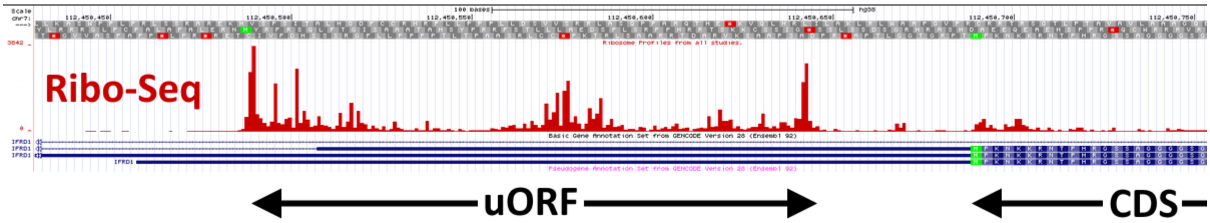
Scale
chr7:
112,450,450
112,450,500
100 bases
112,450,550
hg38
112,450,600
112,450,650
112,450,700
112,450,750
3842 _
Ribosome Profiles from all studies.
Ribo-Seq
0 _
Basic Gene Annotation Set from GENCODE Version 28 (Ensembl 92)
IFRD1
IFRD1
IFRD1
IFRD1
Pseudogene Annotation Set from GENCODE Version 28 (Ensembl 92)
uORF
CDS

Figure 4.1*: A screenshot from GWIPS-Viz showing the human gene IFRD1. The open reading frame architecture above the plot shows start codons in green and stop codons in red. The Ribo-Seq density is clearly higher and localized to within the ORF in the 5' leader.*

Aggregation of reads from multiple Ribo-Seq studies makes uORFs easily visible in cases like *IFRD1*. However in cases where translated ORFs overlap (Michel et al., 2012) this is more difficult. Visualising the Ribo-Seq reads for the human gene *KIAA0100* does not show anything particularly unusual (Figure 4.2), apart from a short translated ORF in the 5' Leader.

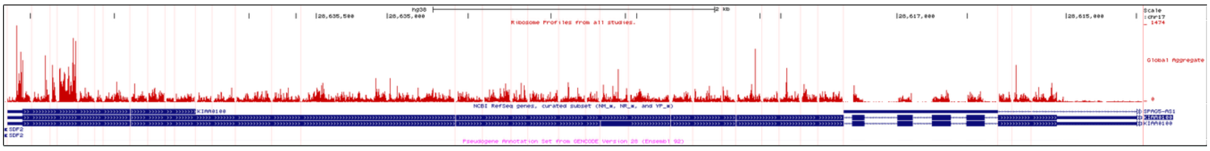
hg38
2 kb
Scale
chr17
28,635,500
28,635,000
28,617,000
28,615,000
Ribosome Profiles from all studies.
1474
Global Aggregate
0
NCBI RefSeq genes, curated subset (NM_*, NR_*, and YP_*)
KIAA0100
SPAG5-AS1
KIAA0100
KIAA0100
SDF2
SDF2
Pseudogene Annotation Set from GENCODE Version 28 (Ensembl 92)

Figure 4.2*: A screenshot from GWIPS-Viz showing aggregated Ribo-Seq data from the human gene KIAA0100. Apart from the increased Ribo-Seq density in the 5' leader nothing else seems out of the ordinary.*

The use of static offsets which GWIPS-Viz and several other resources use, also tends to destroy the periodicity signal when aggregating from multiple studies. Determining an optimal offset per read length and colouring the reads according to the reading frame that the A/P-site aligns to, as is done on Trips-Viz, makes the profile much clearer (Figure 4.3).

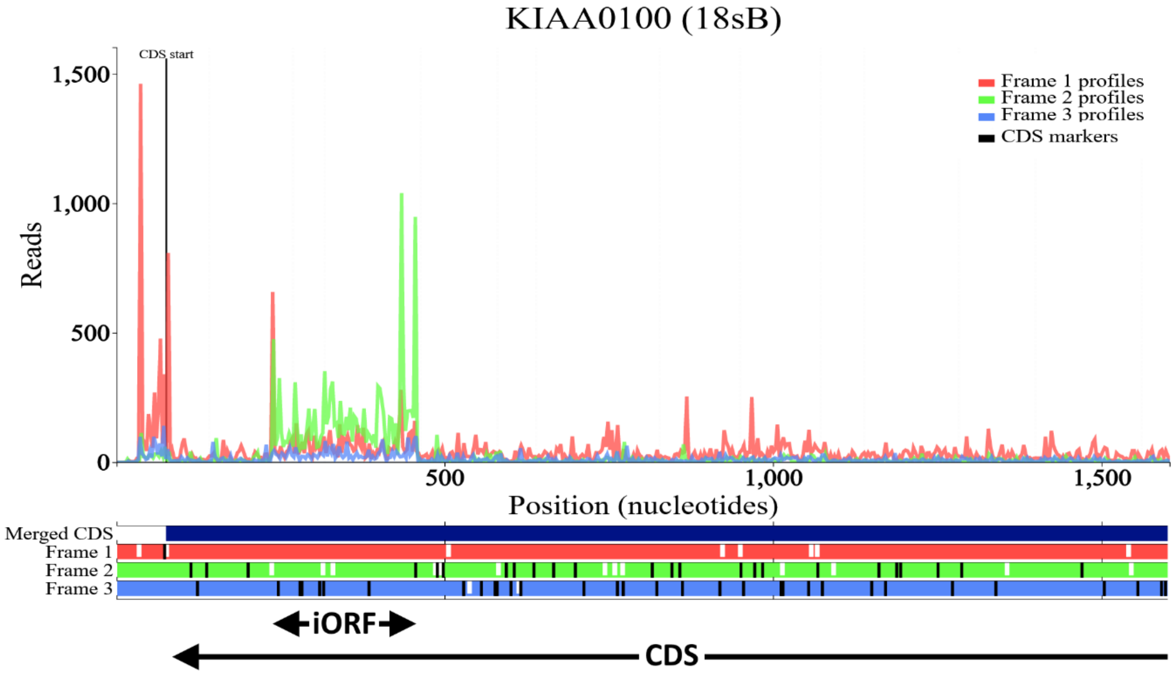

Figure 4.3*: Aggregated ribosome profiling data for the human gene KIAA0100. Reads are coloured red, green and blue corresponding to the open reading frame architecture shown beneath the plot. Most reads are red which corresponds to the annotated CDS (frame 1) as well as a short uORF which is also in frame 1, but the colouring makes it clear that there is also translation of a nested or internal ORF in frame 2 (green) which would be difficult to see otherwise.*

While these types of plots had been used for ribosome profiling data before, they were done for single studies/datasets. There was no publicly available resource that would allow for aggregation of publicly available processed data to generate these types of plots. Creating a resource that would generate coloured Ribo-Seq plots on the fly that would be useful to both myself and the research community in general, was the motivation behind creating Trips-Viz. While the usefulness of three frame colouring is obvious in the case of nested/internal ORFs, it can be beneficial in many other cases too, see Figure 4.4.

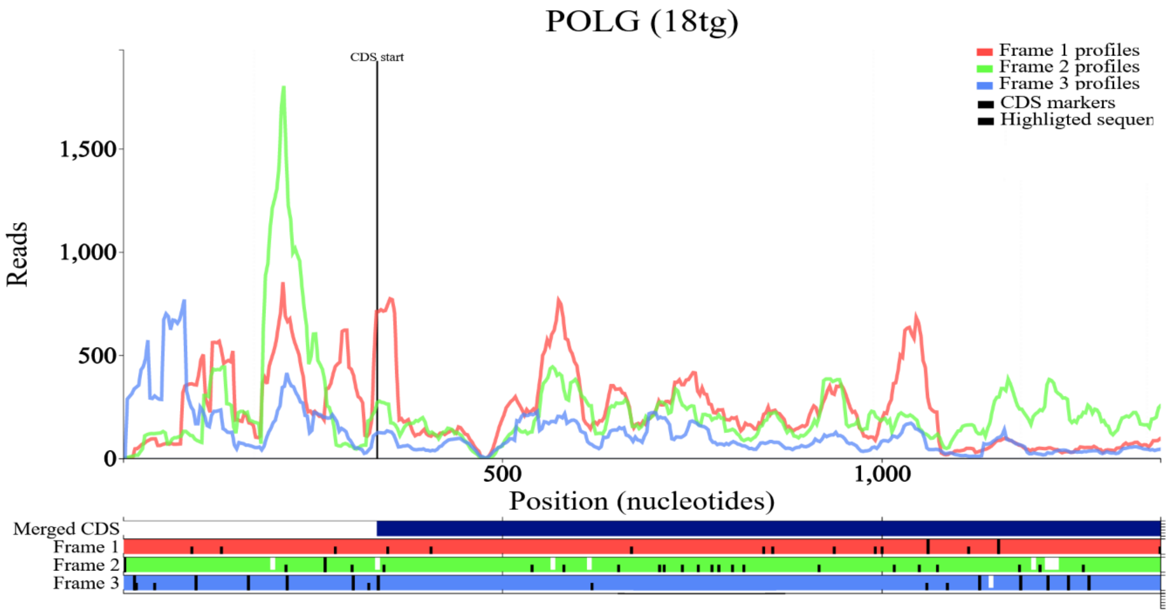

Figure 4.4*: An aggregated set of ribosome profiling data from multiple studies aligned to the human POLG gene. A zoomed in section of the 5' Leader and partial CDS is shown. A long overlapping uORF that initiates at a CUG codon (short black lines in the ORF architecture) can be seen in Frame 1. This would be difficult to see without three frame colouring and is complicated further by the apparent translation of two separate uORFs in the 5' leader in frame 2 (green) and frame 3 (blue).*

Aggregating ribosome profiling data has issues in that the strength of the periodicity can vary between studies and even between datasets of the same study. Trips-Viz includes a meta-information section where users can generate periodicity plots for any dataset, these plots include a periodicity score between 0 (weakest) and 1 (strongest), to allows them to pre-select only strong periodicity datasets. To make this process even faster a filter can be applied when generating plots that will discard all reads below a certain periodicity score, allowing users to quickly increase the periodicity signal in the plots at the expense of read depth (Figure 4.5).

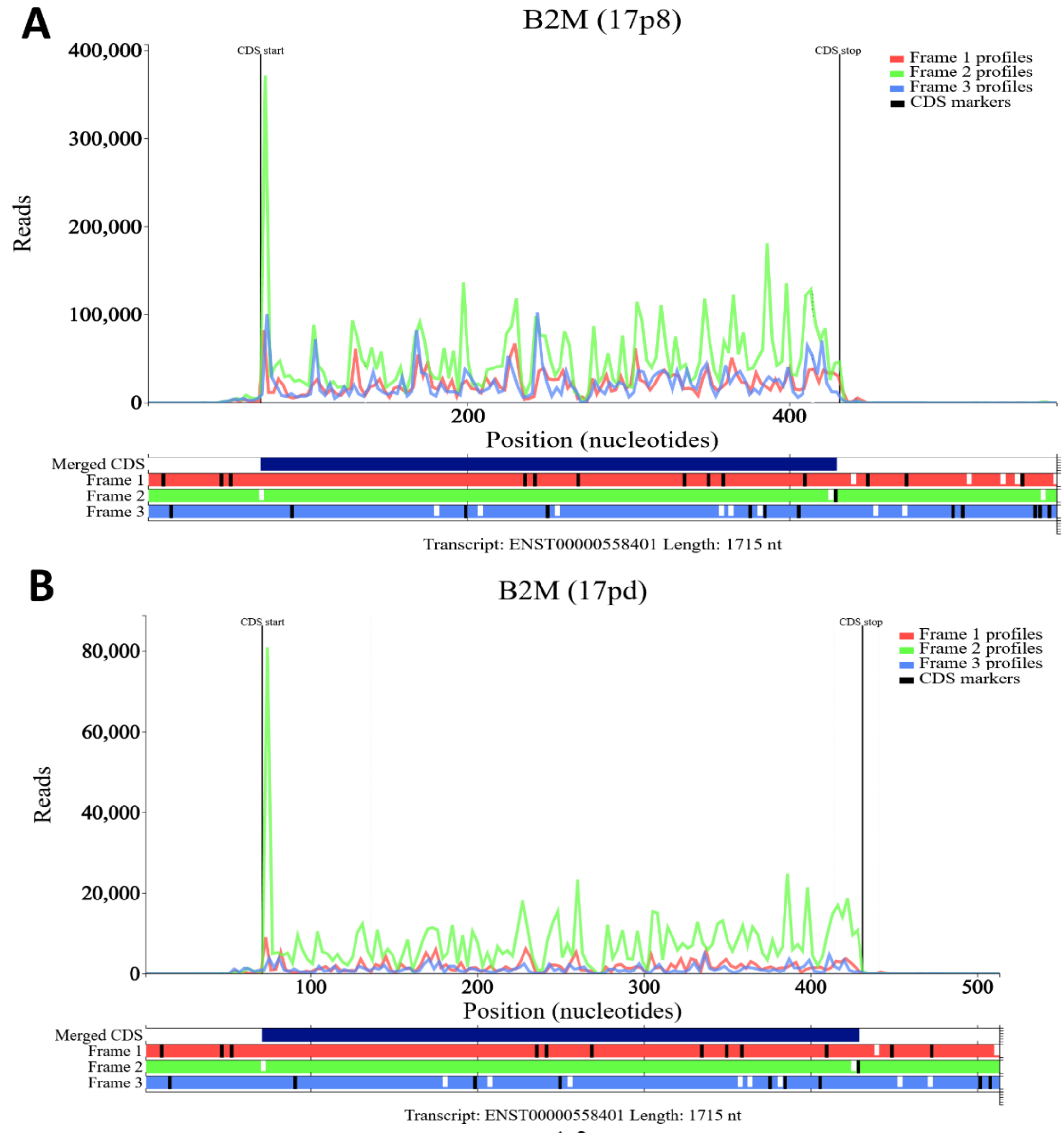

Figure 4.5*: Panel A shows aggregated ribosome profiling data from the human gene B2M with no filter. Panel B shows the same gene with a periodicity filter applied which removes read lengths with weak periodicity. This results in less reads but a better separation between the in-frame reads (frame 2, green) and out of frame reads (frames 1 and 3, red and blue).*

Another issue that can prevent discovery of novel translated ORFs is the exclusion of ambiguously mapped reads. Due to the typically short read lengths present in ribosome profiling data, ambiguous mapping is more prevalent in ribosome profiling data than it is in many other sequencing types. Many pipelines deal with this issue by simply discarding any reads that map ambiguously. This has the advantage of being able to fully trust that any aligned reads cannot be coming from anywhere else in the genome but can mean that some translated ORFs may be missed. Trips-Viz handles this by allowing users to choose whether to turn on or off ambiguously mapped reads on the fly. This must be used with caution but can make some translated ORFs detectable where otherwise they wouldn’t be (Figure 4.6).

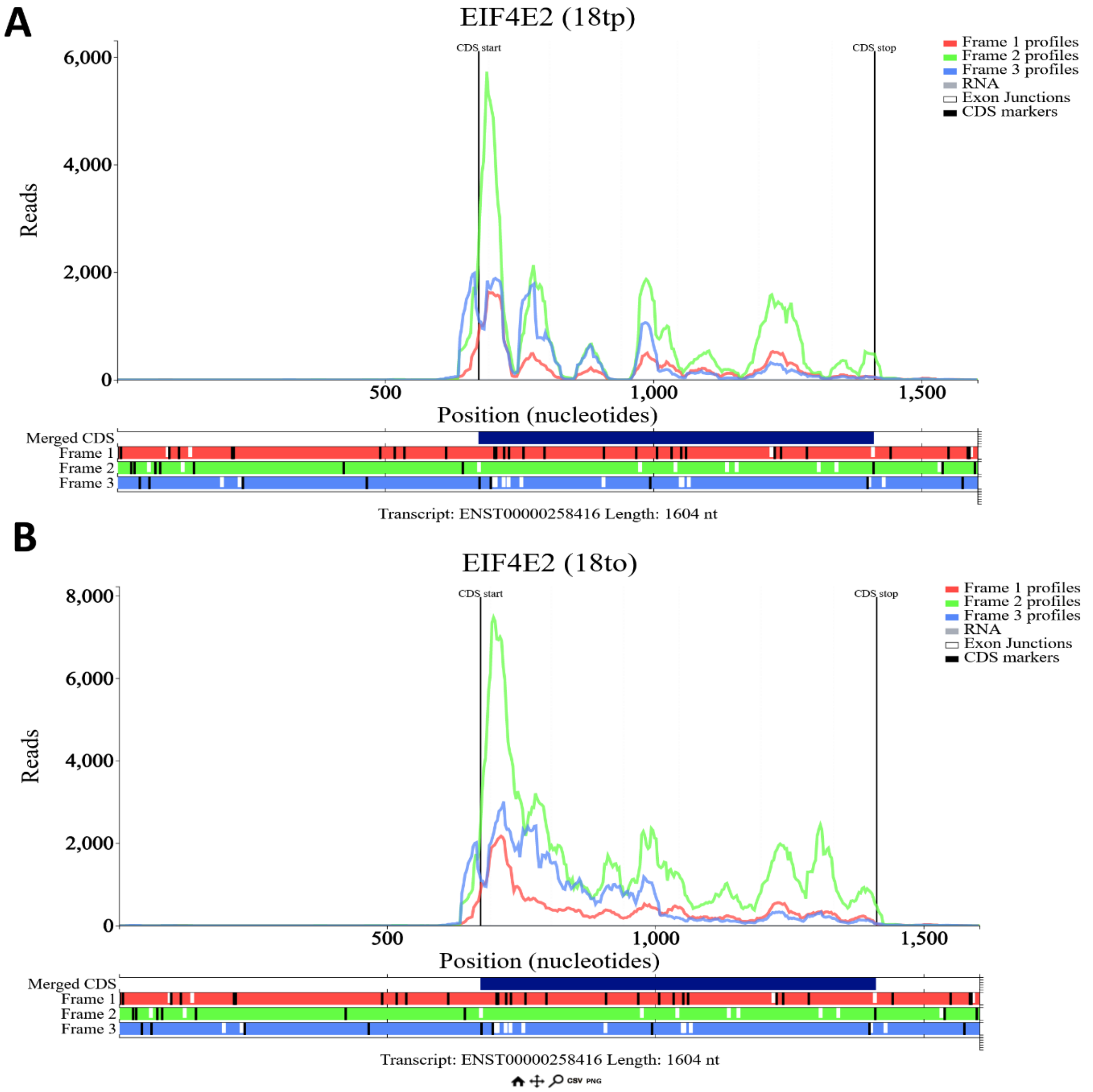

Figure 4.6: *Panel A shows aggregated ribosome profiling data for the human gene EIF4E2. Only unambiguous reads are shown. This results in numerous gaps within the CDS where no ribo-seq reads are mapped. Panel B shows the same gene but with ambiguously mapped reads permitted. This removes the gaps within the CDS and makes the presence of a translated internal ORF in frame 3 (blue) clearly visible.*

While Trips-Viz is also capable of carrying out other types of analysis the visualization is its strongest and most unique feature and is likely largely responsible for its popularity. Currently Trips-Viz has 137 registered users and receives ~160 unique visitors per month, however this number is steadily increasing. To date Trips-Viz has 25 citations and has been used to visualize ribosome profiling data in a number of papers including the overlapping uORF POLG (Loughran et al., 2020), Khan et al., 2020), lncRNA translation (Konina et al., 2021) and uORFs (Filatova et al., 2021).

Trips-Viz has been relatively successful in fulfilling its objective of being a useful resource for the community, however there is still many areas in which it can improve and expand. At the most basic level this simply includes plans to continue processing ribosome profiling data for existing organisms and addition of others. Other related data types may also be included, for example disome-seq (Meydan et al., 2020) and epitranscriptomic data (Jantsch et al., 2018). Given it's relative popularity incorporation of a method to detect pauses, using an existing software such as PausePred is also planned (Kumari et al., 2018). The Meta-Information section will be expanded by the addition of fastq-screen outputs (Wingett et al., 2018) to screen for common contaminants in sequencing data, given that this is a prevalent issue in RNA-Seq data (Olarerin-George et al., 2015) it is likely many of the publicly available Ribo-Seq studies also suffer from this issue. Lastly there are planned improvements to the method of transcript selection on Trips-Viz, moving toward a more data driven approach whereby users will be able to select transcripts based on which ones are best supported by the RNA-Seq data as well as better incorporation with Gwips-Viz (Michel et al., 2018) so that users can choose to view transcripts after viewing RNA-Seq at the genome level.

# Bibliography

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., . . . Blankenberg, D. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res, 46*(W1), W537-W544. doi:10.1093/nar/gky379

Albert, F. W., Muzzey, D., Weissman, J. S., & Kruglyak, L. (2014). Genetic influences on translation in yeast. *PLoS Genet, 10*(10), e1004692. doi:10.1371/journal.pgen.1004692

Andreev, D. E., O'Connor, P. B., Fahey, C., Kenny, E. M., Terenin, I. M., Dmitriev, S. E., . . . Baranov, P. V. (2015). Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *Elife, 4*, e03971. doi:10.7554/eLife.03971

Andreev, D. E., O'Connor, P. B., Loughran, G., Dmitriev, S. E., Baranov, P. V., & Shatsky, I. N. (2017). Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res, 45*(2), 513-526. doi:10.1093/nar/gkw1190

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

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*(7613), 570-574. doi:10.1038/nature18647

Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., . . . Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J, 33*(9), 981-993. doi:10.1002/embj.201488411

Becker, A. H., Oh, E., Weissman, J. S., Kramer, G., & Bukau, B. (2013). Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. *Nat Protoc, 8*(11), 2212-2239. doi:10.1038/nprot.2013.133

Berg, J. A., Belyeu, J. R., Morgan, J. T., Ouyang, Y., Bott, A. J., Quinlan, A. R., . . . Rutter, J. (2020). XPRESSyourself: Enhancing, standardizing, and automating ribosome profiling computational analyses yields improved insight into data. *PLoS Comput Biol, 16*(1), e1007625. doi:10.1371/journal.pcbi.1007625

Birkeland, A., ChyZynska, K., & Valen, E. (2018). Shoelaces: an interactive tool for ribosome profiling processing and visualization. *BMC Genomics, 19*(1), 543. doi:10.1186/s12864-018-4912-6

Blencowe, B. J. (2006). Alternative splicing: new insights from global analyses. *Cell, 126*(1), 37-47. doi:10.1016/j.cell.2006.06.023

Brandman, O., & Hegde, R. S. (2016). Ribosome-associated protein quality control. *Nat Struct Mol Biol, 23*(1), 7-15. doi:10.1038/nsmb.3147

Brar, G. A., & Weissman, J. S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol, 16*(11), 651-664. doi:10.1038/nrm4069

Brown, A., Rathore, S., Kimanius, D., Aibara, S., Bai, X. C., Rorbach, J., . . . Ramakrishnan, V. (2017). Structures of the human mitochondrial ribosome in native states of assembly. *Nat Struct Mol Biol, 24*(10), 866-869. doi:10.1038/nsmb.3464

Brunet, M. A., Brunelle, M., Lucier, J. F., Delcourt, V., Levesque, M., Grenier, F., . . . Roucou, X. (2018). OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res*. doi:10.1093/nar/gky936

Brunet, M. A., Brunelle, M., Lucier, J. F., Delcourt, V., Levesque, M., Grenier, F., . . . Roucou, X. (2019). OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res, 47*(D1), D403-D410. doi:10.1093/nar/gky936

Brunet, M. A., Levesque, S. A., Hunting, D. J., Cohen, A. A., & Roucou, X. (2018). Recognition of the polycistronic nature of human genes is critical to understanding the genotype-phenotype relationship. *Genome Res, 28*(5), 609-624. doi:10.1101/gr.230938.117

Buskirk, A. R., & Green, R. (2017). Ribosome pausing, arrest and rescue in bacteria and eukaryotes. *Philos Trans R Soc Lond B Biol Sci, 372*(1716). doi:10.1098/rstb.2016.0183

Calviello, L., Hirsekorn, A., & Ohler, U. (2019). SaTAnn quantifies translation on the functionally heterogeneous transcriptome. doi:10.1101/608794

Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., . . . Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. *Nat Methods, 13*(2), 165-170. doi:10.1038/nmeth.3688

Calviello, L., & Ohler, U. (2017). Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet, 33*(10), 728-744. doi:10.1016/j.tig.2017.08.003

Calviello, L., Sydow, D., Harnett, D., & Ohler, U. (2019). doi:10.1101/601468

Cao, X., Khitun, A., Na, Z., Dumitrescu, D. G., Kubica, M., Olatunji, E., & Slavoff, S. A. (2020). Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines. *J Proteome Res, 19*(8), 3418-3426. doi:10.1021/acs.jproteome.0c00254

Carja, O., Xing, T., Wallace, E. W. J., Plotkin, J. B., & Shah, P. (2017). riboviz: analysis and visualization of ribosome profiling datasets. *BMC Bioinformatics, 18*(1), 461. doi:10.1186/s12859-017-1873-8

Castelo-Szekely, V., De Matos, M., Tusup, M., Pascolo, S., Ule, J., & Gatfield, D. (2019). Charting DENR-dependent translation reinitiation uncovers predictive uORF features and links to circadian timekeeping via Clock. *Nucleic Acids Res, 47*(10), 5193-5209. doi:10.1093/nar/gkz261

Chew, G. L., Pauli, A., Rinn, J. L., Regev, A., Schier, A. F., & Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development, 140*(13), 2828-2834. doi:10.1242/dev.098343

Chugunova, A., Loseva, E., Mazin, P., Mitina, A., Navalayeu, T., Bilan, D., . . . Dontsova, O. (2019). LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism. *Proc Natl Acad Sci U S A, 116*(11), 4940-4945. doi:10.1073/pnas.1809105116

Chun, S. Y., Rodriguez, C. M., Todd, P. K., & Mills, R. E. (2016). SPECtre: a spectral coherence--based classifier of actively translated transcripts from ribosome profiling sequence data. *BMC Bioinformatics, 17*(1), 482. doi:10.1186/s12859-016-1355-4

Chung, B. Y., Hardcastle, T. J., Jones, J. D., Irigoyen, N., Firth, A. E., Baulcombe, D. C., & Brierley, I. (2015). The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA, 21*(10), 1731-1745. doi:10.1261/rna.052548.115

Clauwaert, J., Menschaert, G., & Waegeman, W. (2019). DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. *Nucleic Acids Res, 47*, e36. doi:10.1093/nar/gkz061

Crappe, J., Ndah, E., Koch, A., Steyaert, S., Gawron, D., De Keulenaer, S., . . . Menschaert, G. (2015). PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res, 43*(5), e29. doi:10.1093/nar/gku1283

da Veiga Leprevost, F., Haynes, S. E., Avtonomov, D. M., Chang, H. Y., Shanmugam, A. K., Mellacheruvu, D., . . . Nesvizhskii, A. I. (2020). Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat Methods, 17*, 869–870. doi:10.1038/s41592-020-0912-y

Delcourt, V., Brunelle, M., Roy, A. V., Jacques, J. F., Salzet, M., Fournier, I., & Roucou, X. (2018). The Protein Coded by a Short Open Reading Frame, Not by the Annotated Coding Sequence, Is the Main Gene Product of the Dual-Coding Gene MIEF1. *Mol Cell Proteomics, 17*(12), 2402-2411. doi:10.1074/mcp.RA118.000593

Desiere, F., Deutsch, E. W., King, N. L., Nesvizhskii, A. I., Mallick, P., Eng, J., . . . Aebersold, R. (2006). The PeptideAtlas project. *Nucleic Acids Res, 34*(Database issue), D655-658. doi:10.1093/nar/gkj040

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Dunn, J. G., & Weissman, J. S. (2016). Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics, 17*(1), 958. doi:10.1186/s12864-016-3278-x

Egorov, A. A., Sakharova, E. A., Anisimova, A. S., Dmitriev, S. E., Gladyshev, V. N., & Kulakovskiy, I. V. (2019). svist4get: a simple visualization tool for genomic tracks from sequencing experiments. *BMC Bioinformatics, 20*(1), 113. doi:10.1186/s12859-019-2706-8

Eichhorn, S. W., Guo, H., McGeary, S. E., Rodriguez-Mias, R. A., Shin, C., Baek, D., . . . Bartel, D. P. (2014). mRNA destabilization is the dominant effect of mammalian microRNAs by the time substantial repression ensues. *Mol Cell, 56*(1), 104-115. doi:10.1016/j.molcel.2014.08.028

Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D. J., Weekes, M. P., . . . Dolken, L. (2018). Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods, 15*(5), 363-366. doi:10.1038/nmeth.4631

Ernlund, A. W., Schneider, R. J., & Ruggles, K. V. (2018). RIVET: comprehensive graphic user interface for analysis and exploration of genome-wide translatomics data. *BMC Genomics, 19*(1), 809. doi:10.1186/s12864-018-5166-z

Fields, A. P., Rodriguez, E. H., Jovanovic, M., Stern-Ginossar, N., Haas, B. J., Mertins, P., . . . Weissman, J. S. (2015). A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Mol Cell, 60*(5), 816-827. doi:10.1016/j.molcel.2015.11.013

Filatova, A. Y., Vasilyeva, T. A., Marakhonov, A. V., Sukhanova, N. V., Voskresenskaya, A. A., Zinchenko, R. A., & Skoblov, M. Y. (2021). Upstream ORF frameshift variants in the PAX6 5'UTR cause congenital aniridia. *Hum Mutat, 42*(8), 1053-1065. doi:10.1002/humu.24248

Fluman, N., Navon, S., Bibi, E., & Pilpel, Y. (2014). mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. *Elife, 3*. doi:10.7554/eLife.03440

Francois, P., Arbes, H., Demais, S., Baudin-Baillieu, A., & Namy, O. (2021). RiboDoc: A Docker-based package for ribosome profiling analysis. *Comput Struct Biotechnol J, 19*, 2851-2860. doi:10.1016/j.csbj.2021.05.014

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., . . . Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res, 47*(D1), D766-D773. doi:10.1093/nar/gky955

Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., . . . Brosch, M. (2012). Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Res, 22*(11), 2208-2218. doi:10.1101/gr.139568.112

Gandin, V., Masvidal, L., Hulea, L., Gravel, S. P., Cargnello, M., McLaughlan, S., . . . Topisirovic, I. (2016). nanoCAGE reveals 5' UTR features that define specific modes of translation of functionally related MTOR-sensitive mRNAs. *Genome Res, 26*(5), 636-648. doi:10.1101/gr.197566.115

Gao, X., Wan, J., Liu, B., Ma, M., Shen, B., & Qian, S. B. (2015). Quantitative profiling of initiating ribosomes in vivo. *Nat Methods, 12*(2), 147-153. doi:10.1038/nmeth.3208

Gerashchenko, M. V., & Gladyshev, V. N. (2017). Ribonuclease selection for ribosome profiling. *Nucleic Acids Res, 45*(2), e6. doi:10.1093/nar/gkw822

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., & Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell, 154*(1), 240-251. doi:10.1016/j.cell.2013.06.009

H. Backman TW, & Girke, T. (2016). systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics, 17*, 388. doi:10.1186/s12859-016-1241-0

Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., . . . Chen, R. (2018). SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief Bioinform, 19*(4), 636-643. doi:10.1093/bib/bbx005

Hardy, S., Kostantin, E., Wang, S. J., Hristova, T., Galicia-Vazquez, G., Baranov, P. V., . . . Tremblay, M. L. (2019). Magnesium-sensitive upstream ORF controls PRL phosphatase expression to mediate energy metabolism. *Proc Natl Acad Sci U S A, 116*(8), 2925-2934. doi:10.1073/pnas.1815361116

Hinnebusch, A. G. (2014). The scanning mechanism of eukaryotic translation initiation. *Annu Rev Biochem, 83*, 779-812. doi:10.1146/annurev-biochem-060713-035802

Hinnebusch, A. G., Ivanov, I. P., & Sonenberg, N. (2016). Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science, 352*(6292), 1413-1416. doi:10.1126/science.aad9868

Inada, T. (2013). Quality control systems for aberrant mRNAs induced by aberrant translation elongation and termination. *Biochim Biophys Acta, 1829*(6-7), 634-642. doi:10.1016/j.bbagrm.2013.02.004

Ingolia, N. T. (2014). Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet, 15*(3), 205-213. doi:10.1038/nrg3645

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J., Jackson, S. E., . . . Weissman, J. S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep, 8*(5), 1365-1379. doi:10.1016/j.celrep.2014.07.045

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*(5924), 218-223. doi:10.1126/science.1168978

Ingolia, N. T., Hussmann, J. A., & Weissman, J. S. (2018). Ribosome Profiling: Global Views of Translation. *Cold Spring Harb Perspect Biol*. doi:10.1101/cshperspect.a032698

Ingolia, N. T., Lareau, L. F., & Weissman, J. S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell, 147*(4), 789-802. doi:10.1016/j.cell.2011.10.002

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., . . . Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods, 11*(2), 163-166. doi:10.1038/nmeth.2772

Ivanov, I. P., Firth, A. E., Michel, A. M., Atkins, J. F., & Baranov, P. V. (2011). Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences. *Nucleic Acids Res, 39*(10), 4220-4234. doi:10.1093/nar/gkr007

Ivanov, I. P., Shin, B. S., Loughran, G., Tzani, I., Young-Baird, S. K., Cao, C., . . . Dever, T. E. (2018). Polyamine Control of Translation Elongation Regulates Start Site Selection on Antizyme Inhibitor mRNA via Ribosome Queuing. *Mol Cell, 70*(2), 254-264 e256. doi:10.1016/j.molcel.2018.03.015

Iwasaki, S., Floor, S. N., & Ingolia, N. T. (2016). Rocaglates convert DEAD-box protein eIF4A into a sequence-selective translational repressor. *Nature, 534*(7608), 558-561. doi:10.1038/nature17978

Jantsch, M. F., Quattrone, A., O'Connell, M., Helm, M., Frye, M., Macias-Gonzales, M., . . . Fray, R. (2018). Positioning Europe for the EPITRANSCRIPTOMICS challenge. *RNA Biol, 15*(6), 829-831. doi:10.1080/15476286.2018.1460996

Ji, Z. (2018). Rfoot: Transcriptome-Scale Identification of RNA-Protein Complexes from Ribosome Profiling Data. *Curr Protoc Mol Biol, 124*(1), e66. doi:10.1002/cpmb.66

Ji, Z., Song, R., Regev, A., & Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife, 4*, e08890. doi:10.7554/eLife.08890

Johnstone, T. G., Bazzini, A. A., & Giraldez, A. J. (2016). Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J, 35*(7), 706-723. doi:10.15252/embj.201592759

Jungreis, I., Lin, M. F., Spokony, R., Chan, C. S., Negre, N., Victorsen, A., . . . Kellis, M. (2011). Evidence of abundant stop codon readthrough in Drosophila and other metazoa. *Genome Res, 21*(12), 2096-2113. doi:10.1101/gr.119974.110

Khan, Y. A., Jungreis, I., Wright, J. C., Mudge, J. M., Choudhary, J. S., Firth, A. E., & Kellis, M. (2020). Evidence for a novel overlapping coding sequence in POLG initiated at a CUG start codon. *BMC Genet, 21*(1), 25. doi:10.1186/s12863-020-0828-7

Kiniry, S. J., Judge, C. E., Michel, A. M., & Baranov, P. V. (2021). Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. *Nucleic Acids Res*. doi:10.1093/nar/gkab323

Kiniry, S. J., Michel, A. M., & Baranov, P. V. (2018). The GWIPS-viz Browser. *Curr Protoc Bioinformatics, 62*(1), e50. doi:10.1002/cpbi.50

Kiniry, S. J., Michel, A. M., & Baranov, P. V. (2020). Computational methods for ribosome profiling data analysis. *Wiley Interdiscip Rev RNA, 11*(3), e1577. doi:10.1002/wrna.1577

Kiniry, S. J., O'Connor, P. B. F., Michel, A. M., & Baranov, P. V. (2018). Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. *Nucleic Acids Res, 47*, D847–D852. doi:10.1093/nar/gky842

Kirchner, S., Cai, Z., Rauscher, R., Kastelic, N., Anding, M., Czech, A., . . . Ignatova, Z. (2017). Alteration of protein function by a silent polymorphism linked to tRNA abundance. *PLoS Biol, 15*(5), e2000779. doi:10.1371/journal.pbio.2000779

Kivioja, T., Vaharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2011). Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods, 9*(1), 72-74. doi:10.1038/nmeth.1778

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods, 14*(5), 513-520. doi:10.1038/nmeth.4256

Konina, D., Sparber, P., Viakhireva, I., Filatova, A., & Skoblov, M. (2021). Investigation of LINC00493/SMIM26 Gene Suggests Its Dual Functioning at mRNA and Protein Level. *Int J Mol Sci, 22*(16). doi:10.3390/ijms22168477

Kumari, R., Michel, A. M., & Baranov, P. V. (2018). PausePred and Rfeet: webtools for inferring ribosome pauses and visualizing footprint density from ribosome profiling data. *RNA, 24*(10), 1297-1304. doi:10.1261/rna.065235.117

Kurian, L., Palanimurugan, R., Godderz, D., & Dohmen, R. J. (2011). Polyamine sensing by nascent ornithine decarboxylase antizyme stimulates decoding of its mRNA. *Nature, 477*(7365), 490-494. doi:10.1038/nature10393

Langmead, B., & Nellore, A. (2018). Cloud computing for genomic data analysis and collaboration. *Nat Rev Genet, 19*(5), 325. doi:10.1038/nrg.2018.8

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(3), R25. doi:10.1186/gb-2009-10-3-r25

Lareau, L. F., Hite, D. H., Hogan, G. J., & Brown, P. O. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife, 3*, e01257. doi:10.7554/eLife.01257

Larsson, O., Sonenberg, N., & Nadon, R. (2010). Identification of differential translation in genome wide studies. *Proc Natl Acad Sci U S A, 107*(50), 21487-21492. doi:10.1073/pnas.1006821107

Lauria, F., Tebaldi, T., Bernabo, P., Groen, E. J. N., Gillingwater, T. H., & Viero, G. (2018). riboWaltz: Optimization of ribosome P-site positioning in ribosome profiling data. *PLoS Comput Biol, 14*(8), e1006169. doi:10.1371/journal.pcbi.1006169

Lecanda, A., Nilges, B. S., Sharma, P., Nedialkova, D. D., Schwarz, J., Vaquerizas, J. M., & Leidel, S. A. (2016). Dual randomization of oligonucleotides to reduce the bias in ribosome-profiling libraries. *Methods, 107*, 89-97. doi:10.1016/j.ymeth.2016.07.011

Lee, S., Liu, B., Lee, S., Huang, S. X., Shen, B., & Qian, S. B. (2012). Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A, 109*(37), E2424-2432. doi:10.1073/pnas.1207846109

Li, G. W., Oh, E., & Weissman, J. S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature, 484*(7395), 538-541. doi:10.1038/nature10965

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, W., Wang, W., Uren, P. J., Penalva, L. O. F., & Smith, A. D. (2017). Riborex: fast and flexible identification of differential translation from Ribo-seq data. *Bioinformatics, 33*(11), 1735-1737. doi:10.1093/bioinformatics/btx047

Liu, Q., Shvarts, T., Sliz, P., & Gregory, R. I. (2020). RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. *Nucleic Acids Res, 48*(W1), W218-W229. doi:10.1093/nar/gkaa395

Liu, T. Y., & Song, Y. S. (2016). Prediction of ribosome footprint profile shapes from transcript sequences. *Bioinformatics, 32*(12), i183-i191. doi:10.1093/bioinformatics/btw253

Liu, W., Xiang, L., Zheng, T., Jin, J., & Zhang, G. (2018). TranslatomeDB: a comprehensive database and cloud-based analysis platform for translatome sequencing data. *Nucleic Acids Res, 46*(D1), D206-D212. doi:10.1093/nar/gkx1034

Loayza-Puch, F., Rooijers, K., Buil, L. C., Zijlstra, J., Oude Vrielink, J. F., Lopes, R., . . . Agami, R. (2016). Tumour-specific proline vulnerability uncovered by differential ribosome codon reading. *Nature, 530*(7591), 490-494. doi:10.1038/nature16982

Lorenz, R., Bernhart, S. H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol, 6*, 26. doi:10.1186/1748-7188-6-26

Loughran, G., Jungreis, I., Tzani, I., Power, M., Dmitriev, R. I., Ivanov, I. P., . . . Atkins, J. F. (2018). Stop codon readthrough generates a C-terminally extended variant of the human vitamin D receptor with reduced calcitriol response. *J Biol Chem, 293*(12), 4434-4444. doi:10.1074/jbc.M117.818526

Loughran, G., Zhdanov, A. V., Mikhaylova, M. S., Rozov, F. N., Datskevich, P. N., Kovalchuk, S. I., . . . Andreev, D. E. (2020). Unusually efficient CUG initiation of an overlapping reading frame in POLG mRNA yields novel protein POLGARF. *Proc Natl Acad Sci U S A, 117*(40), 24936-24946. doi:10.1073/pnas.2001433117

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol, 15*(12), 550. doi:10.1186/s13059-014-0550-8

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet, 17*(1).

Martinez, T. F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M. N., & Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. *Nat Chem Biol, 16*(4), 458-468. doi:10.1038/s41589-019-0425-0

Matsuo, Y., Ikeuchi, K., Saeki, Y., Iwasaki, S., Schmidt, C., Udagawa, T., . . . Inada, T. (2017). Ubiquitination of stalled ribosome triggers ribosome-associated quality control. *Nat Commun, 8*(1), 159. doi:10.1038/s41467-017-00188-1

McGlincy, N. J., & Ingolia, N. T. (2017). Transcriptome-wide measurement of translation by ribosome profiling. *Methods, 126*, 112-129. doi:10.1016/j.ymeth.2017.05.028

Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappe, J., Gevaert, K., & Van Damme, P. (2013). Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol Cell Proteomics, 12*(7), 1780-1790. doi:10.1074/mcp.M113.027540

Meydan, S., & Guydosh, N. R. (2020). Disome and Trisome Profiling Reveal Genome-wide Targets of Ribosome Quality Control. *Mol Cell, 79*(4), 588-602 e586. doi:10.1016/j.molcel.2020.06.010

Michel, A. M., Ahern, A. M., Donohue, C. A., & Baranov, P. V. (2015). GWIPS-viz as a tool for exploring ribosome profiling evidence supporting the synthesis of alternative proteoforms. *Proteomics, 15*(14), 2410-2416. doi:10.1002/pmic.201400603

Michel, A. M., Andreev, D. E., & Baranov, P. V. (2014). Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. *BMC Bioinformatics, 15*, 380. doi:10.1186/s12859-014-0380-4

Michel, A. M., & Baranov, P. V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdiscip Rev RNA, 4*(5), 473-490. doi:10.1002/wrna.1172

Michel, A. M., Choudhury, K. R., Firth, A. E., Ingolia, N. T., Atkins, J. F., & Baranov, P. V. (2012). Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res, 22*(11), 2219-2229. doi:10.1101/gr.133249.111

Michel, A. M., Fox, G., A, M. K., De Bo, C., O'Connor, P. B., Heaphy, S. M., . . . Baranov, P. V. (2014). GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res, 42*(Database issue), D859-864. doi:10.1093/nar/gkt1035

Michel, A. M., Kiniry, S. J., O'Connor, P. B. F., Mullan, J. P., & Baranov, P. V. (2018). GWIPS-viz: 2018 update. *Nucleic Acids Res, 46*(D1), D823-D830. doi:10.1093/nar/gkx790

Michel, A. M., Mullan, J. P., Velayudhan, V., O'Connor, P. B., Donohue, C. A., & Baranov, P. V. (2016). RiboGalaxy: A browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA Biol, 13*(3), 316-319. doi:10.1080/15476286.2016.1141862

Miettinen, T. P., & Bjorklund, M. (2015). Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions. *Nucleic Acids Res, 43*(2), 1019-1034. doi:10.1093/nar/gku1310

Mohammad, F., Woolstenhulme, C. J., Green, R., & Buskirk, A. R. (2016). Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep, 14*(4), 686-694. doi:10.1016/j.celrep.2015.12.073

Ndah, E., Jonckheere, V., Giess, A., Valen, E., Menschaert, G., & Van Damme, P. (2017). REPARATION: ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic Acids Res, 45*(20), e168. doi:10.1093/nar/gkx758

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., . . . Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. *Nature, 543*(7643), 72-77. doi:10.1038/nature21373

O'Connor, P. B., Andreev, D. E., & Baranov, P. V. (2016). Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat Commun, 7*, 12915. doi:10.1038/ncomms12915

O'Connor, P. B., Li, G. W., Weissman, J. S., Atkins, J. F., & Baranov, P. V. (2013). rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. *Bioinformatics, 29*(12), 1488-1491. doi:10.1093/bioinformatics/btt184

Oertlin, C., Lorent, J., Murie, C., Furic, L., Topisirovic, I., & Larsson, O. (2019). Generally applicable transcriptome-wide analysis of translation using anota2seq. *Nucleic Acids Res, 47*, e70. doi:10.1093/nar/gkz223

Olarerin-George, A. O., & Hogenesch, J. B. (2015). Assessing the prevalence of mycoplasma contamination in cell culture via a survey of NCBI's RNA-seq archive. *Nucleic Acids Res, 43*(5), 2535-2542. doi:10.1093/nar/gkv136

Olexiouk, V., Van Criekinge, W., & Menschaert, G. (2018). An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res, 46*(D1), D497-D502. doi:10.1093/nar/gkx1130

Olshen, A. B., Hsieh, A. C., Stumpf, C. R., Olshen, R. A., Ruggero, D., & Taylor, B. S. (2013). Assessing gene-level translational control from ribosome profiling. *Bioinformatics, 29*(23), 2995-3002. doi:10.1093/bioinformatics/btt533

Ozadam, H., Geng, M., & Cenik, C. (2020). RiboFlow, RiboR and RiboPy: an ecosystem for analyzing ribosome profiling data at read length resolution. *Bioinformatics, 36*(9), 2929-2931. doi:10.1093/bioinformatics/btaa028

Perkins, P., Mazzoni-Putman, S., Stepanova, A., Alonso, J., & Heber, S. (2019). RiboStreamR: a web application for quality control, analysis, and visualization of Ribo-seq data. *BMC Genomics, 20*(S5). doi:10.1186/s12864-019-5700-7

Perkins, P., Mazzoni-Putman, S., Stepanova, A., Alonso, J., & Heber, S. (2019). RiboStreamR: a web application for quality control, analysis, and visualization of Ribo-seq data. *BMC Genomics, 20*(Suppl 5), 422. doi:10.1186/s12864-019-5700-7

Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., & Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol Syst Biol, 10*, 770. doi:10.15252/msb.20145524

Popa, A., Lebrigand, K., Paquet, A., Nottet, N., Robbe-Sermesant, K., Waldmann, R., & Barbry, P. (2016). RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Res, 5*, 1309. doi:10.12688/f1000research.8964.1

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics, 32*, 496. doi:10.1038/ng1032

Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., . . . Pritchard, J. K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife, 5*. doi:10.7554/eLife.13328

Rajput, B., Pruitt, K. D., & Murphy, T. D. (2019). RefSeq curation and annotation of stop codon recoding in vertebrates. *Nucleic Acids Res, 47*(2), 594-606. doi:10.1093/nar/gky1234

Rathore, A., Chu, Q., Tan, D., Martinez, T. F., Donaldson, C. J., Diedrich, J. K., . . . Saghatelian, A. (2018). MIEF1 Microprotein Regulates Mitochondrial Translation. *Biochemistry, 57*(38), 5564-5575. doi:10.1021/acs.biochem.8b00726

Reid, D. W., Xu, D., Chen, P., Yang, H., & Sun, L. (2017). Integrative analyses of translatome and transcriptome reveal important translational controls in brown and white adipose regulated by microRNAs. *Sci Rep, 7*(1), 5681. doi:10.1038/s41598-017-06077-3

Reixachs-Solé, M., Ruiz-Orera, J., Alba, M. M., & Eyras, E. (2019). Ribosome profiling at isoform level reveals an evolutionary conserved impact of differential splicing on the proteome. doi:10.1101/582031

Reuter, K., Biehl, A., Koch, L., & Helms, V. (2016). PreTIS: A Tool to Predict Non-canonical 5' UTR Translational Initiation Sites in Human and Mouse. *PLoS Comput Biol, 12*(10), e1005170. doi:10.1371/journal.pcbi.1005170

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi:10.1093/bioinformatics/btp616

Rodriguez, J. M., Rodriguez-Rivas, J., Di Domenico, T., Vazquez, J., Valencia, A., & Tress, M. L. (2018). APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res, 46*(D1), D213-D217. doi:10.1093/nar/gkx997

Schmidt, T., Samaras, P., Frejno, M., Gessulat, S., Barnert, M., Kienegger, H., . . . Wilhelm, M. (2018). ProteomicsDB. *Nucleic Acids Res, 46*(D1), D1271-D1281. doi:10.1093/nar/gkx1029

Schueren, F., Lingner, T., George, R., Hofhuis, J., Dickel, C., Gartner, J., & Thoms, S. (2014). Peroxisomal lactate dehydrogenase is generated by translational readthrough in mammals. *Elife, 3*, e03640. doi:10.7554/eLife.03640

Shirokikh, N. E., & Preiss, T. (2018). Translation initiation by cap-dependent ribosome recruitment: Recent insights and open questions. *Wiley Interdiscip Rev RNA, 9*(4), e1473. doi:10.1002/wrna.1473

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res, 27*(3), 491-499. doi:10.1101/gr.209601.116

Somogyi, P., Jenner, A. J., Brierley, I., & Inglis, S. C. (1993). Ribosomal pausing during translation of an RNA pseudoknot. *Mol Cell Biol, 13*(11), 6931-6940. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8413285

Stefan, D. (2016). RiboPip. Retrieved from https://github.com/stepf/RiboPip

Stern-Ginossar, N., & Ingolia, N. T. (2015). Ribosome Profiling as a Tool to Decipher Viral Complexity. *Annu Rev Virol, 2*(1), 335-349. doi:10.1146/annurev-virology-100114-054854

Tanaka, M., Sotta, N., Yamazumi, Y., Yamashita, Y., Miwa, K., Murota, K., . . . Fujiwara, T. (2016). The Minimum Open Reading Frame, AUG-Stop, Induces Boron-Dependent Ribosome Stalling and mRNA Degradation. *Plant Cell, 28*(11), 2830-2849. doi:10.1105/tpc.16.00481

Tenson, T., & Ehrenberg, M. (2002). Regulatory nascent peptides in the ribosomal tunnel. *Cell, 108*(5), 591-594. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11893330

Tholstrup, J., Oddershede, L. B., & Sorensen, M. A. (2012). mRNA pseudoknot structures can act as ribosomal roadblocks. *Nucleic Acids Res, 40*(1), 303-313. doi:10.1093/nar/gkr686

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE, 70*(9), 1055-1096. doi:10.1109/PROC.1982.12433

Tjeldnes, H., Labun, K., Torres Cleuren, Y., Chyzynska, K., Swirski, M., & Valen, E. (2021). ORFik: a comprehensive R toolkit for the analysis of translation. *BMC Bioinformatics, 22*(1), 336. doi:10.1186/s12859-021-04254-w

Tsai, C. J., Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M. M., & Nussinov, R. (2008). Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol, 383*(2), 281-291. doi:10.1016/j.jmb.2008.08.012

Tunney, R., McGlincy, N. J., Graham, M. E., Naddaf, N., Pachter, L., & Lareau, L. F. (2018). Accurate design of translational output by a neural network model of ribosome distribution. *Nat Struct Mol Biol, 25*(7), 577-582. doi:10.1038/s41594-018-0080-2

Turner, M., Hu, F., Lu, J., Matheson, L. S., Diaz-Munoz, M. D., & Saveliev, A. (2021). ORFLine: a bioinformatic pipeline to prioritise small open reading frames identifies candidate secreted small proteins from lymphocytes. *Bioinformatics*. doi:10.1093/bioinformatics/btab339

Tzani, I., Ivanov, I. P., Andreev, D. E., Dmitriev, R. I., Dean, K. A., Baranov, P. V., . . . Loughran, G. (2016). Systematic analysis of the PTEN 5' leader identifies a major AUU initiated proteoform. *Open Biol, 6*(5). doi:10.1098/rsob.150203

Van Damme, P., Gawron, D., Van Criekinge, W., & Menschaert, G. (2014). N-terminal proteomics and ribosome profiling provide a comprehensive view of the alternative translation initiation landscape in mice and men. *Mol Cell Proteomics, 13*(5), 1245-1261. doi:10.1074/mcp.M113.036442

Vanderperre, B., Lucier, J. F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., . . . Roucou, X. (2013). Direct detection of alternative open reading frames translation products in human significantly expands the proteome. *PLoS One, 8*(8), e70698. doi:10.1371/journal.pone.0070698

Verbruggen, S., & Menschaert, G. (2018). mQC: A post-mapping data exploration tool for ribosome profiling. *Comput Methods Programs Biomed*. doi:10.1016/j.cmpb.2018.10.018

Verbruggen, S., Ndah, E., Van Criekinge, W., Gessulat, S., Kuster, B., Wilhelm, M., . . . Menschaert, G. (2019). PROTEOFORMER 2.0: further developments in the ribosome profiling-assisted proteogenomic hunt for new proteoforms. *Mol Cell Proteomics, 18*, S126-S140. doi:10.1074/mcp.RA118.001218

Wang, H., McManus, J., & Kingsford, C. (2016). Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. *Bioinformatics, 32*(12), 1880-1882. doi:10.1093/bioinformatics/btw085

Wang, H., McManus, J., & Kingsford, C. (2017). Accurate Recovery of Ribosome Positions Reveals Slow Translation of Wobble-Pairing Codons in Yeast. *J Comput Biol, 24*(6), 486-500. doi:10.1089/cmb.2016.0147

Wang, H., Yang, L., Wang, Y., Chen, L., Li, H., & Xie, Z. (2018). RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res, 47*, 230–234. doi:10.1093/nar/gky978

Weatheritt, R. J., Sterne-Weiler, T., & Blencowe, B. J. (2016). The ribosome-engaged landscape of alternative splicing. *Nat Struct Mol Biol, 23*(12), 1117-1123. doi:10.1038/nsmb.3317

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res, 7*, 1338. doi:10.12688/f1000research.15931.2

Woolstenhulme, C. J., Guydosh, N. R., Green, R., & Buskirk, A. R. (2015). High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep, 11*(1), 13-21. doi:10.1016/j.celrep.2015.03.014

Wu, C. C., Zinshteyn, B., Wehner, K. A., & Green, R. (2019). High-Resolution Ribosome Profiling Defines Discrete Ribosome Elongation States and Translational Regulation during Cellular Stress. *Mol Cell*. doi:10.1016/j.molcel.2018.12.009

Wu, W. S., Jiang, Y. X., Chang, J. W., Chu, Y. H., Chiu, Y. H., Tsao, Y. H., . . . Tseng, J. T. (2018). HRPDviewer: human ribosome profiling data viewer. *Database (Oxford), 2018*. doi:10.1093/database/bay074

Wu, W. S., Tsao, Y. H., Shiue, S. C., Chen, T. Y., Tseng, Y. Y., & Tseng, J. T. (2021). A tool for analyzing and visualizing ribo-seq data at the isoform level. *BMC Bioinformatics, 22*(Suppl 10), 271. doi:10.1186/s12859-021-04192-7

Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., & Yang, X. (2018). De novo annotation and characterization of the translatome with ribosome profiling data. *Nucleic Acids Res, 46*(10), e61. doi:10.1093/nar/gky179

Xiao, Z., Zou, Q., Liu, Y., & Yang, X. (2016). Genome-wide assessment of differential translations with ribosome profiling data. *Nat Commun, 7*, 11194. doi:10.1038/ncomms11194

Xie, C., Bekpen, C., Kunzel, S., Keshavarz, M., Krebs-Wheaton, R., Skrabar, N., . . . Tautz, D. (2019). A de novo evolved gene in the house mouse regulates female pregnancy cycles. *Elife, 8*. doi:10.7554/eLife.44392

Xie, S. Q., Nie, P., Wang, Y., Wang, H., Li, H., Yang, Z., . . . Xie, Z. (2016). RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res, 44*(D1), D254-258. doi:10.1093/nar/gkv972

Xu, Z., Hu, L., Shi, B., Geng, S., Xu, L., Wang, D., & Lu, Z. J. (2018). Ribosome elongating footprints denoised by wavelet transform comprehensively characterize dynamic cellular translation events. *Nucleic Acids Res, 46*(18), e109. doi:10.1093/nar/gky533

Yordanova, M. M., Loughran, G., Zhdanov, A. V., Mariotti, M., Kiniry, S. J., O'Connor, P. B. F., . . . Baranov, P. V. (2018). AMD1 mRNA employs ribosome stalling as a mechanism for molecular memory formation. *Nature, 553*(7688), 356-360. doi:10.1038/nature25174

Zafrir, Z., Zur, H., & Tuller, T. (2016). Selection for reduced translation costs at the intronic 5' end in fungi. *DNA Res, 23*(4), 377-394. doi:10.1093/dnares/dsw019

Zhang, P., He, D., Xu, Y., Hou, J., Pan, B. F., Wang, Y., . . . Chen, Y. (2017). Genome-wide identification and differential analysis of translational initiation. *Nat Commun, 8*(1), 1749. doi:10.1038/s41467-017-01981-8

Zhang, S., Hu, H., Zhou, J., He, X., Jiang, T., & Zeng, J. (2017). Analysis of Ribosome Stalling and Translation Elongation Dynamics by Deep Learning. *Cell Syst, 5*(3), 212-220 e216. doi:10.1016/j.cels.2017.08.004

Zhao, D., Baez, W., Fredrick, K., & Bundschuh, R. (2018). RiboProP: A Probabilistic Ribosome Positioning Algorithm for Ribosome Profiling. *Bioinformatics*. doi:10.1093/bioinformatics/bty854

Zhong, Y., Karaletsos, T., Drewe, P., Sreedharan, V. T., Kuo, D., Singh, K., . . . Ratsch, G. (2017). RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics, 33*(1), 139-141. doi:10.1093/bioinformatics/btw585