HUMBOLDT-UNIVERSITÄT ZU BERLIN

Detecting and quantifying the translated transcriptome with Ribo-seq data

Dissertation

zur Erlangung des akademischen Grades
Ph.D.
bzw. Doctor of Philosophy
im Fach Biologie
eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
von
MSc Lorenzo Calviello

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst
Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen:

1 – Prof. Dr. Uwe Ohler
2 – Prof. Dr. Nils Blüthgen
3 – Prof. Dr. Markus Landthaler

Tag der mündlichen Prüfung 08.11.2017

Contents

List of main figures

List of main tables

List of abbreviations

7mG	7-methyl-Guanosine.
APPRIS	Annotation of Principal Isoforms
ATP	Adenosine Triphosphate
AUC	Area Under the ROC Curve
CCDS	Consensus CDS
CDS	Coding Sequence
CHX	Cycloheximide
DFT	Discrete Fourier Transform
DNA	Deoxyribonucleic Acid
ER	Endoplasmic Reticulum
FDR	False Discovery Rate
GTP	Guanosine Triphosphate
HMM	Hidden Markov Model
HPLC	High-Performance Liquid Chromatography
IP	Immuno-Precipitation
IRES	Internal Ribosomal Entry Site
LTM	Lactimidomycin
MS-MS	Tandem Mass Spectrometry
NGS	Next-Generation Sequencing
NMD	Nonsense-Mediated Decay
ORF	Open Reading Frame
P-site	Peptidyl-site
PAR-CLIP	Photoactivatable Ribonucleoside-enhanced Crosslinking and IP
PARS	Parallel Analysis of RNA Structure
PCR	Polymerase Chain Reaction
PSD	Power Spectral Density
PTC	Premature Stop Codon
PTM	Post-Translational Modification
RBP	RNA-Binding Protein
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
RPF	Ribosome-Protected Fragment
RPKM	Reads per Kilobase of exon per Million reads
RT	Reverse Transcription
SILAC	Stable Isotope Labeling with Amino acids in Cell culture
SNP	Single-Nucleotide Polymorphism
TSS	Transcription Start Site
UMI	Unique Molecular Identifier
UTR	Untranslated Region
bp	base pair
iBAQ	intensity-Based Absolute Quantification
lncRNA	long non-coding RNA
mRNA	messenger RNA
miRNA	microRNA
ncRNA	non-coding RNA
nt	nucleotide
rRNA	ribosomal RNA
tRNA	transfer RNA
uORF	upstream ORF

Erklärung

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswis-senschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

Berlin, den

Lorenzo Calviello

Abstract

The study of post-transcriptional gene regulation requires in-depth knowledge of multiple molecular processes acting on RNA, from its nuclear processing to translation and decay in the cytoplasm. With the advent of RNA-seq technologies we can now follow each of these steps with high throughput and resolution.

Ribosome profiling (Ribo-seq) is a popular RNA-seq technique, which aims at monitoring the precise positions of millions of translating ribosomes, proving to be an essential tool in studying gene regulation. However, the interpretation of Ribo-seq profiles over the transcriptome is challenging, due to noisy data and to our incomplete knowledge of the translated transcriptome.

In this Thesis, I present a strategy to detect translated regions from Ribo-seq data, using a spectral analysis approach aimed at detecting ribosomal translocation over the translated regions. The high sensitivity and specificity of our approach enabled us to draw a comprehensive map of translation over the human and *Arabidopsis thaliana* transcriptomes, uncovering the presence of known and novel translated regions. Evolutionary conservation analysis, together with large-scale proteomics evidence, provided insights on their functions, between the synthesis of previously unknown proteins to other possible regulatory roles. Moreover, quantification of Ribo-seq signal over annotated transcript structures exposed translation of multiple transcripts per gene, revealing the link between translation and RNA-surveillance mechanisms. Together with a comparison of different Ribo-seq datasets in human cells and in *Arabidopsis thaliana*, this work comprises a set of analysis strategies for Ribo-seq data, as a window into the manifold functions of the expressed transcriptome.

Keywords: Ribo-seq, translation, transcriptomics, proteomics, bioinformatics, spectral analysis.

Zusammenfassung

Die Untersuchung der posttranskriptionellen Genregulation erfordert eine eingehende Kenntnis vieler molekularer Prozesse, die auf RNA wirken, von der Prozessierung im Nukleus bis zur Translation und der Degradation im Zytoplasma. Mit dem Aufkommen von RNA-seq-Technologien können wir nun jeden dieser Schritte mit hohem Durchsatz und Auflösung verfolgen.

Ribosome Profiling (Ribo-seq) ist eine RNA-seq-Technik, die darauf abzielt, die präzise Position von Millionen translatierender Ribosomen zu detektieren, was sich als ein wesentliches Instrument für die Untersuchung der Genregulation erweist. Allerdings ist die Interpretation von Ribo-seq-Profilen über das Transkriptom aufgrund der verrauschten Daten und unserer unvollständigen Kenntnis des translatierten Transkriptoms eine Herausforderung.

In dieser Arbeit präsentiere ich eine Methode, um translatierte Regionen in Ribo-seq-Daten zu erkennen, wobei ein Spektralanalyse verwendet wird, die darauf abzielt, die ribosomale Translokation über die übersetzten Regionen zu erkennen. Die hohe Sensibilität und Spezifität unseres Ansatzes ermöglichten es uns, eine umfassende Darstellung der Translation über das menschlichen und pflanzlichen (*Arabidopsis thaliana*) Transkriptom zu zeichnen und die Anwesenheit bekannter und neu-identifizierter translatierter Regionen aufzudecken. Evolutionäre Konservierungsanalysen zusammen mit Hinweisen auf Proteinebene lieferten Einblicke in ihre Funktionen, von der Synthese von bisher unbekannter Proteinen einerseits, zu möglichen regulatorischen Rollen andererseits. Darüber hinaus zeigte die Quantifizierung des Ribo-seq-Signals über annotierte Genemodelle die Translation mehrerer Transkripte pro Gen, was die Verbindung zwischen Translations- und RNA-Überwachungsmechanismen offenbarte. Zusammen mit einem Vergleich verschiedener Ribo-seq-Datensätze in menschlichen und planzlichen Zellen umfasst diese Arbeit eine Reihe von Analysestrategien für Ribo-seq-Daten als Fenster in die vielfältigen Funktionen des exprimierten Transkriptoms.

Schlüsselwörter: Ribo-seq, Translation, Transkriptomik, Proteomik, Bioinformatik, Spektralanalyse.

1 Introduction

All the processes that occur inside a cell are the result of millions of interactions between different molecular complexes. Gene expression regulation ensures that such interactions happen in a precise and timely manner, modulating the flow of genetic information at distinct steps.

Gene expression is a cascade of different steps, where genes are first *transcribed* into RNA molecules in the nucleus. RNAs represent a class of highly heterogeneous molecules, with high diversity even when transcribed from a single gene. The functions of RNAs are many, but perhaps their most important one is enable the production of proteins in the cytoplasm, during a process named *translation*. The functions of RNAs are mostly predicted from their sequences, e.g. whether they seem to encode a protein product or not. Such predictions alone are often used to infer whether RNAs in the cell undergo translation into proteins. However, the actual translation status of thousands of RNAs is difficult to monitor, and, in many cases, the protein-coding abilities of thousands of RNAs are unknown.

Next Generation Sequencing technologies allows for detection and quantification of nucleic acids like DNA and RNA, allowing us to fill the gap between gene sequences and their biological functions. By employing RNA isolation coupled to sequencing (RNA-seq), it is possible to interrogate different aspects of the RNA life cycle, from transcription to post-transcriptional aspects of gene expression.

Computational analysis of RNA-seq data provides identification and quantification of RNAs in our samples, allowing us to investigate their biological functions and their dynamics in different conditions. As many variations of the RNA-seq protocol exist, tailored analysis strategies must be applied to extract meaningful information from the data, with a variety of analysis tools being developed for different experimental protocols.

Thanks to the development of a new protocol, named Ribosome Profiling, we can now monitor translation at high resolution for thousands of RNA molecules, potentially revealing the protein-coding ability of entire transcriptomes. Understanding the technology and the analysis strategies required is thus key to extrapolate meaningful results on a genome-wide scale. Moreover, precise information about the translation status of different RNAs can complement the information coming from other RNA-seq protocols, allowing for integration of multiple data sources for a more complete understanding of the expressed transcriptome.

1.1 Thesis outline

In Section 2.1 I will give a brief introduction to the molecular basis of RNA biology, highlighting the main steps in the gene regulatory cascade, with an emphasis on translation. A survey on the main methods used to interrogate the translation status of the transcriptome is presented in Section 2.2, with a detailed analysis on the analysis strategies in Section 2.3. Emphasis on the Ribo-seq protocol and data analysis is presented in these two sections.

Section 3.1 describes our interdisciplinary approach to detect translation in Ribo-seq data, while its application on new data from a human cell line appears in Section 3.2. Section 3.3 deals with the application of our strategy in new data coming from the plant *Arabidopsis Thaliana*, together with a comparison with different available Ribo-seq datasets. Work in progress is described in Section 3.4, where we extend our strategy to detect and quantify translation for different RNAs coming from a single gene. Finally, our results are discussed in Section 4, accompanied by a list of references in Section 5 and few additional material in the Appendix sections.

2 Background

2.1 The Molecular Biology of RNA processing

2.1.1 Life and the central dogma

Distinguishing living organisms from inanimate matter is a non-trivial task, especially when thinking about a pathogenic virus, or a robot able to learn from the environment. However, when zooming at the molecular and cellular level, we can appreciate some common aspects of living organisms. For instance, structural characteristics (the presence of a cell membrane), or phenomenological behaviors (metabolism, cell division) are common in all the life kingdoms, and thus enable us to define some fundamental properties which can aid us defining a living entity[1]:

1) Compartmentalization, the definition of boundaries between the organism and the environment
2) Autopoiesis, the ability to self-sustain
3) Reproduction, the capacity to produce new living organisms

The first property derives from the definition of a minimal unit, the cell, as the universal building block of all life. The second encompasses a plethora of different metabolic processes and their regulation, which enables cell growth and adaption to environmental stimuli. Reproduction provides new living entities of the same organisms (the *offspring*), which also inherit the ability to reproduce themselves. As the ability to reproduce is inheritable, life safeguards its continuity, and allows for the emergence of more complex processes in time, like adaptation of a population of organisms to changes in the environment.

While compartmentalization might be achieved via spontaneous aggregation of lipid molecules (which can be provided by the environment[2]), autopoiesis and reproduction are the result of the complex interactions of molecular entities within the cell. The ability to internally synthesize some of the necessary molecules needed for different biological processes defines the cell as a semi-independent entity. However, for a specific biosynthetic reaction to occur, cells require the presence of a template containing all the necessary information (called *genetic* *information*), and a machinery able to read and put into action the information within. This concept can be extended from cells to entire organisms: for example, during organismal development in vertebrates, a single fertilized embryo can give rise to a complex organism, with many specialized cells and tissues. All the information and the molecules who can read this information and translate it into dynamic molecular interactions are present into a single cell, and must be inherited by the offspring to continue the cycle of life.

With the first observations of chromosomal structure made of DNA and proteins inside cells, biologists started investigating whether DNA is a suitable carrier of heritable information. In the 1940s, experiments carried on different *Streptococcus* strains showed how DNA is the only molecule able to transform different strain of bacteria into each other, thus conferring cell identity, while molecules like proteins or other metabolites could not[3]. The main blocks defining the network flow of genetic information was then defined: DNA was found to produce an analog molecule, RNA, during a process called *Transcription*, while RNAs (or *transcripts*) are used to synthesize proteins, during the *Translation* process.

Thanks to the discovery of the double helix structure from Watson and Crick[4], DNA properties were being discovered, pointing at its ability to carry genetic information but also at its replicative nature, which can ensure the hereditary nature of life. As the chemistry of nucleotides (building blocks of DNA and RNA) and amino acids (the building blocks of proteins) had already been characterized before the 1950s, theories linking the genetic information in DNA and the composition of synthesized proteins were being suggested. One theory described how a triplet of nucleotides (a *codon*) can specify one amino acid, hinting at the presence of a *genetic code* linking a DNA sequence to an aminoacidic sequence[5]. In a famous publication in 1961 by Crick *et al*, [6], a bacteriophage T4 DNA sequence was mutated in different positions, and the authors observed how deletion or insertion of nucleotides in the DNA sequence were disrupting the coding information, as they resulted in a non-functional protein product; deletions or insertions of 3 bases in DNA were not causing the production of an aberrant protein, confirming the triplet-nucleotide pattern in the genetic code. In parallel, it was shown by Nierenberg and Matthaei how a RNA molecule with a poly-U sequence would produce a phenylalanine polypeptide, suggesting that UUU is the codon encoding for phenylalanine[7].

Additional experiments then completed the map of the *genetic code* (Figure 1), assigning to each codon its corresponding amino acid. Following studies then helped understanding how this information is parsed during the Translation process, also thanks to the discovery of the tRNA structure by Holley[8]. As shown in Figure 1, the genetic code is *degenerate*, as multiple (*synonymous*) codons correspond to the same amino-acid. Synonymous codons only differ in the third codon position: this phenomenon is linked to the physical interaction between each codon and its corresponding tRNA sequence (the *anticodon*) during the elongation process (Section 2.1.5), where the third position (also called the *wobble* position) has a lower interaction energy, and thus lower importance in defining the genetic code.

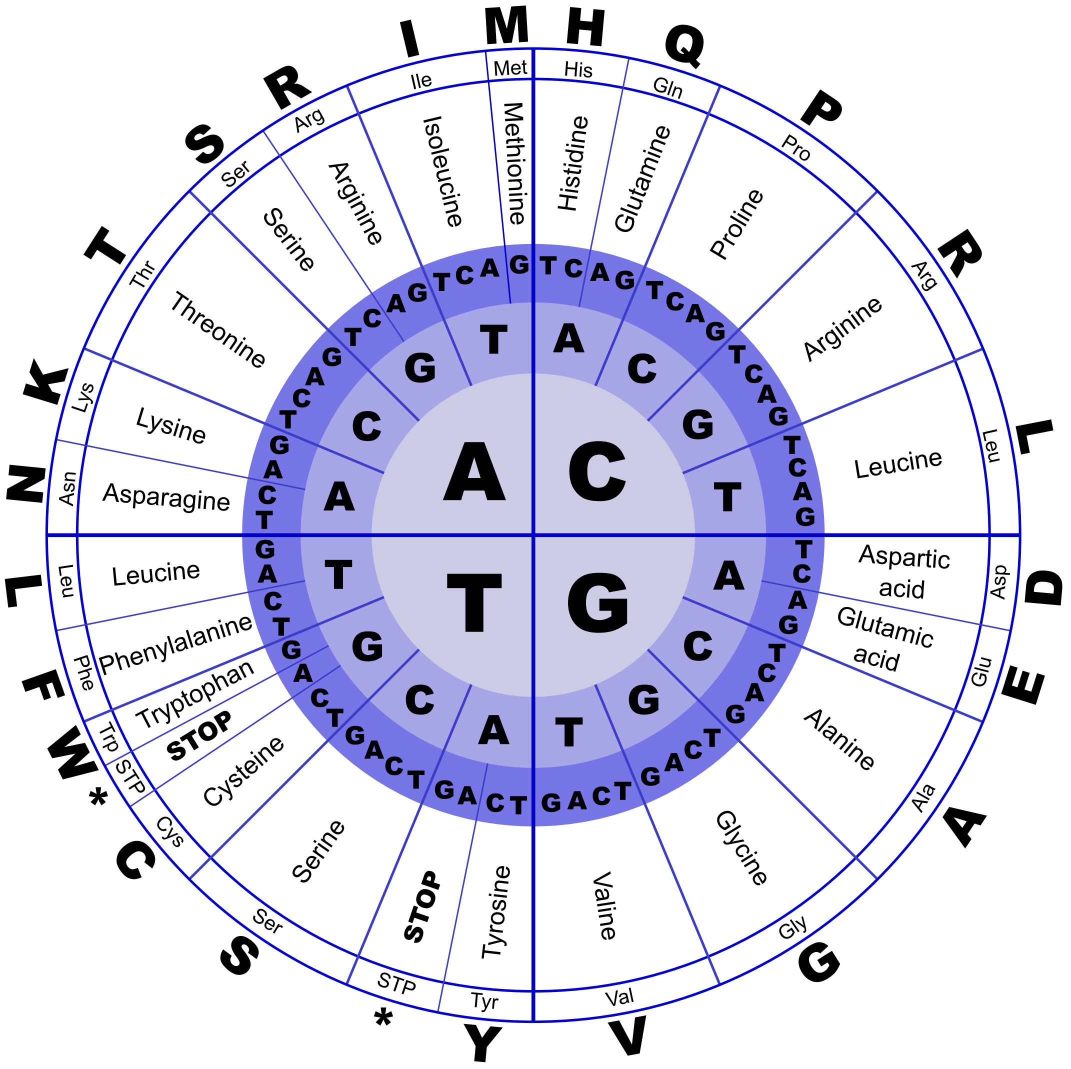


Figure 1: The genetic code. Starting from the center, a triplet of nucleotides specifies an amino acid or a stop codon. Author: J. Alves, Creative Commons Zero 1.0 License.

The presence of a genetic code to translate RNA sequences into protein proved to be a property present in all organisms in the different kingdoms of life, with minor differences between different organisms. The universal nature of these discoveries led to the formulation of the so-called Central Dogma of molecular Biology by Crick [9], [10], which states the directional flow of information from DNA to Protein (Figure 2).

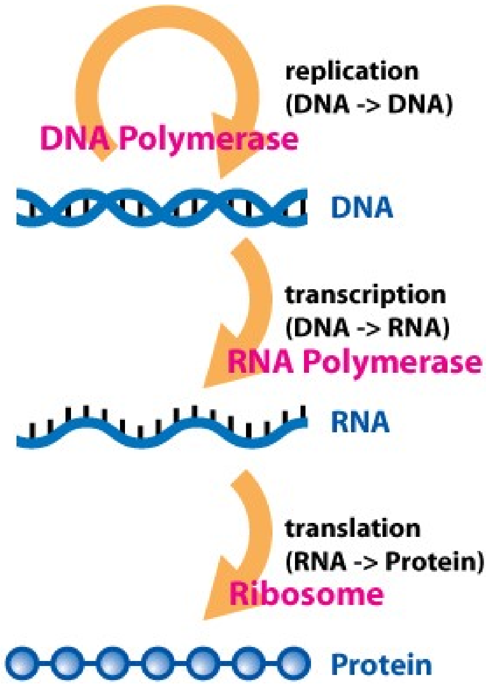


Figure 2: The central dogma and its main molecular actors. From DNA, the genetic information is replicated (via the DNA polymerase), or transferred to RNA (during transcription with RNA Polymerase) and proteins (during translation with the Ribosome). Source=http://en.wikipedia.org/wiki/File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg , GPL license 1.2+, Author=Dhorspool.

The formulation of the Central Dogma posed the molecular basis to understand the link between the information encoded in the DNA (the genotype) and its biological manifestation (the phenotype). As shown in Figure 2, multiple steps are required to de-code the information present in the complete set of DNA sequences (the *genome*) to produce RNA and proteins. The study of the regulation happening at each one of these steps represents a wide area of active research. However, to approach the study of gene regulation in a human cell we must first appreciate the complexity of the human genome and the staggering diversity of its molecular product.

2.1.2 A multitude of RNA species

Genomes may wildly differ between organisms, both in terms of size and composition. A common feature of mammalian genomes is the low percentage of DNA sequence containing coding genes, defined as DNA loci encoding for functional proteins. It is estimated that half or more of the human genome is made of repetitive sequences, which mostly do not code for protein[11]. Only around 1% of the human genome (which is ~3.5 billion base pairs) is made of coding sequences, and ~20.000 human genes encode for distinct protein product (estimates from GENCODE annotation, version 19). However, despite their great diversity and importance in all cell types known, protein coding genes represent only a fraction of the total number of human genes. An increasing number of non-coding genes with many different functions permeates the human genome. Of those, many exert well studied regulatory functions via their short RNA products (>200 nt, Figure 3), while others are currently being investigated by the research community. Genes can be divided in different categories, mostly based on the function of their RNA product (Figure 3, definitions the GENCODE[12] annotation, version 19):

Protein-coding – encode for proteins;
rRNAs – ribosomal RNA, component of the ribosome;
tRNAs – transfer RNA, involved in protein synthesis (see Section 2.1.5);
snRNAs – small nuclear RNA, involved in pre-mRNA splicing;
snoRNAs – small nucleolar RNA, involved in rRNA, tRNA and snRNA processing;
miRNAs – microRNAs, small RNAs involved in regulating RNA stability and translation;
lincRNAs – long (>200 bp) intergenic non-coding RNAs, involved in many regulatory processes (unknown for most of them);
processed transcripts – non-coding RNAs who do not contain an ORF;
antisense RNAs – RNAs overlapping with protein-coding genes but on the opposite genomic strand;
pseudogenes – transcripts with strong sequence similarity to other known genes but often with a disrupted coding sequence; can be derived from gene duplication or mature RNA retro-transposition (*processed pseudogenes*);

Other less characterized classes may be defined for other RNAs based on their genomic position (e.g. sense intronic RNA, 3’overlapping RNA), or not well understood structural properties (e.g. vault RNA), or involvement in small RNA metabolism (e.g. scaRNA).

According to the GENCODE annotation, in >87% of the cases, the transcription of human genes creates an RNA product whose final form contains only some section of the original transcript.

From the full-length transcribed RNA (the *pre-mRNA*), short sequences called *exons* are joined to form a mature transcript, while longer sequences between exons (the *introns*) are removed, during a process named *splicing*. Splicing can happen for any transcribed gene, but mostly happens in protein-coding genes, as >94% of coding transcripts are spliced. For mature protein coding transcripts, three distinct elements can be further defined based on their coding capability: the first element is a 5'UTR (Untranslated Region), then a CDS (Coding Sequence, which contains the translated RNA sequence, known as the ORF, or Open Reading Frame), and a 3'UTR. The length of these regions varies from transcript to transcript. Overall, 5'UTR are around 300 nt long, while CDS and 3'UTR are longer (Figure 3).

As specified above, splicing joins exonic sequences from a pre-mRNA molecule. However, splicing can join different exons combinations from the same pre-mRNA molecule, taking the definition of *alternative splicing*. Different transcripts (or *isoforms*) coming from a single gene can undergo different processing fates (Section 2.1.4): transcript isoforms can code for different proteins, or represent non-coding variants of a protein coding transcript, and be subjected to different localization, translation, or decay processes. As expected, the number of possible mature transcript structures grows exponentially with the number of exons, thus greatly increasing the complexity of the human transcriptome. The number of exons varies for different transcripts, with 4 being the median of exons per transcripts; some genes, like the *TTN* gene, contain transcripts formed by more than 350 exons.

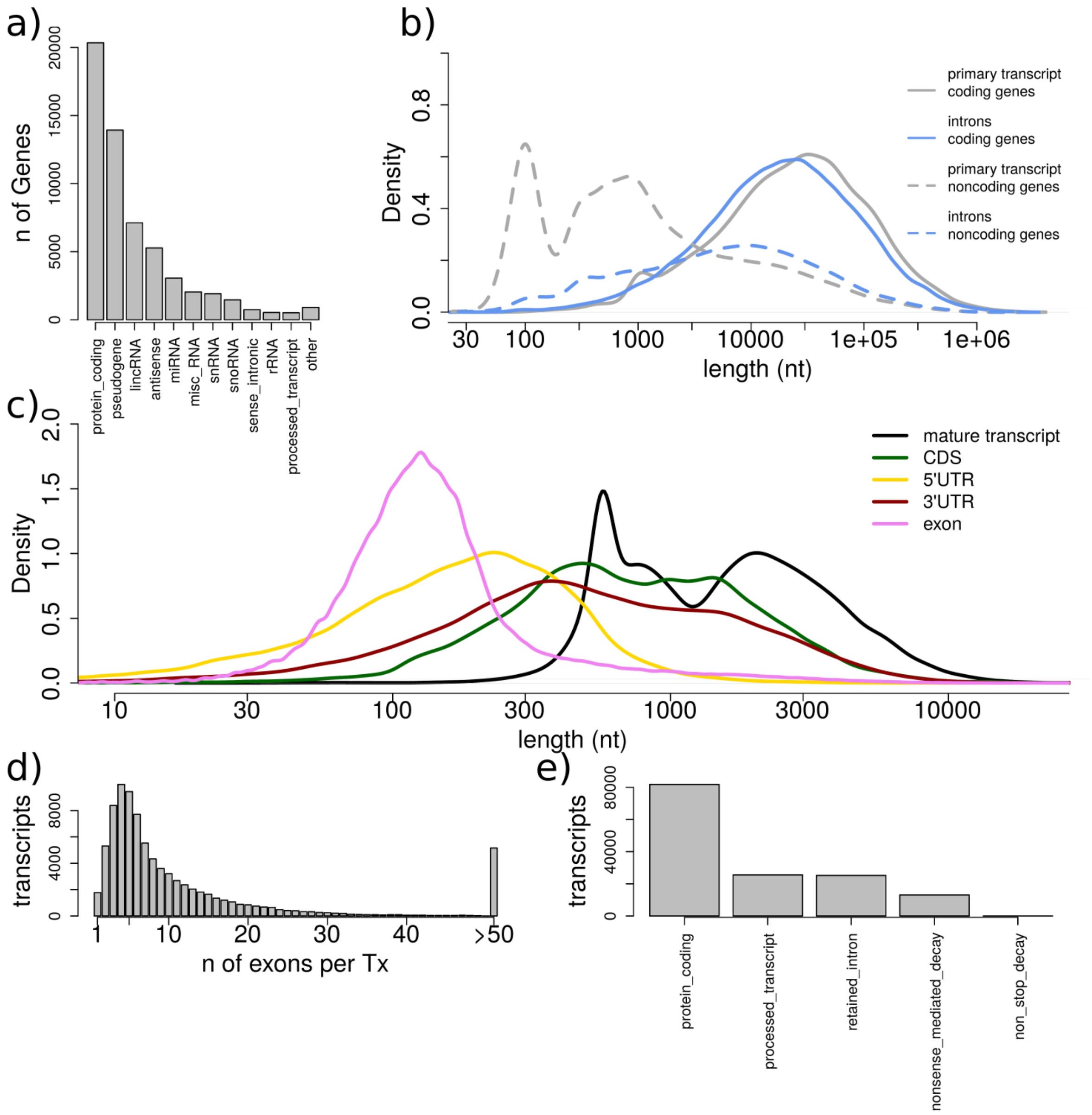


Figure 3: An overview of the human transcriptome. a) Different gene biotypes assigned to the known genes. Only protein-coding genes have annotated CDS. b) Length distribution of introns and gene lengths, for coding and non-coding genes. c) Length distribution of UTRs and CDS in protein-coding genes, compared to exon lengths. d) Number of exons per annotated coding transcripts. e) Different transcript biotypes in protein-coding genes. All presented data comes from GENCODE annotation, version 19.

Such complexity created by genes and transcript diversity enables the specification of multiple cells and tissues. In fact, in a given cell, only a subset of genes is actively expressed. Some genes (named *housekeeping* genes) are constitutively expressed across different tissues, as they encode for proteins fundamental to the basal cell metabolism. Moreover, genes are expressed at different quantitative levels in different cell types: genes encoding for a skeletal muscle protein, such as the Dystrophin, is highly expressed (together with its regulators) in skeletal muscle cells, while a gene encoding a synaptic protein is highly expressed in neurons. Different biological conditions, such as oxidative stress conditions or a differentiation process, can display specific gene expression profiles. Moreover, additional specificity is granted by the specific expression of different RNA isoforms[13], which further increases the level of specification achieved by gene expression regulatory mechanisms.

The incredible RNA diversity from both coding and non-coding genes is mirrored by their complex molecular life, from biogenesis to translation and decay, where different categories can undergo different processing steps, from the nucleus to the cytoplasm.

2.1.3 Nuclear processing

The DNA molecule is bound by DNA binding proteins in a complex called chromatin. For DNA to be transcribed, the chromatin complex must be opened, allowing the RNA polymerase (together with different cofactors) to start pre-mRNA synthesis. The exact position where the RNA polymerase starts transcribing (the Transcription Start Site, or TSS) can also vary, creating different possible 5’ ends of a transcript[14]. The different molecular reactions involved in chromatin remodeling and transcription are tightly regulated, and their regulation (Transcriptional regulation) is a wide area of intensive study.

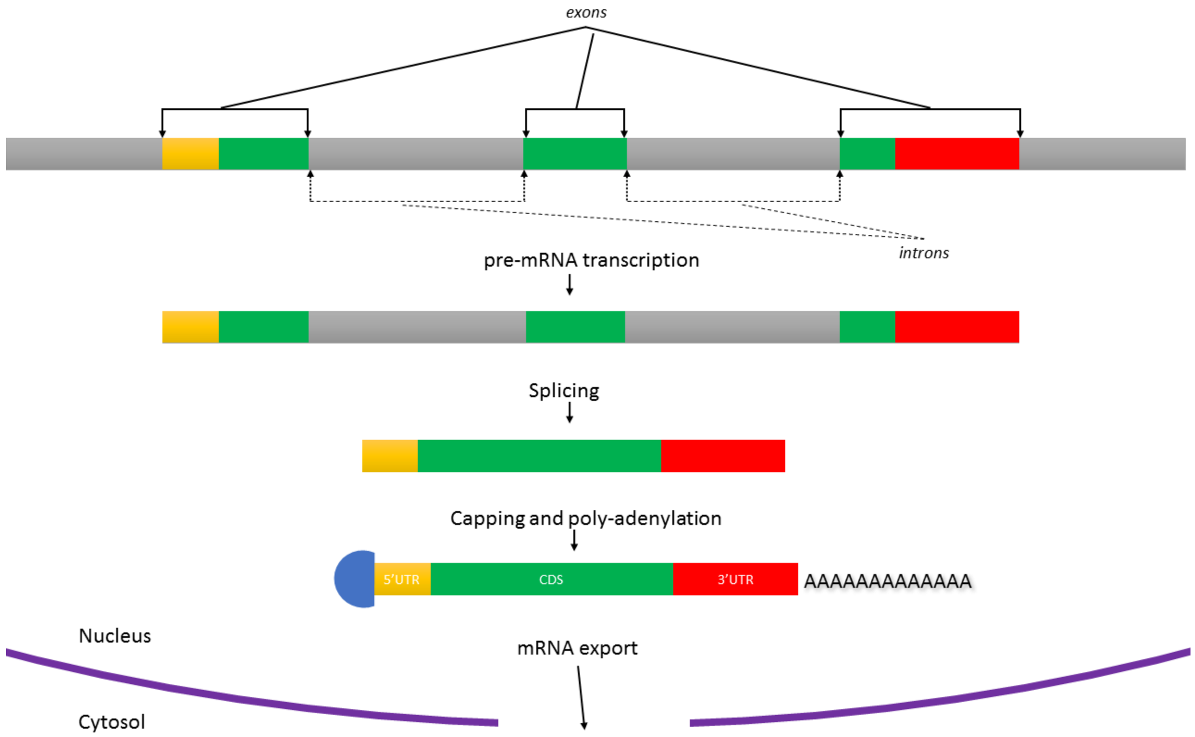


Figure 4: mRNA nuclear processing. Exons and introns of a gene are shown. After transcription, a pre-mRNA molecule is produced; intronic sequences are removed during splicing; after the addition of a cap and a poly-A tail, the mature transcript is exported to the cytoplasm.

Already during its synthesis, a nascent RNA molecule is bound by RNA-binding proteins (RBPs) who regulate subsequent processing steps, outlined in Figure 4. A 7-methyl-guanosine (7mG) "cap" at the 5'end of a transcript is added, which will facilitate translation in the cytoplasm and protect RNA from degradation (Section 2.1.4). During splicing, which also occurs largely co-transcriptionally[15], RBPs bind to exonic and intronic sequences and splice introns out of the pre-mRNA molecule, regulating the production of different exons combinations from the same original transcript. Subsequently, at the 3' end of the transcript, a stretch of Adenosines (poly-A tail) is added, which is also important in the regulation of transcript stability and translation[16]. The exact position where the poly-A tail is added also varies among tissues and condition, representing another regulatory step which creates different transcripts with different functions[17]. A nuclear RNA-surveillance pathway (the *exosome*) degrades erroneous RNA products, ensuring that transcripts are correctly processed. The kinetics of the above steps are very important, as different transcripts can processed with different efficiencies[18]. For instance, many RNA molecules can be selectively retained in the nucleus, thus limiting their export in the cytoplasm and promoting their interaction with the nuclear apparatus[19]. All the processing steps here briefly mentioned are extremely important, as they can have an impact on the downstream steps of mature RNA metabolism, in the cytoplasm.

2.1.4 The cytoplasmic fates of an RNA molecule

As the RNA, together with bound RBPs, is exported in the cytoplasm, it interacts with different protein complexes which determine its function (Figure 5).

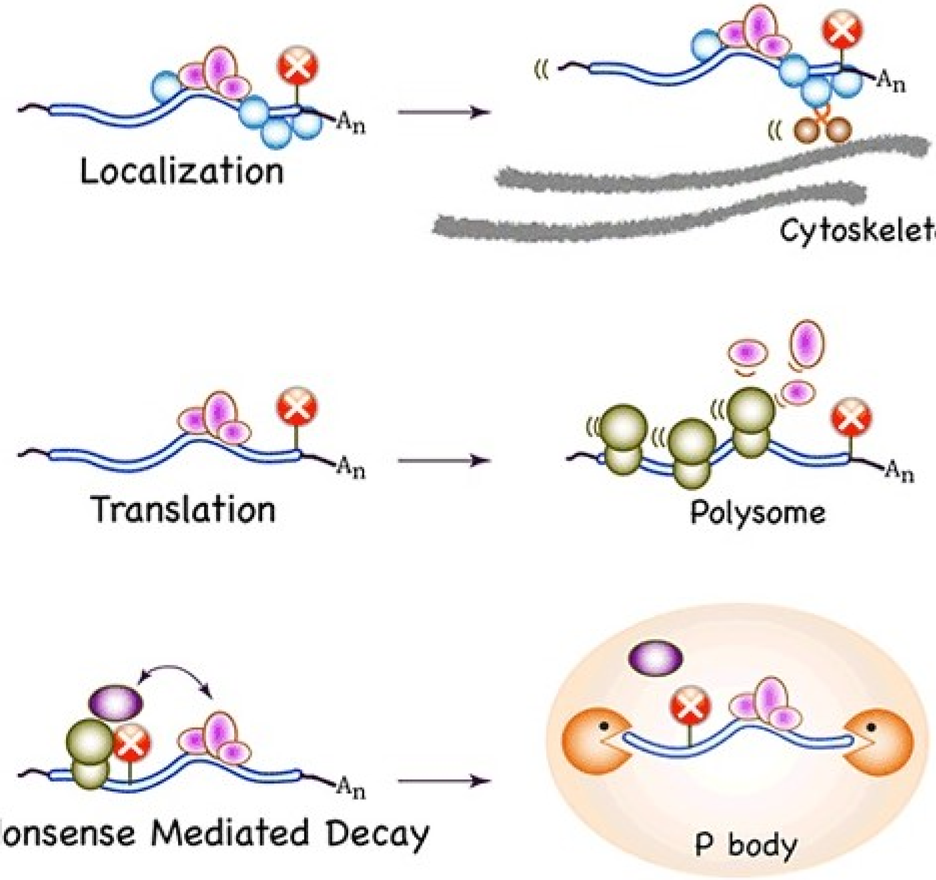


Figure 5: Different RNA cytoplasmic fates. Transcripts can be localized to different compartment (top), translated and thus form a polysome structure (middle), or degraded by different complexes like the Nonsense-Mediated Decay machinery (bottom). Both localization and degradation are linked to the translation status of a transcript (see text). Originally adapted from: http://www.hhmi.org/research/rna-processing-and-ribonucleoprotein-complexes

RNAs can be localized in different sub-cellular compartments, as the endoplasmic reticulum[20] (ER) or other locations in the cellular periphery. Based on the cell morphology and function, RNA localization can be crucial to ensure local processing of RNAs in specialized cellular compartments, like pre- or post-synaptic compartments in neuronal cell types[21]. In the oocyte in *Drosophila melanogaster,* specific RNA transcripts are recognized (also thanks to their secondary structures) and localized along the anterior-posterior and dorsal-ventral axes, where they contribute to the correct spatial patterning of the developing embryo[22].

RNA turnover in the cytoplasm represents an important layer of regulation of gene expression: many RBPs are involved in triggering transcript degradation or in promoting its stability[23]. RNA degradation can occur in specialized cytoplasmic foci, like in processing bodies (P bodies) [24]. Alterations in the RBP binding and function can promote malfunctions at the level of RNA stability, and participate in the onset of several diseases[25]. RBP binding can occur on different regions of the transcript and depend on the transcript translation state (Section 2.1.7). 5’->3’ exonucleases can degrade “decapped” RNAs from the 5’end, while 3’->5’ exonucleases act on de-adenylated RNAs, where the poly-A tail has been previously removed. For this specific degradation events to occur, the m7G cap or the poly-A tail must usually be removed from the transcript molecule, and this is often triggered by binding of miRNAs or other RBPs, often in the 3’UTRs of target transcripts[26]. Another important mode of RNA degradation is represented by the RNA-surveillance pathway, which will be examined further in Section 2.1.7.

The primary function of RNA is to engage with the ribosomal machinery to synthesize protein, and this process will require a more in-depth explanation about its single molecular steps and its relevance in the gene expression cascade.

2.1.5 The Translation process

Translation is an ancient biological process, present throughout all the three kingdoms of life. The high degree of similarity among organisms is reflected by the presence of a common catalytic machine, the ribosome. The eukaryotic ribosome is a ribozyme, as its catalytic function is carried by small RNAs and dozens of proteins, and can be divided in 2 subunits: a small subunit, also known as the 40S subunit (S stands for Svedberg, a coefficients measuring its sedimentation time during centrifugation), and a large 60S subunit, while the fully assembled complex is known as 80S. Three additional structures can be identified inside the ribosome, named Aminoacyl-site (A-site), Peptidyl-site (P-site), and the Exit site (E-site). These different sub-ribosomal structures are binding pockets for tRNAs, small non-coding RNA molecules able to carry amino acids, which play a fundamental role in the different steps of translation (Figure 6).

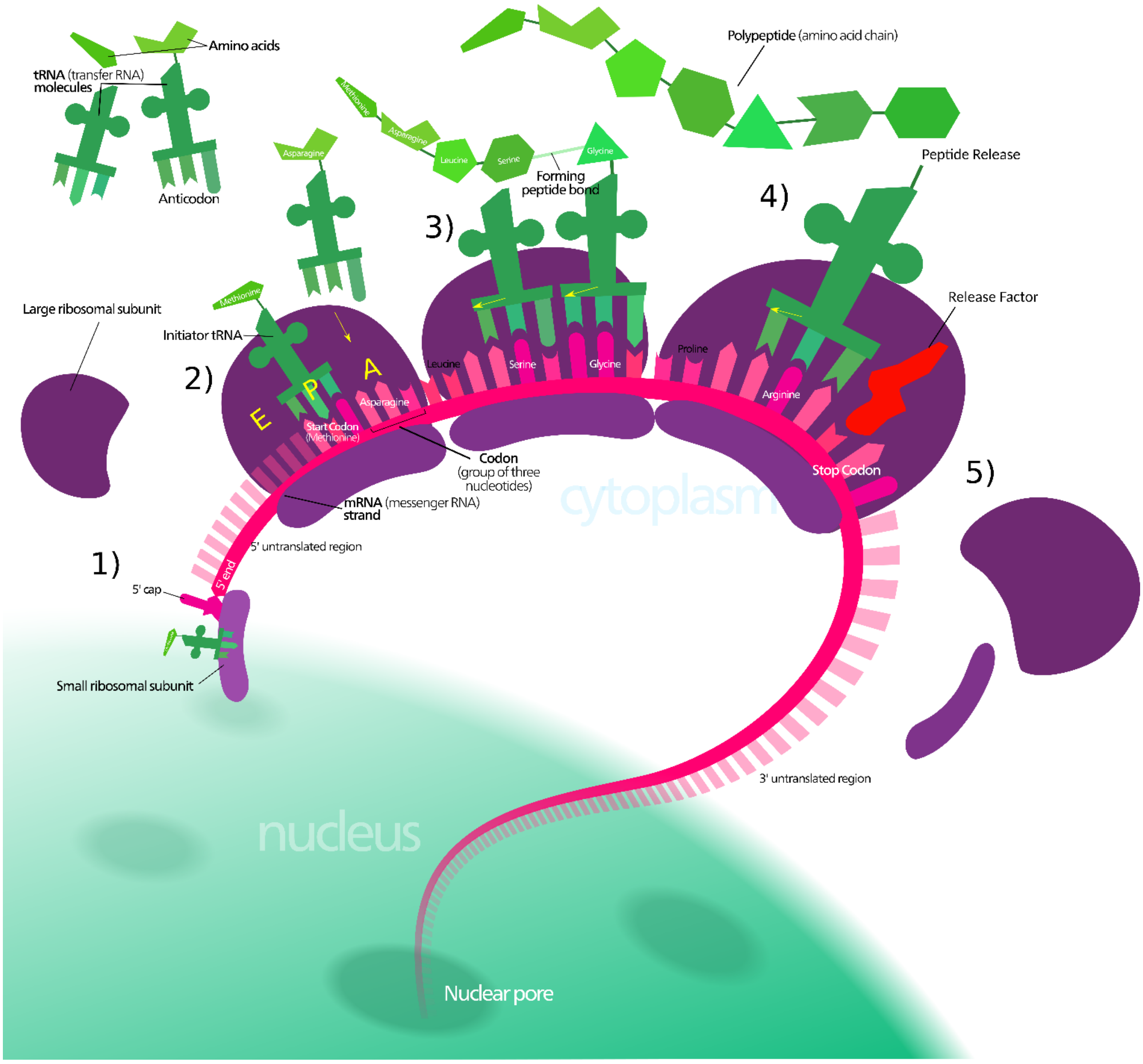


Figure 6: The main steps of the translation process. 1) Cap recognition: the pre-initiation complex binds to the cap. 2) 80S Assembly: As the start codon is recognized the 80S is assembled. Elongation begins and a loaded tRNA binds in the A-site. 3) Elongation: a peptide bond is formed on the nascent chain in the A-site, and the ribosome moves one codon towards the 3'end. As the ribosome translocates, the empty tRNA goes in the E-site, the tRNA with the nascent peptide moves to the P-site, and a vacant A-site can accept a new loaded tRNA. 4) Termination: when the ribosome hits a stop codon, a release factor binds in the A-site. The polypeptide chain is released, the empty tRNA moves on the E-site and 5) the 80S disassembles. Adapted from https://en.wikipedia.org/wiki/File:Protein_synthesis.svg, Author: Kelvinsong, License Creative Commons Attribution 3.0 Unported.

During translation initiation, a complex of initiation factors, GTP and a special methionine-tRNA (called initiator tRNA) binds to the 40S ribosomal subunit, in what will become the P-site compartment. This pre-initiation complex is now able to recognize the cap on a transcript (favored by the presence of a poly-A tail) and start scanning the 5'UTR, looking for a start codon. Alternatively, cap-independent mechanisms of translation initiation can also occur (Section 2.1.6). If a start codon is detected in a non-favorable sequence context, the small subunit will keep scanning to the next start codon candidate, in a process known as “leaky scanning”[27]. At this point, the pre-initiation complex undergoes a conformational change, initiation factors dissociate, and the large subunit joins to form the 80S fully assembled ribosome, keeping the initiator tRNA in the P-site compartment.

The ribosome complex can now enter the translation elongation steps, where it synthesizes proteins along the ORF on the transcript, fueled by GTP hydrolysis and helped by the action of different elongation factors. At the start codon, an aminoacyl-tRNA binds to the next codon in the A-site of the ribosome; the first peptide bond is formed, and the Methionine carried by the initiator tRNA is transferred to the tRNA in the A-site. At this point the ribosome moves 3 nucleotides (1 codon) forward, shifting the empty initiator tRNA to the E-site, the tRNA with the nascent peptide to the P-site, and leaving the A-site empty (**Figure 6**). The empty tRNA exits from the ribosomal compartment, and a new aminoacyl-tRNA can bind to the A-site for a new cycle of elongation. At the end of the ORF, a release factor binds to the stop codon at the A-site, and triggers the dissociation of the full-length peptide chain, allowing for the empty ribosome to detach from the mRNA.

While translating, the ribosome machinery is tightly bound to the RNA, covering a portion of the mRNA molecule; estimates of the size of such ribosomal “footprint” were attempted in the late ‘60s using RNA fingerprinting assays[28]. The size of the ribosome footprint is usually around 29nt, but can slightly vary between organisms (and organelles[29]) and it is dependent on the precise ribosome conformation[30]. The relative position of the different ribosomal compartments can also be inferred from the footprint position[31] (Figure 7). Moreover, given the codon-by-codon movement of a translocating ribosome, it should be in theory possible to observe a 3nt shift in ribosomal protection on a translated mRNA, where the precise location of the footprint follows the translated frame.

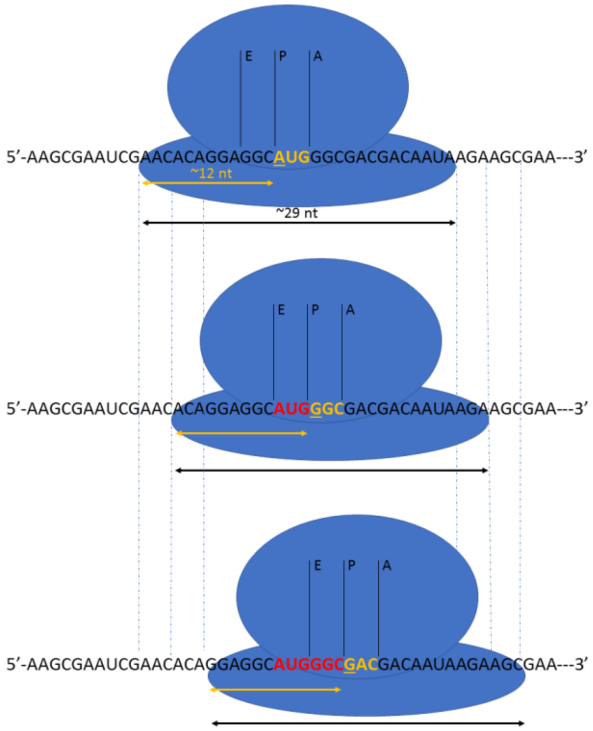

E
P
A
5'-AAGCGAAUCGAACACAGGAGGCAUGGGCGACGACAAUAAGAAGCGAA---3'
~12 nt
~29 nt
E
P
A
5'-AAGCGAAUCGAACACAGGAGGCAUGGGCGACGACAAUAAGAAGCGAA---3'
E
P
A
5'-AAGCGAAUCGAACACAGGAGGCAUGGGCGACGACAAUAAGAAGCGAA---3'

Figure 7: Ribosomal translocation. Given the ribosomal footprint location and the mRNA sequence, it is possible to identify the codon processed by the different sub-ribosomal compartments (top). During each translocation step, the ribosome moves 3nt towards the 3' end of the mRNA, and such movement is reflected in a shift of the footprint position (middle and bottom).

Multiple spaced ribosomes (a *polysome*) can simultaneously translate on a single ORF, and as their number and efficiency in translating dictate the amount of protein synthesis, fine tuning of the multiple steps in the translation cycle represents an important step in regulating the gene expression cascade[32].

2.1.6 Translation regulation

All the single steps of the translation cycle can be regulated in response to external stimuli, often through binding of regulators on the RNA molecule. Regulation at the level of initiation can happen thanks to the binding of RBPs and microRNAs[26]. In this case, RBPs can interact with other proteins who are in turn able to interact with the translation initiation complex, and thus trigger translational repression on the target mRNA[33]. The initiation rate can also be regulated by other signaling pathways: during stress conditions, the Integrated Stress Response pathway inhibits the formation of the pre-initiation complex[34], thus impeding translation initiation on thousands of transcripts. This mode of regulation can in turn promote alternative translation initiation pathways, which can also recognize different start codons[35]. The canonical ORF is defined as starting with an AUG codon, which has been shown to induce the formation of the initiation complex with high efficiency[36]. However, few examples of efficient non-AUG start codons are known in the literature and have been experimentally identified. Thanks to high-throughput techniques (Section 2.2.3), thousands of non-AUG start codons were identified and proposed as *bona-fide* translational start sites[37]. However, a global confirmation of all these non-canonical start sites is still lacking. Little is known about how different regulators can influence start codon recognition. Many genes, including *PTEN*, a famous onco-suppressor[38], can use alternative translation initiation sites, producing N-terminal extension or truncation of the original protein. Start codon recognition has important implications, as the N-terminal sequences are very important for protein localization and function[39].

Comparably little is known about regulation at the level of elongation. As shown in Section 2.1.1, the specificity for each codon is mostly depending on its first two nucleotides, given the base-pairing between codons and anticodons. Given the same tRNA, differences in the 3rd nucleotide position can slightly alter the kinetics of tRNA recognition, and thus modulate the efficiency of translation elongation[40]. Similarly, the presence of rare codons (recognized by less abundant tRNAs) and stable mRNA secondary structures have been proposed as efficient mechanisms who can “stall” elongating ribosomes[40], sometimes with effects on the nascent protein folding and stability[41]. Additional RNA structures can also lead to ribosomal frameshifting during elongation, which lead to mRNA degradation via the Nonsense-Mediated Decay (NMD) pathway[42] (see Section 2.1.7). However, despite the extensive literature on codon-mediated regulation on both translation and RNA stability, the underlying molecular mechanisms are yet to be fully elucidated, and a quantitative estimate of its impact on translation is still lacking.

Even less is known about regulation at the level of termination, despite some reports about rare events of ribosomal read-through as a possible way to modify the C-terminus of the encoded protein[43], which is also important for the protein localization and function. Interestingly, the UGA codon, normally a stop codon (Figure 1), can code for selenocysteine, an additional amino acid which is incorporated in few important mRNA, mostly coding for metabolic enzymes[44].

The importance of translation for cell survival, together with our knowledge about the structural differences between eukaryotic and prokaryotic ribosomes, allowed us to use efficient natural compounds as antibiotics against several bacterial species. Other translational inhibitors can act on eukaryotic translation, allowing us to block ribosomes at different stages of the translation cycle: Cycloheximide, for instance, can bind to the E site of the elongating ribosome, blocking the exit of an empty tRNA and thus the translocation step[45]. Other inhibitors, such as Harringtonine or Lactimidomycin (Section 2.2.3), can "lock" the ribosome in the initiation complex formation step, allowing us to study translation initiation.

In addition to *trans*-acting elements able to regulate translation, also *cis*-regulatory elements (present on the RNA transcript itself) can regulate translation. An example for *cis*-regulatory elements are small ORFs present in the 5'UTRs of transcripts, analyzed in many analyzed eukaryotic species[46]. Such upstream ORFs (uORFs) are believed to repress translation of the main ORF, as their translation reduces the number of available ribosomes for the main ORF translation[46], [47]. The putative short peptide encoded by the small uORF translation is thought to be a by-product of such regulatory event, despite some known contradictory examples[48]. It has been proposed that several thousand candidate uORFs exist and can regulate the main ORF translation in different species[46], but the actual usage of all these putative regulatory elements in different systems is still a matter of discussion.

Additional elements in the 5'UTR, called Internal Ribosomal Entry Sites (IRES), are able to bind the small ribosomal subunit which can then start scanning and synthesizing proteins, thus bypassing the recognition of the cap at the 5' of the transcript[49]. This is of great importance, especially when considering that during viral infection the cell undergoes stress, decreasing canonical cap-dependent translation and favoring the cap-independent translation of the IRES-containing viral transcripts[50].

Another extremely interesting aspect to consider is ribosome heterogeneity: ribosomes can differ between cell types, and sub-populations of ribosomes can also be distinguished within the same cell, raising the possibility that translation regulation might be a much more heterogeneous process acting on specialized ribosomes[51], [52].

As mentioned before, the binding of RBPs and miRNAs can repress translation on the target mRNAs, but also trigger de-capping and poly-adenylation of the transcript, thus triggering susceptibility to exonucleases and thus degradation. Different studies, especially focusing on miRNA-mediated regulation, tried to disentangle the differences between these two modes of regulatory action (translational repression and RNA degradation), also considering the temporal kinetics of this two processes[53], [54]. Additional regulation over transcript stability is achieved via other mechanisms, which again act on the ribosome to achieve specificity.

2.1.7 Translation and RNA decay

The connection between translation and RNA metabolism becomes even more intricate when studying the RNA surveillance pathway. Thought to be evolved to ensure the clearance of aberrant transcription and splicing events, the RNA surveillance pathway can trigger endonucleolytic cleavage and degradation on RNA molecules, and this process has been shown to be dependent on the translational status of the transcript[55], [56]. Of these pathways, the Nonsense-Mediated Decay (NMD) is one of the most studied. Several studies pointed out NMD acts when recognizing a Premature Termination Codon (PTC) as a sign of an aberrant transcript. The definition of a PTC usually includes the presence of an exon-exon junction downstream the stop codon, where specialized protein being part of the Exon Junction Complex (EJC) are binding to members of the NMD pathway (like the members of the UPF family) and can thus trigger transcript degradation. One of the proposed modes of action for NMD-mediated degradation[55] explains how the endonucleolytic cleavage (by the SMG6 protein) takes place close to the stop codon of the to-be-degraded transcript, and entails the interaction between several proteins and the terminating ribosome[55]. Additional decapping mechanisms can also be triggered, and 5’-3’ exonucleases can ultimately degrade the cleaved/decapped transcript[57]. Additional proposed mechanisms for NMD action can be independent of EJC binding, and involve the recognition of long 3’UTR sequences[58].

Additional RNA surveillance mechanisms, such as Non-stop Decay or No-Go Decay[55], also act on actively translated transcripts, pointing out again at the importance of translation in the entire cytoplasmic life of an RNA molecule.

Given the importance of translation regulation, a specific, local concentration of RBPs can strongly influence the translational output. Thus, the coupling between RNA localization (Section 2.1.4) and local translation is an important process which can ensure additional specificity in the regulation of protein synthesis. Sub-cellular compartmentalization of translation regulation is of course relevant for specialized cell types like neurons, but arguably for any other cell type, as ribosomes can translate free in the cytoplasm or on the surface of the endoplasmic reticulum (ER) [20], [59], and a different molecular environment can modulate translation in very specific ways.

In the light of the numerous mechanisms of post-transcriptional gene regulation, and their crosstalks (Section 2.1.4), one can imagine the wide range of possible regulation happening at the level of individual RNAs, where elements like uORFs, PTCs, or other elements along the transcript structure shape individual regulatory programs. As outlined above (Section 2.1.2), alternative splicing adds an additional layer of heterogeneity, where transcripts differ between each other only in some elements, while sharing most of the sequence. It has been recently shown that alternative transcript isoforms are translated[60] and can display distinct translational outputs[61], [62]. The function of alternative splicing thus not only aims at increasing proteome diversity, but also at directing gene expression towards transcripts with possibly very different functions (Figure 8).

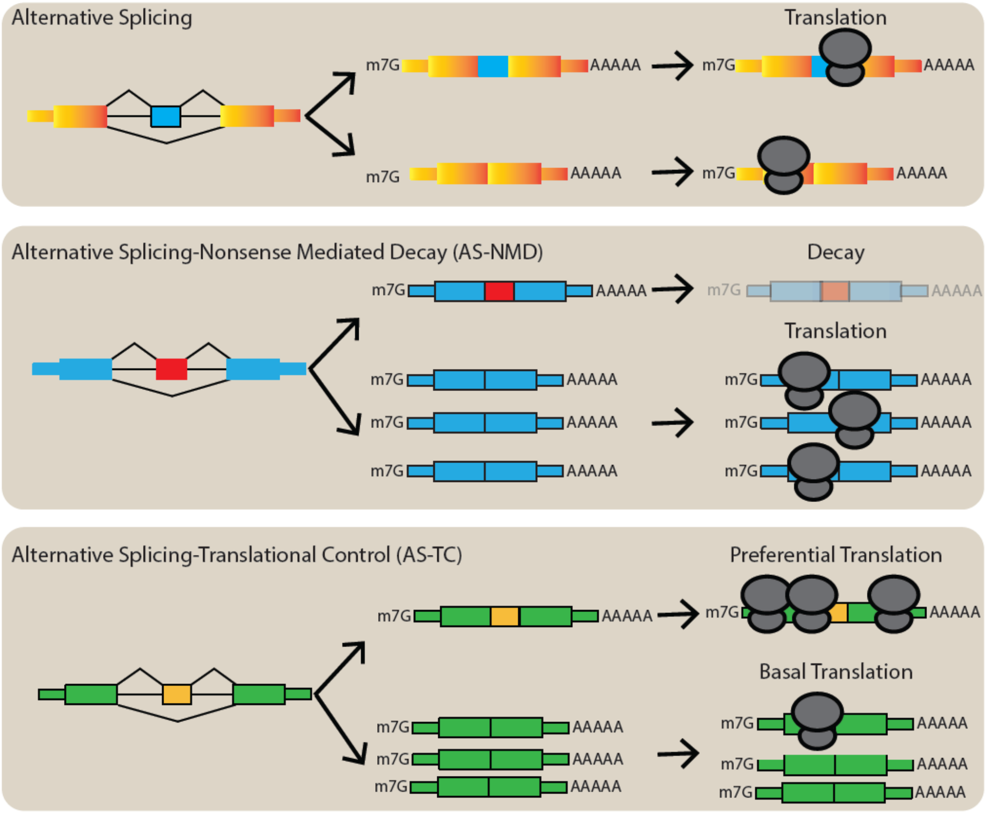

Alternative Splicing
Translation
m7G
AAAAA
Alternative Splicing-Nonsense Mediated Decay (AS-NMD)
Decay
Translation
Alternative Splicing-Translational Control (AS-TC)
Preferential Translation
Basal Translation

Figure 8: Functional heterogeneity of the alternative transcriptome. From a single gene, alternative splicing can create transcripts coding for different proteins (top), transcripts which can be selectively degraded (middle), or transcripts translated at different levels (bottom). Image from Sterne-Weiler et al, ref. 61. Creative Commons License (Attribution-NonCommercial 3.0 Unported).

The ability to switch RNA processing programs towards non-translated transcripts enables the cell to regulate gene expression without the need to tune of the amount of pre-mRNA produced. For instance, during macrophage differentiation, a subset of highly expressed genes switches to the production of NMD-target transcript isoforms, thus down-regulating protein synthesis.

Such program, which is independent from nascent RNA production, is necessary to ensure the correct differentiation program, and confer macrophages their peculiar shape and function[63]. In a scenario where the exact structures of thousands of transcripts can vary, together with their functions, we need to gather transcriptome-wide information about the pool of RNAs present in our system of interest. A detailed understanding of modern technologies, together with their applications, promises and limitations, is thus required to query the functional status of entire transcriptomes.

2.2 Omics techniques to understand RNA biology

2.2.1 Next-generation sequencing

The detection of an RNA transcript in a cell can be accomplished by using reverse transcription (RT) coupled with PCR (RT-PCR). The obtained DNA product can then be visualized and quantified using agarose gel electrophoresis. This procedure can be run in parallel to detect dozens of transcripts, but it requires precise knowledge of their sequence (for the reverse transcription reaction), and lacks the sensitivity to detect lowly abundant products. Imaging techniques, such as single-molecule FISH can also help us identifying the presence of RNA molecules, together with their spatial location in the cell. Unfortunately, both imaging and RT-PCR can only give us information about few transcripts at a time, while cells simultaneously transcribe and translate tens of thousands of RNAs.

After the complete sequencing of the human genome, different companies started manufacturing microarrays, sets of thousands of DNA probes, who could selectively capture and quantify different DNA molecules. When used on a pool of retro-transcribed RNAs, microarrays could give us information about thousands of known RNA transcripts, representing a big step forward in the study of the entire set of transcripts (the *transcriptome*). In the meantime, a tremendous improvement in DNA sequencing techniques allowed the sequencing of large pools of DNA molecules (a *library*) with high precision, giving rise to Next-generation sequencing (NGS), revolutionizing genomics and all its applications[64]. One of the most successful sequencing chemistry is the one adapted by the Illumina company, which is the one employed for most of the data presented in this dissertation.

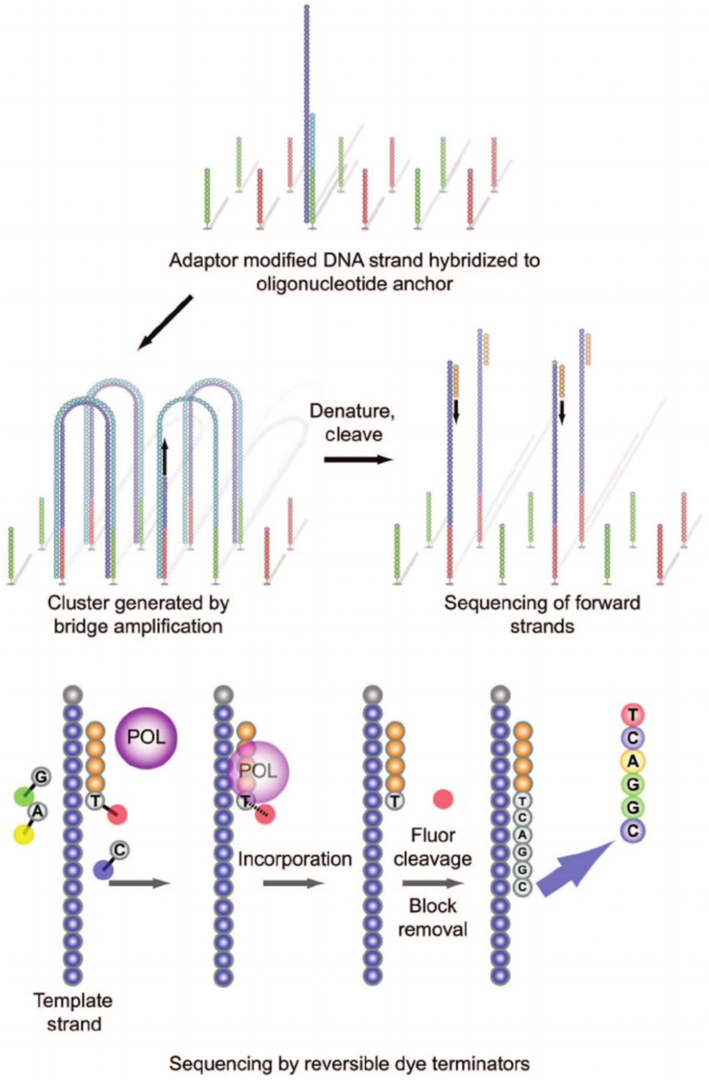

Adaptor modified DNA strand hybridized to oligonucleotide anchor
Denature, cleave
Cluster generated by bridge amplification
Sequencing of forward strands
POL
G
A
T
C
Incorporation
Fluor cleavage
Block removal
Template strand
T
C
A
G
G
C
Sequencing by reversible dye terminators

Figure 9: Illumina sequencing-by-synthesis approach. DNA with attached adapter sequences hybridize to the surface (top). The opposite end of the DNA fragment hybridizes to another proximal anchor (middle), thus forming a platform for DNA amplification with the help of polymerases, dNTPs and primers (not shown). After generating a cluster of identical DNA fragments, sequencing of one or two extremities of the fragments can be performed (here shown for only one strand). This sequencing reaction is carried using modified nucleotides which allow the polymerization of only one nucleotide. At each cycle, a labeled nucleotide is incorporated, and its attached fluorophore is detected, revealing the original sequence. The number of cycles determines the length of the sequences fragment. Taken from ref 64. Usage allowed by the “Fair Usage” description, as described by copyright laws adopted by the publisher.

Special sequences, the *adapters*, are ligated at the two extremities of our DNA fragments. The DNA is then inserted in a flow cell of the sequencing machine. On each flow cell, millions of DNA fragments are spotted on a glass surface. As shown in Figure 9, these “anchor” fragments hybridize with one of the adapter sequences on our DNA molecules, thus immobilizing the DNA fragments. Fragments can be amplified, forming clusters of identical DNA oligonucleotides. Primers can now be used to specifically sequence one extremity (or two, see below) of our fragment. Using labeled nucleotides together with a fluorimeter, we can reveal the original DNA fragment sequence. The incorporation (and detection) of one nucleotide at a time is performed at each step in parallel for all the DNA fragments, until reaching the desired length. At the end of this procedure, we have a sequence of intensities per fluorophore, which can be decoded to yield millions of nucleotide sequences, called *reads*, representing fixed-length segments of the initial pool of input DNA fragments. This sequencing protocol produces, per flow cell (for a HiSeq 2000 machine), around 200 million reads (it can vary depending on the sequencer). This means we can achieve a substantial transcript coverage over a wide range of expression values, even when combining multiple samples per flow cell (*multiplexing*). However, the sequenced read length is around 100nt (for the HiSeq 2000), which would allow us to sequence only a tiny segment of each transcript. To overcome this limitation, the input DNA (after RT) can be fragmented to ensure a more uniform sequence coverage over different sub-segment of the original fragment. Alternatively, a modification of the protocol can sequence both ends of an anchored DNA fragment, producing two short sequences from the two different ends of the same molecule, going under the name of *paired-end sequencing*.

Despite some technological limitations[65], [66], NGS methods allows us to sequence entire pools of retro-transcribed RNAs, in a process named RNA-seq, which resulted in a superior alternative for the study of entire transcriptomes[67], allowing us to quantify the presence of known and novel RNA molecules, and proving to be extremely versatile in studying different aspects of RNA biology.

2.2.2 RNA-seq applications

As described in Section 2.1.3 and 2.1.4, an RNA molecule undergoes multiple processing steps, both in the nucleus and in the cytoplasm. The ability to couple NGS technologies with the isolation of RNA molecules in different stages of the RNA life cycle resulted in a tremendous explosion of RNA-seq technologies, which are allowing us to greatly advance our understanding of the dynamics of gene expression regulation.

A common RNA-seq protocol consists in isolating polyadenylated transcripts using oligo-dT beads, followed by reverse transcription, fractionation and sequencing. This procedure avoids the amplification of rRNA, which is by far the most abundant RNA in the cytoplasm, and at the same time it enriches for polyadenylated transcripts, which in most cases represent stable and translated RNA molecules (Section 2.1.5). A slightly different procedure consists in skipping the poly-A selection, and using different rRNA removal strategy, using beads (RiboZero[68]) or oligo probes followed by selective degradation (RNAse H), followed by fragmentation and sequencing. The population of RNAs coming from this protocol consists of a more heterogenous transcriptome, including transcripts lacking a poly-A tail, and other unstable RNA products, like unspliced nuclear transcripts[69].

From the rRNA-depleted pool of RNAs, one can also isolate smaller RNA fragments using gel electrophoresis or alternative methods. We can thus enrich for small RNAs, like miRNAs, together with other small RNA fragments derived by other experimental protocols. To get a clearer picture of cytoplasmic and nuclear RNA abundance, the input RNA for the library preparation can also come from cellular fractionation, from either the nucleus or the cytoplasm[70]. Other subcellular fractionation method can give us a view over the localized transcriptome, in compartments like the endoplasmic reticulum[20] or neuronal projections[71].

RNA fragments can also be selected for the presence of the cap at the 5'end, and then subjected to sequencing, to have a global view on the capped transcriptome and on the position of the transcription start sites[72]. Similarly, it is possible to enrich for the 3'ends of an mRNA to gain knowledge about the exact cleavage and poly-adenylation sites[73], or even about the poly-A tail length and composition[74].

RNA-seq can also be coupled to immunoprecipitation (IP) to pinpoint the precise binding location of an RBP. Different protocols, such as PAR-CLIP (Photoactivatable Ribonucleoside-enhanced Crosslinking and IP) [75] or eCLIP (enhanced Crosslink and IP) [76], introduce a cross-linking reaction followed by immunoprecipitation to isolate the bound pool of RNAs. RNA digestion followed by sequencing can pinpoint the exact binding site of an RBP, at single nucleotide resolution.

In the PARS (Parallel Analysis of RNA Structure) protocol, RNA can be digested by enzymes which selectively cut double-stranded RNA or hairpin structures. The cleaved fragments from the two treatments can be isolated before sequencing, yielding a map of genome-wide RNA secondary structures[77].

Thanks to pulse-labeling of RNA molecules it is possible to extract nascent transcripts at different time points, and follow their dynamics of synthesis, splicing and decay, giving us an unprecedented view on the RNA life cycle over the entire transcriptome[78].

As many RNA species can be degraded either in the nucleus or cytoplasm, it is possible to deplete members of the degradation machinery, to enrich for a pool of unstable RNAs, which can now be detected and sequences. This strategy enables us to zoom into unstable nuclear RNAs produced by the pervasive transcription of non-coding regions[79], or into unstable cytoplasmic transcripts degraded by the NMD machinery[57].

Many modifications can be added to improve the quantification estimates of the sequenced RNA species. For instance, additional oligonucleotides with randomized ends can be added to the adapter sequences, which are ligated to the RNA fragments prior to the PCR amplification step. This way, possible biases introduced at the amplification step can be greatly reduced, by collapsing the sequences coming from the same fragment[80]. The use of such molecular labels (called *unique molecular identifiers*, or UMI) has been shown to greatly reduce the technical noise in sequencing data, thus improve quantification estimates, using RNA-seq or any other sequencing strategy.

To investigate the translational status of different transcripts, researchers have historically made use of polysome profiling: after ultracentrifugation over a sucrose gradient, transcript will distribute over different fractions based on their association with polysomes. A sedimentation profile towards the heavier polysomes can be used as a proxy to define high rates of protein production, and shifts in the polysomal fractions can be used to monitor the different translation status of different transcripts across conditions. Very recently, this technique has been coupled to RNA-seq, obtaining a transcriptome-wide view of polysome association with different RNA species, showing differential translation output across different isoforms per gene[61], [62].
In 2009, a new technique, named Ribosome Profiling[31], was developed in the Weissman lab by Nicholas Ingolia, which revolutionized the field of functional transcriptomics, allowing us to map the position of millions of translating ribosomes over the entire transcriptome.

2.2.3 Ribosome Profiling

The Ribosome Profiling (or Ribo-seq) technique aims at isolating the RNA fragments translated by the ribosomes, and it is comprised of several steps[81], [82], summarized here (Figure 10):

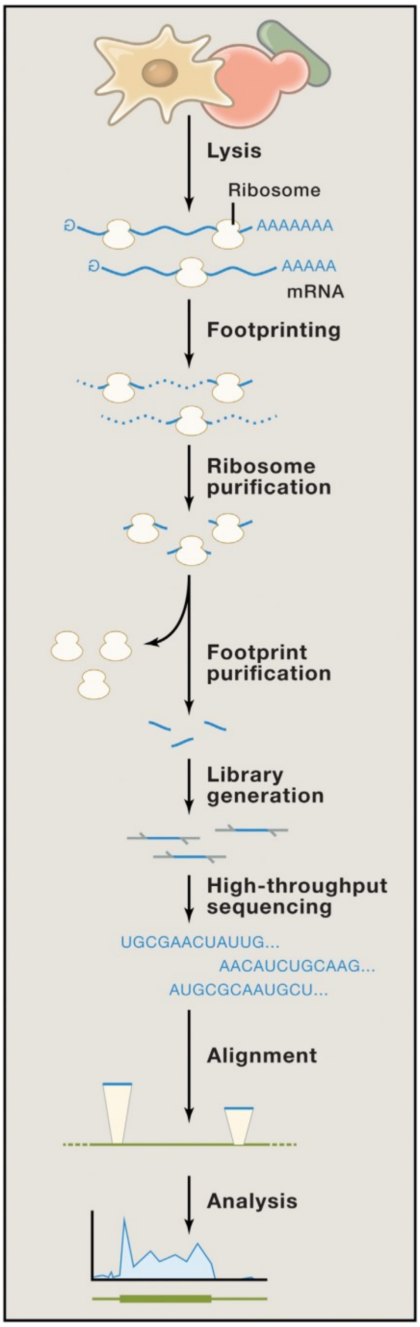


Figure 10: The Ribo-seq protocol. Cells are lysed, and the recovered RNA is subjected to footprinting. Ribosome footprints are purified and subjected to sequencing, followed by computational analysis. Image from ref. 82, released under license.

1) Cell Lysis

   Cells are lysed using a lysis buffer containing Cycloheximide (CHX). Thanks to the action of Cycloheximide, we can block translation elongation, thus “freezing” ribosomes on their original position on the mRNA. The use of alternative translational inhibitor is possible, as outlined in the sections below.

2) Nuclease Footprinting

   RNA digestion, using an endonuclease such as RNAse I, ensures that RNAs are uniformly cut, with exception to fragments protected by the ribosome.

3) Purification of protected fragment

   Ribosome-protected fragments (RPFs) can be purified using a sucrose density gradient or sucrose cushion ultracentrifugation. Alternatively, RPFs can be recovered using size exclusion chromatography.

4) rRNA depletion

   As we are purifying ribosomes together with their underlying RNA sequences, depletion of rRNA is crucial, and can be achieved by using different strategies, like the use of specific beads with the RiboZero method.

5) Size selection

   The expected 28-30 nt RNA fragments coming from the protocol can be extracted using PAGE (poly-acrylamide gel electrophoresis).

6) Library preparation and Sequencing

   The 3’ end of the fragments must be phosphorylated, in order to add adapter sequences for sequencing (Section 2.3.2). During this step, it is possible to use UMIs (Section 2.2.2) to obtain a more quantitative representation of the isolated pool of footprinted RNA.

Reverse transcription is then performed, with subsequent circularization and PCR amplification. The sample can now be sequenced.

A major variant in the protocol consists in adding Harringtonine[37] or Lactimidomycin[83] (LTM) in step 1 instead of Cycloheximide. This way it is possible to enrich for initiating ribosomes and quantify the usage of different translation start sites under different conditions. To further enrich for initiating ribosomes, the QTI-seq protocol[84] introduces a Puromycin treatment after adding the LTM. After puromycin treatment, elongating ribosomes stop elongating and fall off, thus allowing us to further enrich for initiation complexes.

Modification of this protocol allows for isolation of mitochondrial ribosomes[29], or ribosomes translating in subcellular compartments[59]. Very recently, a modified protocol able to isolate scanning ribosomes in the 5’UTR has been established in yeast, yielding a detailed picture of different ribosomal states during translation[85].

To yield a comprehensive view of the translatome, the Ribo-seq protocol needs a high amount of RNA material, usually in the µg range, usually corresponding to tens of millions of cells. A recent modification to the protocol consists in skipping the adapter ligation step by employing a polyadenylation strategy coupled to purification with oligo-dT beads, enabling sequencing of ribosome footprints from a much lower amount of input RNA material[86].

Despite the high amount of input required and a lengthy protocol (>4 days), Ribo-seq has been established in a different number of organisms, from bacteria to a whole range of eukaryotic organisms. However, differences in the choice of nuclease[87] and different drug treatments[88] heavily impact the obtained translational profiles in different published experiments, resulting in an overall poor standardization of the method[89], accompanied by a lack of uniform analysis strategies (Section 2.3.6).

2.2.4 Proteomics approaches

Together with the advancement in DNA sequencing, proteomics approaches also evolved from low throughput techniques (e.g. western blotting) to methods able to simultaneously detect thousands of proteins, leveraging on the unique biochemical properties of proteins. 2D gel electrophoresis is a technique used to separate and identify proteins by running a polyacrylamide gel on 2 dimensions, where one separates proteins based on the mass, and the other based on their charge. However, the dynamic range of the method is very limited (it is impossible to detect small proteins or big complexes) and it is very difficult to fully separate proteins with similar mass and charge properties.

Identifying and detecting the entire proteome in a cell can be extremely difficult, and this lead to the development of techniques able to focus on the detection of peptides coming from fragmented proteins, followed by computational reconstruction of the full-length proteins the peptides come from. This *shotgun proteomics* approach proved to be very successful in identifying thousands of proteins in a sample, giving rise to a revolution in the study of the proteome[90].

In a shotgun proteomics experiment proteins are isolated from a sample, and then digested using a proteolytic enzyme to produce a mixture of small peptides belonging to different proteins; trypsin is an ideal candidate to digest proteins as it is able to cut very often on a protein sequence with high specificity, and it creates charged peptides which are easier to detect; peptides are then isolated using high-performance liquid chromatography (*HPLC*) and analyzed using tandem mass spectrometry (or *MS/MS*).

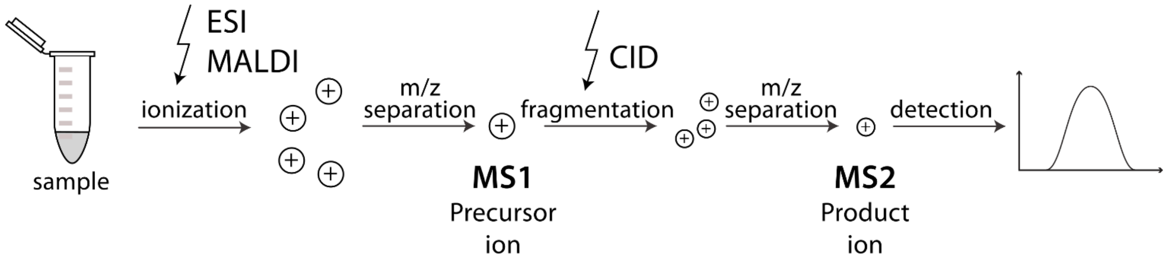


Figure 11: Shotgun proteomics example workflow. A peptide sample, previously digested with proteases like trypsin, is ionized, using techniques like ESI coupled to liquid chromatography, or MALDI when using a solid matrix. Ions are separated based on their mass and charge, and each ion (Precursor ion) is subjected to fragmentation, using collision-induced dissociation (CID). Fragments are subsequently separated based on their mass and charged and their intensity is detected on a membrane. Taken from: https://commons.wikimedia.org/wiki/File:TandemMS.svg, Author: Hannes Röst and M. Steiner. Creative Commons Attribution-Share Alike 3.0 Unported License.

A mass spectrometer consists of three different parts: an ion source, a mass analyzer and a detector membrane. In MS/MS peptides have to be first ionized (Figure 11). A common method is electrospray ionization, which ionizes peptides as they are being eluted by the HPLC column. As the peptides are eluted and ionized (here called *precursor ions*), they are sent to a collision chamber where they undergo a second fragmentation step. The fragmented ions are now sent to a detector which records the absolute mass-to-charge ratio (*m/z*) for each fragment. At the end, a spectrum of m/z values per each peptide is produced by the MS/MS procedure. This peptide-centric approach enables the detections of thousands of peptides which are then used to infer the presence and abundance of the cellular proteome.

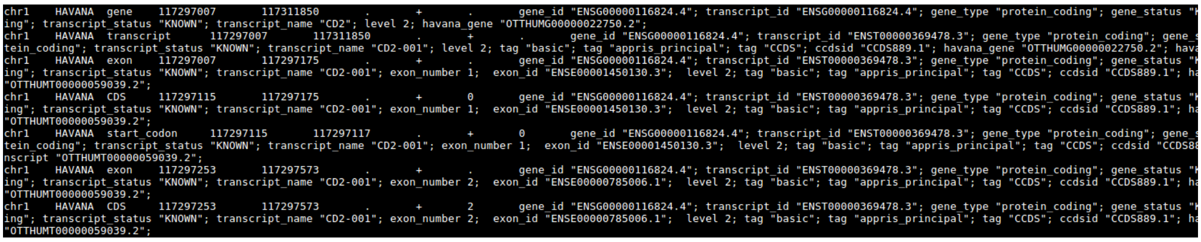
To infer accurate quantification in different conditions, proteins can be labeled using labeling compounds[91], [92], or labeled amino acids which can be added to the culture medium, as in the SILAC (Stable Isotope Labeling with Amino acids in Cell culture) approach[93]. The labeling can also be performed at different time points in a pulse-and-chase fashion (like in pSILAC[94]), to infer the dynamics of protein synthesis and degradation rates over the detected proteins[32].

2.3 Computational analysis of -omics data

2.3.1 Genomes and transcriptomes

The end result of a sequencing experiment is to extract signal from the experiment over genomic regions. The human genome comprises ~3.5 x 10^9 base pairs, and to perform analysis on such a large space standard data formats must be defined and understood.

A genome is represented by a sequence of nucleotides divided by chromosome, in a *.fasta* format, where each sequence has a name, a *header*. The ensemble of known functional gene structures in a genome (the gene *annotation*) is provided by a *gtf* file, which contains the genomic coordinates of different genomic features (Figure 12).



```
chr1    HAVANA  gene    117297007       117311850       .       +       .       gene_id "ENSG00000116824.4"; transcript_id "ENSG00000116824.4"; gene_type "protein_coding"; gene_status "K
ing"; transcript_status "KNOWN"; transcript_name "CD2"; level 2; havana_gene "OTTHUMG00000022750.2";
chr1    HAVANA  transcript      117297007       117311850       .       +       .       gene_id "ENSG00000116824.4"; transcript_id "ENST00000369478.3"; gene_type "protein_coding"; gene_s
tein_coding"; transcript_status "KNOWN"; transcript_name "CD2-001"; level 2; tag "basic"; tag "appris_principal"; tag "CCDS"; ccdsid "CCDS889.1"; havana_gene "OTTHUMG00000022750.2"; hava
chr1    HAVANA  exon    117297007       117297175       .       +       .       gene_id "ENSG00000116824.4"; transcript_id "ENST00000369478.3"; gene_type "protein_coding"; gene_status "K
ing"; transcript_status "KNOWN"; transcript_name "CD2-001"; exon_number 1;  exon_id "ENSE00001450130.3";  level 2; tag "basic"; tag "appris_principal"; tag "CCDS"; ccdsid "CCDS889.1"; ha
"OTTHUMT00000059039.2";
chr1    HAVANA  CDS     117297115       117297175       .       +       0       gene_id "ENSG00000116824.4"; transcript_id "ENST00000369478.3"; gene_type "protein_coding"; gene_status "K
ing"; transcript_status "KNOWN"; transcript_name "CD2-001"; exon_number 1;  exon_id "ENSE00001450130.3";  level 2; tag "basic"; tag "appris_principal"; tag "CCDS"; ccdsid "CCDS889.1"; ha
"OTTHUMT00000059039.2";
chr1    HAVANA  start_codon     117297115       117297117       .       +       0       gene_id "ENSG00000116824.4"; transcript_id "ENST00000369478.3"; gene_type "protein_coding"; gene_s
tein_coding"; transcript_status "KNOWN"; transcript_name "CD2-001"; exon_number 1;  exon_id "ENSE00001450130.3";  level 2; tag "basic"; tag "appris_principal"; tag "CCDS"; ccdsid "CCDS88
nscript "OTTHUMT00000059039.2";
chr1    HAVANA  exon    117297253       117297573       .       +       .       gene_id "ENSG00000116824.4"; transcript_id "ENST00000369478.3"; gene_type "protein_coding"; gene_status "K
ing"; transcript_status "KNOWN"; transcript_name "CD2-001"; exon_number 2;  exon_id "ENSE00000785006.1";  level 2; tag "basic"; tag "appris_principal"; tag "CCDS"; ccdsid "CCDS889.1"; ha
"OTTHUMT00000059039.2";
chr1    HAVANA  CDS     117297253       117297573       .       +       2       gene_id "ENSG00000116824.4"; transcript_id "ENST00000369478.3"; gene_type "protein_coding"; gene_status "K
ing"; transcript_status "KNOWN"; transcript_name "CD2-001"; exon_number 2;  exon_id "ENSE00000785006.1";  level 2; tag "basic"; tag "appris_principal"; tag "CCDS"; ccdsid "CCDS889.1"; ha
"OTTHUMT00000059039.2";
```

Figure 12: Example from the GENCODE 19 GTF file. Column 3 identifies the feature (exon, CDS, ...), columns 1,4,5,7 denote its genomic coordinates, while column 9 contains additional information about the gene and the transcript the feature belongs to.

For each element, a *transcript_id* and a *gene_id* column map the element to a specific transcript and gene, and a *gene_type* column reports the annotation category, or *biotype* (Section 2.1.2). Additional columns can be useful for cross-reference with other databases (like for the CCDS database, for *Consensus Coding Sequences*[95]), or to filter high-quality transcript structures (e.g. defined by the APPRIS system[96]), using tags like *ccdsid* or *appris*. The choice of annotation is an important step in a genomics workflow. The RefSeq database[97] offers a catalog of curated, non-redundant transcripts, while the GENCODE consortium, an integrated effort between ENSEMBL and HAVANA, contains a more comprehensive set of transcript/gene variants[12].

A set of transcript sequences can also be used for analysis, without any information about the exonic structures in the genome. This strategy can be successful when estimating transcript abundances[98] or when a genome assembly is not available[99].

2.3.2 NGS data pre-processing & mapping

The output of an NGS experiment in a sample is a sequence of intensities for each of the four reading channels (Section 2.2.1, Figure 9). Those intensities can be converted, for each position, into a nucleotide, forming a sequence of nucleotides (a *sequence read*, or simply a *read*) for each sequenced fragment. Depending on the number of cycles used in our sequencing reaction, we will obtain a longer or shorter sequence length (or *read length*). The sequence for each read is provided in a *fasta* file. When also a quality score is present for each position, the more common *fastq* format is used. The quality of our sequences is encoded in a *phred* score, which represents a probability for each position to have an erroneous intensity-to-nucleotide conversion. The phred score can be used to verify the overall quality of our sequencing results. During library preparation adapter sequences are added at the ends of our RNA fragments. Adapters must then be "trimmed" from our reads, using tools like *cutadapt*[100] or others. Depending on the length of the original RNA fragment, this might represent a crucial step in our analysis workflow (Section 2.3.6).

The obtained reads can now be mapped to extract the signal coming from the RNA-seq experiment. Mapping reads to a reference genome can be performed using software like *Bowtie*. Mapping algorithms can efficiently map millions of short sequences by building an index of our reference sequence using techniques like the burrows-wheeler-transform[101], suffix arrays[102], or the FM index[103]. Due to the action of splicing (Section 2.1.2), RNA-seq reads contain also transcript sequences mapping to sections of distant exons, which further complicate the mapping procedure. To solve this problem, different aligners were developed to be able to map reads spanning exon-exon junctions with high efficiency. TopHat[104] and STAR[102] are two of the most popular tools able to map RNA-seq reads also on exon-exons junctions, which use modified indexing strategies that include sequences formed by splicing. RNA-seq reads can also be mapped to transcript sequences instead of a genome sequence, and it is the starting point of many popular RNA-seq workflows[98], [105], [106].

To maximize the mapping performance, it is possible to allow some nucleotides to not map to the reference perfectly (*mismatches*): numerous Single Nucleotide Polymorphisms (SNPs) and the presence of sequencing errors favor the use of mismatches when mapping NGS data. A single sequence can also map to multiple positions in the genome, due to the pervasive presence of repetitive elements in the human genome. Of importance is the presence of thousands of annotated pseudogenes, which can be derived from gene duplications and subsequent inactivation, or by retro-transposition on a full mature transcript in the genome (*processed pseudogene*). Depending on the read length and sequencing strategy (single- or paired-end), different regions of the genome cannot be mapped uniquely. Therefore, the choice of parameters for read mapping must be adapted to the RNA-seq protocol used.

After the mapping is complete, we get the alignments reported in a BAM file, containing the genomic position(s) where our reads mapped, together with their mapping statistics (presence of mismatches, multimapping statistics, etc…). The BAM file can be now parsed to extract gene- or region- level signal statistics.

2.3.3 Quantification and normalization strategies

In an RNA-seq experiment, the number of reads mapping to a genomic locus (e.g. a gene) is proportional to the steady-state expression of that locus. Most of the RNA-seq (and NGS in general) analysis focuses on the analysis of *read counts* (or simply *counts*) per genomic position. Counts can be extracted from alignment files using the genomic coordinates of the regions of interest. Gene coordinates can be downloaded from different databases, or extracted from a GTF file. Genomic coordinates are usually stored in a BED file. Using tools like different Bioconductor packages[107] or *bedtools*[108] it is possible to count the number of reads mapping to each gene/region.

After extracting the counts per gene, different normalization steps must be applied, as the number of counts per gene will be proportional to the length of the gene. Moreover, the overall sequencing depth will influence the total number of counts. To give an unbiased estimation of steady-state expression for different genes and different samples, the RPKM[67] (Reads Per Kilobase of exon per Million reads) metric was introduced, with the following formula for the gene g (Equation 1):

$$RPKM_g = \frac{C_g}{D/10^6} * \frac{10^3}{L_g} \quad (1)$$

The RPKM value normalizes the counts per gene C_g based on the gene length L_g and on the library depth D, representing one of the most used metrics to estimate relative gene expression.

However, the RPKM metric might not be suitable when comparing different experiments, as the sum of RPKM values is different in different samples, making comparisons unreliable. To overcome the issue, the TPM (Transcript Per Million, Equation 2) metric was introduced[98]:

$$TPM_g = \frac{C_g}{L_g/10^3} * \frac{10^6}{\sum_g^{\#g} \frac{C_g}{L_g/10^3}} \qquad (2)$$

In the TPM calculation the normalization by length comes first, and the normalized values are scaled to a total sum of 1 million, making TPM values more reliable when trying to compare gene expression in different samples. However, those measures can give us relative quantification estimates, e.g. how much one gene is expressed compared to others. To obtain absolute quantification estimates (e.g. RNA molecules per cell), internal RNA standards with known concentration (spike-ins) can be introduced in the sample and their RNA-seq coverage can be used to scale values for the endogenous genes[109].

Despite the proven quantitative nature of sequencing data, multiple sources of variability can introduce artifacts in quantification estimates. This introduces a series of artefacts when trying to assess differential expression between different biological conditions. To overcome these issues, many computational methods have been developed to model the distribution of counts per region using replicated data and different statistical models. The use of the negative binomial distribution as a model for count data allows the estimation of a technical component of variability and a variability induced by the experimental condition, which enables for identification of high-confidence differentially expressed genes (with tools like DESeq, DESeq2 or edgeR [110]–[113]). A similar strategy can be used to detect differential usage of single exons, where read counts on exonic regions are used to estimate condition-specific exon usage (as in the DEXSeq strategy [114]).

After initial attempts to quantify transcript expression from RNA-seq counts[115], more accurate quantification estimates were obtained by more complete statistical models of RNA-seq coverage, with tools like RSEM[98], which proved to be one of the most accurate expression quantification method for RNA-seq experiments[116], [117]. In the RSEM strategy, reads are first mapped directly to the transcriptome using Bowtie, allowing for multiple mapping position. Next, a directed graphical model is derived to calculate, for each sequenced RNA fragment, a maximum-likelihood estimate of its originating transcript[98]. The modeling consists of sets of different random variables, which correspond to the fragment length, mapping position in the transcript and position-specific quality scores of the sequence. The parameters for this model are estimated using the Expectation-Maximization (EM) algorithm, where the estimates are calculated and updated at each cycle. At convergence, the EM gives a vector of parameter estimates which are converted into transcript fractions for each fragment. The cyclic nature of the EM procedure ensures a better estimation of expression values for repetitive sequences, which represents a major advantage over other approaches.

Mapping to the transcriptome dramatically reduces the search space when compared to a full genome, but on the other hand makes it impossible to capture signal coming from introns or intergenic regions. Having a set of transcript sequences as a reference might be an advantageous solution, for example when the quality of the genome assembly is not satisfactory. A new wave of RNA-seq quantification methods also uses a transcriptome reference for RNA-seq quantification. Tools like Kallisto[105] or Salmon[106] employ a k-mer alignment strategy using k-mer hash built from transcript sequences. The advantage of such approach consists in a severe reduction of computational time and resources, with a minimal loss of accuracy when compared to tools like RSEM. However, any spatial information about transcript coverage is lost, making these approaches extremely useful when the sole purpose of the analysis is to estimate expression levels.

2.3.4 Beyond count-based methods

Aggregating the number of counts per gene gives us a quantitative representation of RNA expression, but it can also represent a simplification of the multiple steps of RNA processing happening during the RNA life cycle. Moreover, the analysis of RNA-seq signal along different transcript positions can reveal specific biases in our protocol, and help us improving our quantification estimates.

A known positive bias towards GC-rich sequences is present during PCR amplification in the library preparation for any NGS technique. Only recently, computational methods are trying to correct for it when quantifying gene expression[106]. Analysis of poly-A RNA-seq (Section 2.2.2) data revealed a bias towards the 3’end of transcripts, as the protocol selectively enriches for transcripts (or possible other fragmented products[118], [119]) with a poly-A tail. Very recently, it has been shown how poly-A RNA-seq, but also total RNA-seq after RiboZero treatment (Section 2.2.2) contain specific biases towards different sequences along the internal exonic sequences[120].

In addition to protocol-specific technical biases, the analysis of RNA-seq signal along the transcript space can also help us interpret interesting biological aspects hidden in the data. As an example, the RNA-seq coverage over a genomic region represents a mixture of signals, which are the results of the expression of different transcript isoforms. To solve these issues, algorithms like Hidden Markov Models (HMMs) became very popular in the study of genomic signals. In a nutshell, with HMMs we train a direct acyclic graphical model where each node corresponds to a different *state* in our signal. Each state corresponds to a different signal profile, and the result of the HMM is a sequence of states over our signal, which will enable us to distinguish between different patterns. Using RNA-seq and genomic sequence, HMMs have been used to solve the mixture of RNA-seq signal coming from different expressed transcript, and discover new splice isoforms[121].

However, the patterns in the coverage at the single nucleotide resolution must be carefully interpreted, as they might be the product of both technical biases and interesting biology. Different RNA-seq protocols will produce a signal profile which is dependent on the RNA pool we are isolating, creating the need for tailored analysis strategies. For example, in the PAR-CLIP protocol (Section 2.2.2), reads harboring a T->C mutation reflect the binding of an RBP to the RNA; methods such as PARAlyzer use a Kernel Density Estimate (KDE) to estimate the signal coming from the T->C reads against a background signal, to infer precise binding events[122]. Other RNA-seq techniques, like PARS[77] are able to map RNA secondary structure, which follows a specific 3-nt cyclical pattern in CDS regions, caused by the wobbling of the 3rd codon position (Section 2.1.1). Cyclical (or *periodic*) patterns in NGS data can thus reveal fundamental mechanisms of biological mechanisms. To reveal the presence of periodic patterns in a signal, spectral analysis methods can be used, like the *Fourier transform*.

2.3.5 The Fourier transform and the Multitaper method

The main idea behind Fourier methods is the representation of a signal into a series of oscillatory components, which can be mathematically described by sines and cosines functions (or *sinusoids*). Sinusoids can be represented using complex notation thanks to the Euler's formula (Equation 3):

$$e^{ix} = \cos(x) + \ \mathrm{i}\sin(x) \quad (3)$$

Using this compact notation, we can now fully describe a signal with an infinite series of periodic components. The mapping between the original signal (in its *time* domain) and its representation in terms of periodic components (the *frequency* domain) is achieved by the Fourier transform, whose idea stemmed in the early 19th century by work of Joseph Fourier, building on previous work by Lagrange, Gauss and others. In mathematical terms (Equation 4):

$$f(t) = \int_{-\infty}^{\infty} F(s)\, e^{i2\pi st} ds \qquad (4)$$

a continuous signal $f(t)$ can be represented by an infinite series of sinusoids $e^{i2\pi st}$ multiplied by their coefficients $F(s)$, integrating over all the possible frequencies ds. The same principle applies in the other direction, where we declare that (Equation 5):

$$F(s) = \int_{-\infty}^{\infty} f(t)\, e^{-i2\pi st} dt \qquad (5)$$

the coefficients $F(s)$ can be calculated by multiplying the original signal $f(t)$ to a series of periodic frequencies, this time integrating over time dt.

However, in many real-life applications, we want to apply our transformation to a discrete, finite signal. In this case, integrating over infinite frequencies or infinite time is impossible. The discrete version of the Fourier transform, now called *Discrete Fourier transform* (DFT), deals with finite signals and frequency *bins*, which depend on the resolution and length of our signal. The DFT formula now states (Equation 6):

$$x_n = \sum_{f_0=0}^{N-1} B(f_0)\, e^{i2\pi f_0\, n/N} \qquad (6)$$

Each data point x_n is obtained by the sum, per each frequency f_0, of the contribution of the function $e^{i2\pi f_0\, n/N}$, which is given by its coefficient $B(f_0)$. Inversely (Equation 7):

$$\hat{B}(f_0) = \sum_{n=0}^{N-1} x_n\, e^{-i2\pi f_0\, n/N} \qquad (7)$$

The (estimated) coefficient $\hat{B}$ for the frequency bin f_0 is calculated summing up the contribution of each data point x_n spinning at that frequency. As shown in Figure 13 the DFT enables us to switch between two different representations of the same signal, and to quantify the contribution of periodic components in our data.

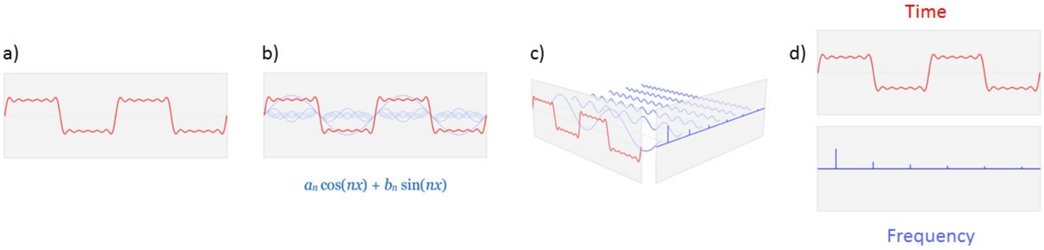


Figure 13: A schematic of the Fourier transform. In a) we can see our original signal, which can be explained by b) a series of sinusoids of different amplitudes (which can be also represented with Eq. 3). c) A vector of coefficients per each sinusoid is calculated from the original signal, thus enabling us d) to switch between the original signal representation (Time domain) and its spectral representation (Frequency domain). Adapted from: https://commons.wikimedia.org/wiki/File:Fourier_transform_time_and_frequency_domains.gif Author: Lucas V. Barbosa, released to the Public Domain.

Using the DFT to quantify the energy of each frequency component (the power spectral density, or PSD) in finite data is very challenging, representing an intense area of study in signal processing theory. The two main problems in PSD estimation are the presence of high levels of noise in the estimated spectrum, and the bias coming from the *leakage* of important frequencies. In presence of finite data, energy from some important frequencies can be detected (or *leak*) in other nearby portions of the spectrum, leading to incorrect estimates of the true PSD[123].

To reduce the noise in the estimated spectrum, one of the possible solutions is to lower the resolution of our spectral estimates, thus averaging estimates between nearby frequencies. Another possibility is to apply a window $\boldsymbol{a}$ to the original signal, which can lower the variability of the obtained spectrum[123]. The modified formula (Equation 8) now reads:

$$\hat{B}(f_0) = \sum_{n=0}^{N-1} \boldsymbol{a_n} x_n \, e^{-i2\pi f_0 \, n/N} \qquad (8)$$

Unfortunately, different windows have a big impact on the PSD estimation and can increase spectral leakage (Figure 14), posing the additional problem of choosing between the dozens of window functions known to date.

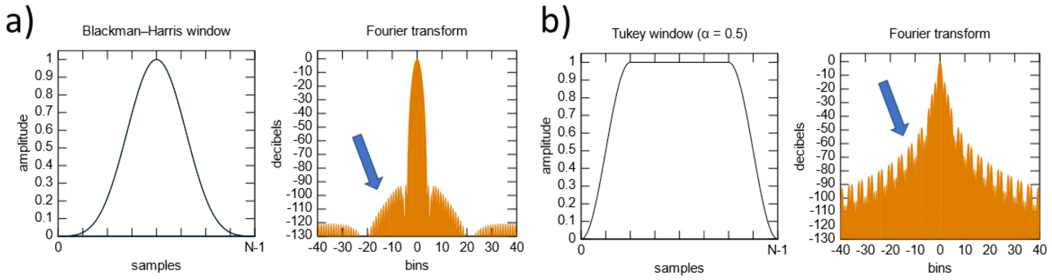


Figure 14: Different window functions and their spectral leakage. In a) the Blackman-Harris window is shown, together with its DFT. In the DFT plot, zero represents the period of the window. Ideally, the signal should be concentrated only around 0. As pointed by the blue arrow, the signal leaks over adjacent frequencies. In b) the Tukey window is shown, which displays a different spectral leakage pattern. Adapted from: https://en.wikipedia.org/wiki/Window_function

Trying to minimize noise and leakage in the power spectrum, the multitaper method was proposed by David Thomson[124]. The main idea behind the multitaper is the use of multiple windows (or *tapers*)

applied to the same signal, and average the periodogram over different windowed signals to reduce noise (Equation 9).

$$\hat{B}(f_0) = \frac{\mathbf{1}}{\boldsymbol{K}} \sum_{k=1}^{K} \sum_{n=0}^{N-1} a_{\boldsymbol{kn}} x_n \, e^{-i2\pi f_0 \, n/N} \qquad (9)$$

By applying multiple windows, the multitaper method proved to be very efficient in reducing the variability in the frequency spectrum, as also shown in Figure 15.

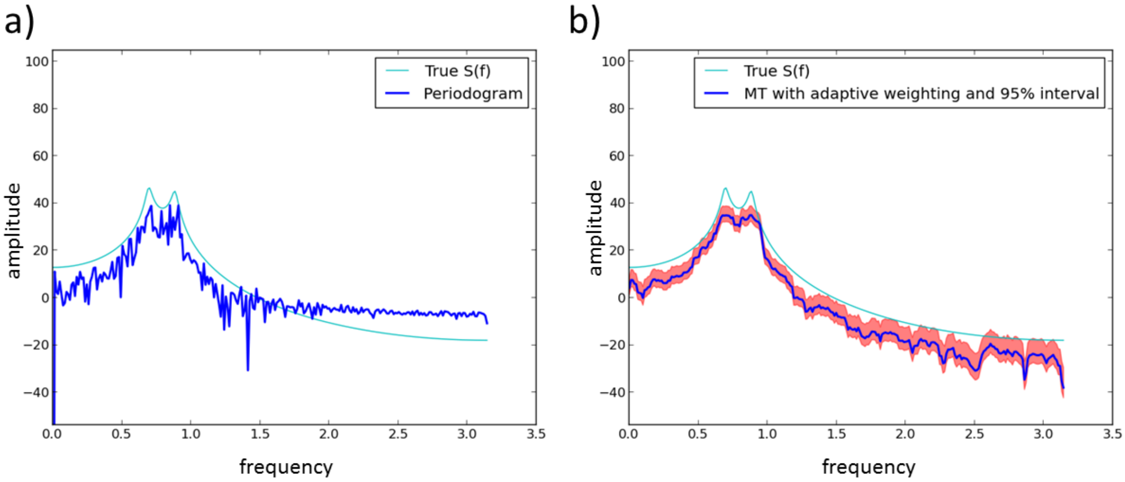


Figure 15: Example of multitaper PSD estimation. In a) we can observe the true signal and the periodogram obtained by the DFT. In b) the same true signal plotted together with its multitaper spectral estimate. It is possible to observe both a reduction in the noise and a reduced spectral leakage (signal does not “leak” to higher frequencies as in the raw DFT). Taken from: http://nipy.sourceforge.net/nitime/examples/multi_taper_spectral_estimation.html

The tapers used in the multitaper analysis are orthogonal window functions called Slepian sequences[125], initially studied by David Slepian, which allow the multitaper method to reduce noise, also providing additional useful properties. As stated earlier, a known problem in spectral estimation is caused by the presence of spectral leakage, which causes the spread of amplitude from one frequency over the neighboring sections of the spectrum (**Figure 14**). The multitaper method proves also to be efficient in reducing spectral leakage, as it maximizes the energy around a specific frequency resolution[124], [126] (Figure 16).

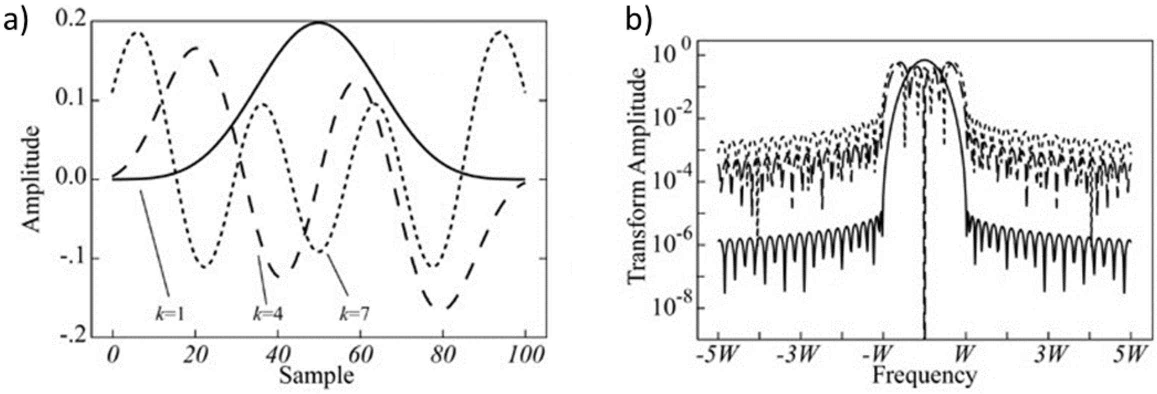


Figure 16: Example of Slepian sequences. a) Three slepian sequences depicted over a signal of 100 samples. b) The DFT of the 3 slepian sequences in a). Despite the presence of leakage, the energy is maximized around a desired frequency resolution. Taken from ref. 126. Usage granted by the journal's policy.

Additionally, the orthogonal nature of Slepian sequences allows us to derive an estimation of statistical confidence over the presence of frequency components in our signal[124], [127], representing a unique tool for the analysis of discrete signals.

With its unique ability to reduce both noise and leakage in the spectrum, the multitaper method represents an extremely valid analysis approach[128], and was successfully implemented in different programming languages[129]. Given its relevance to this work, a more complete explanation of the multitaper strategy and its mathematical formulation is presented in the Appendix A.

2.3.6 Ribosome profiling data analysis

The Ribo-seq protocol produces a set of short RNA sequences which were protected by translating ribosomes (Section 2.2.3). These sequences (also called RPFs) are very short, and adapter sequences need to be stripped from the read sequences, to obtain the exact footprinted RNA fragment. Depending on the efficiency of the rRNA removal step (Section 2.2.3), a strong percentage of reads consists of rRNA. It is advisable to remove those, as their massive presence can skew subsequent quantification estimates. tRNAs and snoRNAs can also take up a significant percentage of the obtained reads, and they should be filtered out.

As we start from a pool of RNA molecules, reads will also map to exon-exon junction sequences, thus a split-aware alignment like STAR[102] can be used. The mapping strategy must also take into account the short nature of RPFs, as many regions in the genome cannot be mapped uniquely when using a ~30 nt long sequence. To limit the effect of multimapping reads, which can take around ~20% of the total mapped reads, only the primary alignment (which can be extracted from the FLAG in the BAM file) per each read can be considered.

In a good Ribo-seq library, reads mostly map to CDS regions (usually >80%) and 5’UTR (~5-10%), and very little to 3’UTRs. Signal coming from introns and intergenic regions are usually the results of multi-mapping fragments.

A distinct read length distribution is observed in Ribo-seq libraries, which usually peaks at 29nt, as a result of the physical occupancy of a translating ribosome on the RNA[28]. A broader distribution of reads has been observed in variants of the protocol, for example using MNase instead of RNAse I as the nuclease used in the footprinting step[87] (Section 2.2.3), but usually followed by a loss of resolution in individual translation profiles. Different read lengths might also represent signal coming from distinct ribosomes, or different ribosome “states”: it has been shown in yeast[30] that a shorter footprint of ~20nt is visible when performing Ribo-seq in absence of CHX. This shorter footprint represents an alternative state of the elongating ribosome in a different conformation, which exposes a smaller surface towards the RNA, thus producing a shorter protected fragment. Distinct read length distributions can also be found in ribosomes belonging to different compartments. Mitochondrial ribosomes have been shown to display a bimodal distribution of read lengths, peaking at 27 and 33 nt, thus showing a clear difference when compared to cytoplasmic-derived RPFs[29].

The Ribo-seq signal over the translated ORF is dependent on the kinetics of the translation process: as the formation of the initiation complex (Section 2.1.5) is a relatively slow process, an accumulation of signal around the start codon can be observed in Ribo-seq data. In most datasets, an additional accumulation can be visualized at the last codon of the ORF, representing the ribosome in its termination state. When plotting only the 5'ends of the RPFs over the known start codons, it is possible to appreciate the single nucleotide resolution of Ribo-seq data, especially in more recent datasets[130], [131] (**Figure 17**). Such aggregate profiles over start and stop codons might greatly differ for different read lengths, so it is always advisable to investigate each read lengths and separate mitochondrial/chloroplast RPFs from the analysis.

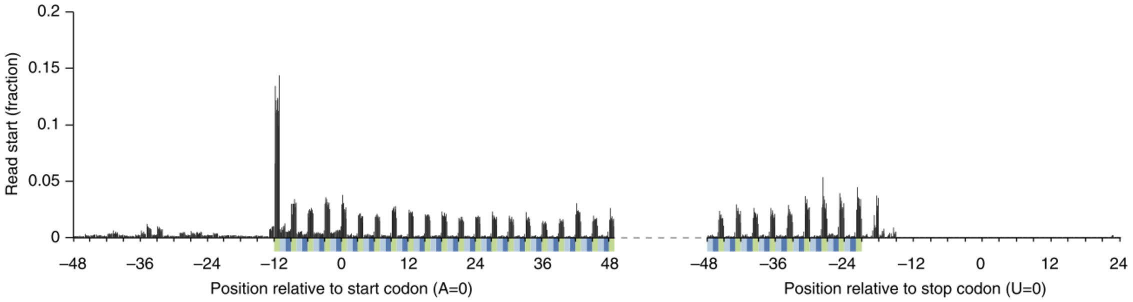


Figure 17: Sub-codon resolution in Ribo-seq data. The 5'ends of Ribo-seq reads are plotted over annotated start and stop codons. A peak of distance at 12nt from the annotated AUG can be observed, together with a clear 3nt periodicity along the translated frame, until the stop codon. 10 samples are plotted for each nucleotide, showing high consistency across samples. Adapted from ref. 131. Creative Commons Attribution 4.0 International License.

As shown in Figure 17, the 5'ends of RPFs pile-up at a distance of 12nt from the annotated start codons. Ribosomes initiate translation by locking the initiator tRNA in the P-site compartment (Section 2.1.5). This means that by adding 12nt from the 5'end of 29nt RPFs we can map the positions of the P-site ribosomal compartment for each RPF. Together with a clear peak of distance from the AUG, it is possible to observe a clear preference for the 5'ends to map on the translated frame, with little signal coming from the other 2 frames. This phenomenon clearly allows us to monitor the active movement of the elongating ribosomes, 1 codon (3nt) at a time (as drawn in Figure 7).

Using the multiple sources of information provided by Ribo-seq (RPF abundance, sharp read length distribution etc…), diverse analysis strategies have been proposed to extract biologically meaningful results from Ribo-seq data analysis, here summarized:

Quantifying translation over transcript levels:

The Translation Efficiency measure (TE) was introduced in the original Ribo-seq publication[31]. The TE metric tries to yield a quantitative measure of translation per transcript, dividing the Ribo-seq signal by the RNA-seq signal. Despite its usage in vast number of publications, the consistency of the TE measure in indicating translation has been discussed by following studies[132], [133], and a few alternatives have been proposed when trying to understand how translation changes in a differential analysis setting (Section 2.3.3). Tools like Xtail use the generalized linear model strategy of DESeq2[113] to model the Ribo-seq and RNA-seq read counts and obtain distributions of fold changes between conditions. Modeling the two distributions enables the significance testing for genes to belong to a concordant vs. discordant mode of regulation (on the translation or expression level) in the assayed conditions. Different approaches can be applied to the analysis of the distribution of RNA-seq and Ribo-seq fold changes, and models representing different modes of regulations can be subsequently tested against each other to distinguish between differences at the level of steady-state transcript abundance or differences at the level of translation[131].

Identifying translated regions:

By mapping the positions of ribosomes, Ribo-seq represents the most suitable technique for the annotation of CDS region. However, due to the intrinsic noise of NGS data and the complexity of the transcriptome/translatome (Section 2.1.2), the identification of high-confidence translated regions from Ribo-seq data is far from trivial. Especially in the early papers[31], [37], coverage plots of interesting transcript regions were showing how the Ribo-seq signal can be used to detect the presence of uORFs, non-canonical start codons or translation on non-coding RNA. As the sequencing of transcriptomes unraveled the presence of thousands of long non-coding RNAs (lncRNAs), different metrics were developed trying to detect differences of ribosome binding in lncRNAs with respect to known coding genes.

TE (publication date: 12.02.2009)

Many works initially took advantage of the TE metric to define actively translated transcripts[134], [135], but subsequent studies pointed out how ribosome abundance (even when normalized by transcript expression) cannot be used as a good proxy to define *bona-fide* translated regions without including a large number of false positive, both in ncRNAs and 3’UTRs[136]. A number of additional strategies were then employed to improve on the identification of the translated transcriptome.

RRS (publication date: 03.07.2013)

The Ribosome Release Score[136], distinguishes translated from non-translated regions by exploiting the release of translating ribosomes at the stop codon. This phenomenon creates a sharp decrease in coverage in Ribo-seq data at the end of the CDS. The RRS score is calculated as the ratio (normalized by RNA-seq reads) of RPFs in the CDS with RPFs in the 3'UTR. At the global scale, the RRS score successfully retains many coding regions and discards known ncRNA regions. However, only when combined with the TE metric the RRS shows a clear separation between CDS and non-coding regions of then transcriptome (e.g. 3'UTRs). Moreover, the definition of a CDS and a 3'UTR is challenged by the presence of multiple translated ORFs per transcript (e.g. uORFs), and the performance of the RRS score in such, very common, cases has never been explored. Additionally, an evaluation of the RRS score sensitivity and specificity in detecting translated regions (e.g. over different expression regimes, using simulations, negative data) has never been tackled. The RRS score lacks a proper documentation and its computational requirements and running time are unknown.

TOC (publication date: 11.07.2013)

The idea of using multiple metrics to detect high-confidence translated regions is the basis of the Translation ORF Classifier (TOC), proposed by Chew *et al*[137]. Four different features are extracted from the Ribo-seq coverage on different regions: the TE metric, for quantification of translation; Inside vs Outside, a metric containing the number of nt covered by Ribo-seq inside and outside the ORF; Fraction Length, representing the size of the ORF over the transcript length; the Disengagement Score, which is the same as the RRS score but without the RNA-seq normalization. A random forest classifier is trained on the 4 different features to understand whether Ribo-seq signal over lincRNAs resembles translation over known protein-coding genes. The output of the classifier is a label per each locus, which distinguishes between coding-like, trailer-like (3'UTR, no reads) and leader-like (5'UTR) loci. The majority of lincRNAs with Ribo-seq signal were assigned a leader-like label, thus presenting features not resembling *bona-fide* protein coding active translation, but still leaving the functional relevance of their translation unanswered. Given the high sequencing depth of the Ribo-seq datasets used (>300 Million mapped reads), the classifier showed good performance also on lowly expressed transcripts, but its performance on other datasets is unknown. It was not never released as a software tool for the community.

FLOSS (publication date: 11.09.2014)

During the Ribo-seq protocol, a size around 29nt is cut after PAGE (Section 2.2.3), to isolate the RPFs. In addition, different contaminants such as rRNA, snoRNA and other structured RNA fragments can survive the next purification steps and thus be sequenced. The idea behind the FLOSS score[138] is to learn a distribution of Ribo-seq fragment lengths on protein-coding region, which represent actively translating ribosomes. Fragment length distribution over each region in the transcriptome is then compared to the reference one, to derive a similarity score indicative of its coding-like validity, taking into account the total Ribo-seq coverage. As expected, the FLOSS scores globally distinguish coding versus non-coding genes. However, even for some predominantly nuclear lincRNAs like MALAT1[139], short elements along the transcript might exhibit a coding-like behavior, thus being masked by the total signal over the transcript. Despite multiple lines of evidence showing the sensitivity of the FLOSS score in detecting new *bona-fide* translation events, an in-depth analysis of the FLOSS score performance was not tackled. The method, available as a set of annotated scripts in a supplementary file, was applied to a very deep Ribo-seq dataset in a mouse cell-line (>250 Million mapped reads).

ORF-Rater (publication date: 03.12.2015)

In the ORF-Rater strategy[140], aggregate profiles over start and stop codons are used to identify translated regions. As these profiles become prominent in Harringtonine or LTM-treated Ribo-seq datasets, the usage of multiple Ribo-seq protocols over the same biological samples produces distinct profiles for many translated ORFs. The core of the ORF-Rater method is a regression fit of the Ribo-seq coverage (coming from the multiple Ribo-seq protocols) along the transcript against its expected coverage given the presence of one (or multiple) translated ORFs. The presence of ORF translation is indicated by a positive regression coefficient of the fit. To evaluate the statistical confidence of the ORFs translation, a random forest classifier is trained on regression results coming from known ORFs and used to score the regression fits for the ORFs candidates, yielding a high-quality set of translated ORFs, covering known and novel genomic regions.

Leveraging on expected profiles at start and stop codons, the ORF-Rater method is able to identify ORF truncations/extension, out-of-frame ORFs and very small ORFs (>20 codons), bypassing some of the limitation imposed by other approaches (see Discussion). However, despite the high quality of the detected candidates, it is not clear whether the high requirements of the method are met by the entire set of translated ORFs in the transcriptome. Different kinetics of initiation and termination might produce coverage profiles different from the expected ones, especially in lowly expressed genes. The general applicability of the method is also challenged by the high data requirements, as the omission of some of the Ribo-seq variants can dramatically reduce the algorithm performance[140]. On a very deep Ribo-seq dataset in a mouse cell-line (~150 Million reads per each of the 4 Ribo-seq variants) the method was run on a high-computing cluster using 256 Gigabytes of RAM and multiple processors, with a runtime of a couple of days. The method is implemented as a software freely available on a Github repository, with documented scripts and detailed usage description.

riboHMM (publication date: 27.05.2016)

Despite the presence of a different modeling framework, the riboHMM[141] strategy to detect translated ORFs uses a similar idea to ORF-Rater. An Hidden Markov Model (HMM) is trained to recognize distinct Ribo-seq profiles over different ORFs positions, leveraging on the distinct pattern of Ribo-seq over start and stop codon, and inside the translated CDS. The model also explicitly models the contribution of each Ribo-seq read length, and sums them over to increase sensitivity. The trained HMM is used to parse the Ribo-seq signal transcriptome-wide, yielding predictions for ORF translation. Using a very deep Ribo-seq dataset in human (580 Million reads) and stringent filtering, riboHMM identified ~36K translated transcripts, covering ORF annotation for 7801 annotated protein-coding genes and thousands of novel candidate ORFs. At lower library depths, the algorithms showed to be robust in terms of False Positive Rate, despite a marked decrease in sensitivity. Runtime information and computational requirement are not provided. The method is available as a software free to use for the community and it is well documented.

ORFscore (publication date: 04.04.2014) & similar studies

As the protocol became more popular, also the overall data quality increased. In 2014, Bazzini *et al*,[130] produced a massive Ribo-seq dataset with precise sub-codon resolution following the Zebrafish early development. Having precise information about the translated frame, they scored different ORFs based on the number of reads falling on the translated frame, compared to a uniform distribution of signal over the three frames. This scoring method, named ORFscore, allowed them to identify a set of translated small ORFs (<100 aa) which were overlooked by automatic annotation pipelines. Despite its high sensitivity on a deep Ribo-seq dataset in Zebrafish (~200 Million reads), the specificy of ORFscore and its performance on different datasets is unknown.

Similarly, two other studies in yeast used the sub-codon resolution of Ribo-seq reads to identify translated ORFs[142], [143]. In one of the two studies, a False Discovery Rate on the ORF identification was calculated using a randomized distribution of P-sites over the three frames[143], drawing from the same assumption behind the ORFscore. None of these approaches were originally implemented in a software available for the community.

RibORF (publication date: 19.12.2015)

RibORF[144] is a method which builds on the ability of the sub-codon resolution of Ribo-seq reads to identify translation. In addition to the amount of Ribo-seq reads in frame, the method uses the Percentage Maximum Entropy (PME) metric to ensure a more uniform coverage of reads along the ORF. The percentage of reads in frame and the PME metric are calculated for each ORF in the transcriptome, and a Support Vector Machine classifier is used to separate good ORF predictions from unreliable results. Around 10.000 translated genes were detected applying RibORF to two average Ribo-seq datasets in human cell-lines (~40 Million reads). The tool is implemented in a software free to use for the community, including essential usage instructions.

RiboTaper (publication date: 14.12.2015, Section 3.1)

Also in the RiboTaper method[145], the sub-codon resolution is key to identify translation. The method identifies regions where Ribo-seq reads display a 3nt periodic behavior consistent with ribosomal translocation, using the statistical test from the multitaper method, a known spectral analysis method. ~12.000 genes are detected using RiboTaper on an average depth HEK293 datasets (~30 Million reads). The algorithm was run on datasets of different quality and from different organisms. Its runtime is ~1 day. Documentation and usage guide are available.

Spectre (publication date: 25.12.2016)

In the Spectre[146] method, spectral coherence (which measures the correlation between two different frequency spectra) is used to indicate whether the periodic components in the P-sites profile match an ideal profile where reads map only to the translated frame. Using quantification estimates from Cufflinks[115] to normalize the P-sites tracks, the algorithm uses coherence values to classify individual transcripts into translated or not translated. Sensitivity and specificity are addressed at different degrees of expression levels. The code is made available and well documented. Runtime is expected to be less than 1 day.

Rb-Bp (publication date: 25.01.2017)

The Rb-Bp[147] strategy uses a probabilistic graphical model to predict translation from P-sites profiles. The model is trained to recognize a pattern where a clear enrichment over one translated frame is observed, and it scores ORFs whether they resemble such pattern or a null uniform model. As with RiboTaper, the algorithm's predictions were validated with proteomics support and QTI-seq data. The algorithm was run on different datasets of modes depth. An evaluation of the method specificity or sensitivity was not extensively presented. Documentation is available and the runtime is expected to be less than 1 day.

Given the scarcity (until recently) of solid analysis methods, several Ribo-seq studies attempted to identify translated ORFs using custom analysis pipelines, from read mapping to ORF identification and variant calling. For example, in the PROTEOFORMER[148] pipeline, translation is identified by looking at Ribo-seq counts over ORF boundaries in CHX and Harringtonine/LTM-treated samples, also taking into account the presence of sequence variants. Of particular note is a study in murine myoblast differentiation where the entire analysis workflow is freely available online[149].

Detecting alternative translation events

Different features of Ribo-seq signals along the translated frames have also been used to identify the presence of alternative translation events. Using a change-point algorithm, Zupanic *et al* [150] detected sharp changes in the Ribo-seq coverage to identify novel initiation sites, premature stop codon usage and novel splice junctions. Despite the presence of new interesting events, it is not clear how the thousands of change-point events all reflect the presence of true alternative translation events, especially considering the high non-uniformity of Ribo-seq signal. Leveraging on the sub-codon resolution of Ribo-seq reads, Michel *et al*,[151] developed a strategy to identify regions where translation occurs on multiple frames. The authors identified ~100 candidates where the ribosomal coverage switches between two different frames along a single transcript. Among the candidates there are two genes with known ribosomal frameshifting sites, >40 of overlapping small ORFs (mostly uORFs), regions where a mixture of signals from multiple RNA isoforms occurs, 13 unexplained cases and 33 (manually verified) false positives. After careful removal of false positive results, the authors could show how these regions are indeed coding in two frames, using evolutionary conservation over the dual coding regions.

As the Ribo-seq protocol became more popular, different datasets have been published, and slight variation in the protocol were observed to have an impact on the obtained signal profile. Together with a plethora of exiting discoveries, also different surveys about possible biases present in Ribo-seq data began to appear, highlighting the need of protocol standardization and more careful approaches in the data interpretation. The kinetics of CHX intake, for instance, can distort the Ribo-seq coverage profile, creating artefactual, but reproducible, patterns. A low concentration of CHX can produce an enrichment of signal around the start codon[88], created by a slower drug intake by the elongating ribosomes.

The choice of nuclease for the footprinting step has a marked effect on the overall resolution of the data. Micrococcal Nuclease (MNase) has been shown to have strong sequence biases, while RNAse I can disrupt the native monosome structure, possibly creating biases in the footprint recovery[87]. It has been recently proposed that a mix of different nucleases might represent a good compromise between high resolution and footprint integrity[87], but its validity over different species remains to be tested.

Other variables in the experimental protocol have an impact over the Ribo-seq signal over different transcripts. The CircLigase A, often used during the library preparation step, has a bias towards A-starting footprints, thus enriching for specific RNA fragments and creating stronger signals at specific transcript positions. Softwares like RUST[89] have been recently developed to quantify the presence of bias in different Ribo-seq libraries, allowing the community to improve over the published protocol and the interpretation of the obtained data.

However, non-uniformity in the Ribo-seq signal also derives from the biology of the translation process. As seen before, enrichment of signal over start and stop codons derives from the slow kinetics of translation initiation and termination. During translation elongation, Proline codons can stall elongating ribosomes, due to its inefficient incorporation during the peptide bond formation[152]. Additional codon pairs also seem to efficiently stall elongating ribosomes[153]. Co-translational folding of the nascent protein is also a determinant of ribosome movement along the ORF, and it is possible to observe different ribosomal speed whether the nascent peptide sequence folds into a coiled-coil structure or an alpha helix[154]. However, such results are very recent and a clear confirmation of the possible mechanisms at the molecular level is still lacking.

In a summary, a high degree of variability is introduced both by variables intrinsic in the experimental protocol and by the yet unresolved molecular mechanisms of translation. The unmet assumption of uniformity of Ribo-seq signal can thus greatly impact the validity of the obtained results, and multiple lines of evidence from different sources must be presented to gain confidence in the analysis of Ribo-seq data.

2.3.7 Evolutionary signatures on genomic regions

To gain insights on the possible functional role(s) of genomic elements, the analysis of evolutionary conservation patterns over different species proved to be a very successful approach[155]. With the increasing number of sequences and entire genomes available from different species, statistical methods of nucleotide substitutions allowed the possibility of defining phylogenetic trees across multiple species. The alignment of sequences and entire genomes between different species enabled the detection of elements which remained relatively “unchanged” during evolution in different organisms. The presence of evolutionary constraints over a sequence can be used as an indicator of molecular functionality, allowing the detection of important transcriptional enhancers, miRNA binding sites[156] or unannotated small proteins[157]. The idea behind the popular PhastCons[158] program is the use of a two-states HMM, which parses through a multiple species alignment and uses a phylogenetic tree to recognize the sequences as conserved or non-conserved. For each nucleotide, the phastCons model emits a probability for that nucleotide to belong to a conserved element. These probabilities (which represent the posterior probabilities of the HMM), are the popular phastCons scores widely used in genomics to evaluate the conservation of nucleotide sequences (Figure 18).

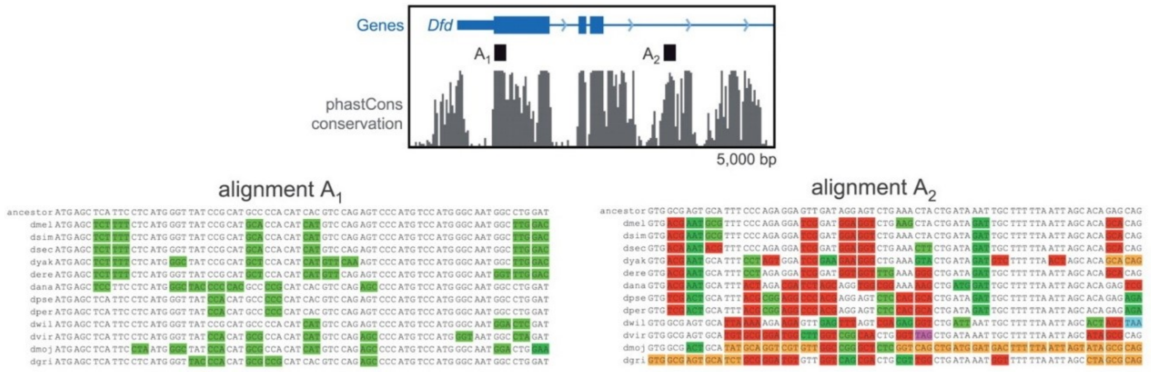


Figure 18: Evolutionary signatures of genomic regions. Different genomic regions (top) can exhibit high level of nucleotide conservation. However, they might drastically differ in terms of codon substitution rates, where coding regions (left) retain the aminoacidic sequence, thus being enriched for neutral or synonymous substitutions, while non-coding regions (as the intron on the right) present high rates of non-synonymous substitutions. Adapted from ref. 159, released under license.

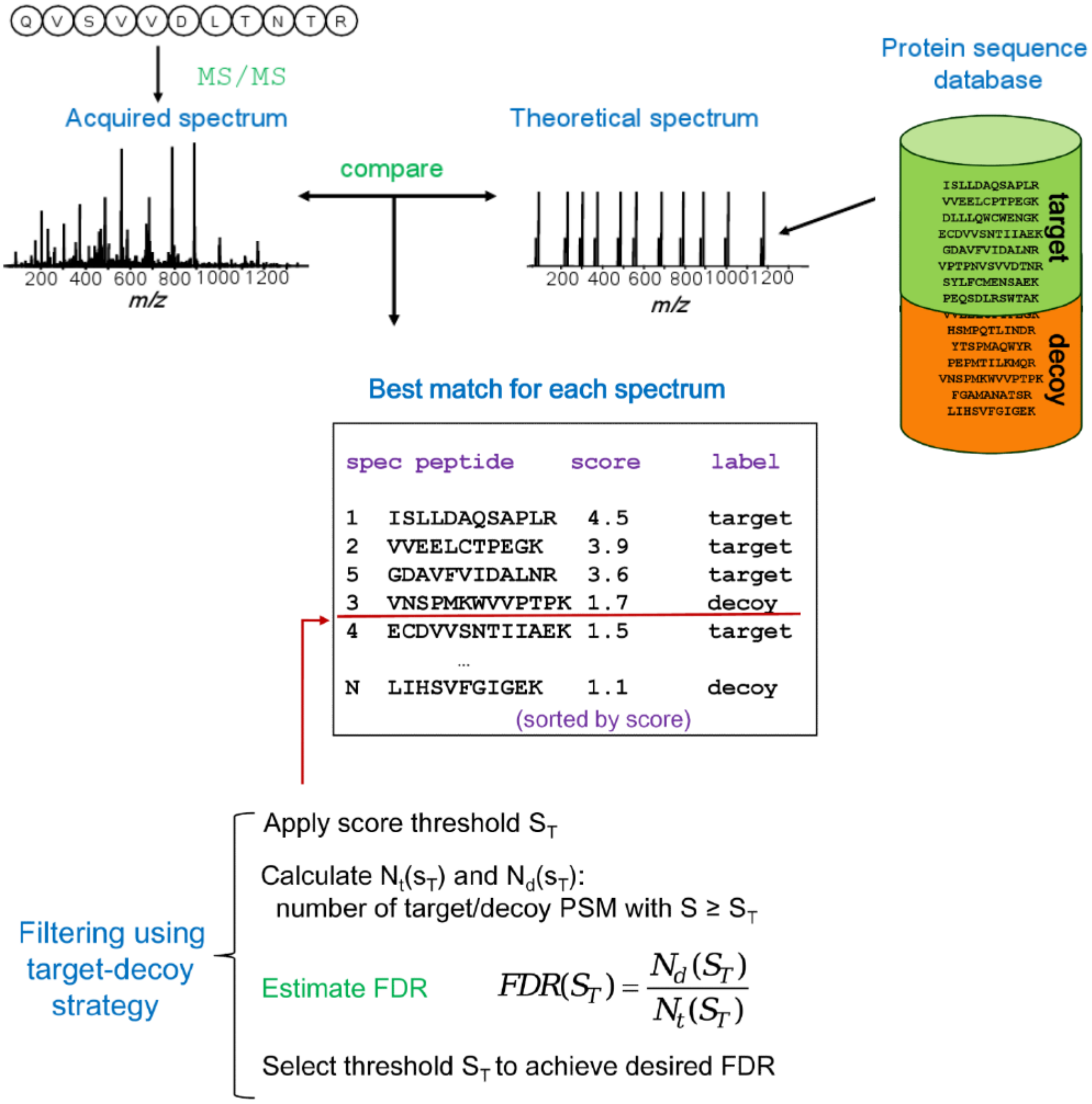
Evolutionary constraints on the nucleotide sequence might reflect a conserved binding activity from a regulator, which might be important in the regulation of transcription or cytoplasmic processing. To understand whether the encoded protein sequence (or parts of it) is under selective pressure, additional methods were introduced in the analysis of evolutionary conservation of nucleotide sequences, as selection on the protein sequence poses constraints on the composition of its coding sequence.

The PhyloCSF[159] approach calculates, for each nucleotide triplet (codon) in protein-coding and non-coding genes, two separate codon substitution models from multiple sequence alignments. For a genomic sequence, the likelihood ratio between the two codon substitution models can be used to understand whether the sequence belongs to a conserved coding locus or not (**Figure 18**). As seen in Section 2.1.1, the degeneracy of the genetic code allows for the presence multiple codons per amino acid. Positive selection for mutations which do not disrupt the encoded amino acid sequence (synonymous mutations) is observed in conserved protein-coding sequences, when compared to disrupting mutations (non-synonymous). The ratio between synonymous versus non-synonymous mutation (Ka/Ks), has been used to identify protein-coding sequences using evolutionary sequence alignments[160], [161]. Sequence variation (as single nucleotide polymorphisms, or SNPs) in the population also show a similar pattern (ratio named dN/dS), as synonymous mutations (dS) are higher in coding than in non-coding genes.

Additional approaches use the sequence composition of known coding and non-coding genes to derive a set of sequence features. Without the use of sequence alignments, a classifier is then trained on those features to distinguish between coding and non-coding genes, showing good performance when compared to alternative approaches[162].

2.3.8 Shotgun proteomics data analysis

In an MS/MS experiment (Section 2.2.4), we obtain an ensemble of spectra containing m/z values for each detected ion. The spectra of values will result from the elution of one peptide, whose identity is unknown. To resolve the initial mixture of peptides in our sample, we can map the obtained MS/MS spectra to proteome-wide *theoretical* spectra calculated from a sequence database[163]. In this approach, annotated proteins in a sequence database are *in silico* digested, and theoretical spectra are calculated. Different methods have been implemented to derive a meaningful matching between theoretical and real spectra: one of the most popular algorithms is represented by the Andromeda engine[164] (part of the popular MaxQuant software[165]), or by MS-GF+ [166], only to name a few. While the different algorithms have marked differences in their scoring systems, they use a similar strategy to evaluate the quality of the Peptide-Spectrum-Match (PSM); a decoy database is built on the original one (usually by reverting the protein sequences), and a measure of statistical significance (usually FDR-based) can be calculated from the match between the experimental data and the two (real and decoy) databases[167] (Figure 19). To further increase confidence in the identification, a recent approach combines the possible strengths and weaknesses of different search engine in a unified platform to identify the detectable proteome[168].



spec	peptide	score	label
1	ISLLDAQSAPLR	4.5	target
2	VVEELCTPEGK	3.9	target
5	GDAVFVIDALNR	3.6	target
3	VNSPMKWVVPTPK	1.7	decoy
4	ECDVVSNTIIAEK	1.5	target
	...		
N	LIHSVFGIGEK	1.1	decoy

(sorted by score)
Filtering using target-decoy strategy
Apply score threshold S_T
Calculate $N_t(s_T)$ and $N_d(s_T)$:
number of target/decoy PSM with $S \geq S_T$
Estimate FDR
$FDR(S_T) = \frac{N_d(S_T)}{N_t(S_T)}$
Select threshold S_T to achieve desired FDR

Figure 19: Analysis workflow for MS/MS data. Acquired spectra are compared with a scoring function to theoretical spectra, obtained from a sequence database which also contains decoy sequences (top). The scores from target and decoy matches are analyzed, and a measure of False Discovery Rate is derived by the hits on target vs. decoy sequences. Different cutoffs on the scores will determine the FDR. Adapted from ref. 163, released under license.

Important parameters to set in the database search step can depend on the experimental conditions: a variable amount of mass tolerance must be specified in order to correctly match the spectra; such values are technology-specific, depending on the mass spectrometer or on the resolution of the HPLC column in LC-based methods. Post-translational modification (PTMs) on peptides will produce different spectra, and the addition (or omission[169]) of multiple possible PTMs can have a great impact on the identification results. Similarly, the use of an incorrect database will also bias the identification results[170], making database search a fundamental step in proteomics data analysis. Analysis of Ribo-seq datasets can allow the identification of novel peptides (Section 3.2.3), despite its limited contribution in providing novel, identifiable proteins in recent efforts[148].

A large amount of unidentified spectra is present in a mass-spec experiment, and more tolerant searches might help enriching the catalog of identified peptides in a single experiment[171]. However, as public databases become increasingly richer and easier to use[172], and different technologies combine the power of targeted proteomics with high-throughput discovery[173], [174], the combination of different proteomics approaches will help us defining the functions of the cellular proteome [175], [176].

Quantitative estimates of protein abundance can be obtained from MS/MS data, for instance by measuring the number of spectra mapping to a specific protein, as in the NSAF (Normalized Spectral Abundance Factor) approach[177]. Another approach sums, for each protein, its peptides intensities obtained in the first MS run, and normalizes such value by the number of theoretical peptides mapping to such protein[32]. Such approach, named iBAQ (intensity-Based Absolute Quantification), became very popular for label-free quantification of MS/MS data, and its calculation is included in the MaxQuant package[165]. However, especially when comparing different biological conditions, the use of labeling techniques and internal standards[178], [179] might represent a superior alternative when trying to quantify protein expression.

3 Results

3.1 A novel approach to Ribo-seq data analysis

Contribution Statement:

Lorenzo Calviello performed all the sequencing data analysis, tested the multitaper performance and implemented the RiboTaper strategy, supervised by Uwe Ohler. Material appearing in this Section has been copied or adapted from our publication[145].

3.1.1 Spectral analysis of P-sites profiles

To obtain a comprehensive view on the detectable translatome, Ribo-seq was performed in HEK293 cells (Appendix B.1), yielding ~30 Million sequences reads. After removing adapters and rRNA reads, we mapped reads to genome using STAR (See Section 2.3.2) supplied with the GENCODE 19 annotation (Appendix B.2). Additionally, we analyzed previously published RNA-seq and Ribo-seq data from different sources, highlighted in Table 1.

accession	condition	Non-rRNA trimmed reads	Aligned reads	Uniquely aligned	read_length for P-sites calculation	Offsets For P-sites calculation	N of P-sites/RNA_sites
This_study	Ribo-seq	29,299,392	25,268,289	20,014,470	26,28,29	9,12,12	15,893,765
GSE49831	RNA-seq	33,701,799	27,688,698	26,289,844	NA	25	27,688,698
SRA160745	ribo_control	10,487,124	6,809,992	5,001,513	26,27,28,29	12,12,12,12	5,047,204
SRA160745	RNA_control	39,548,815	33,154,640	28,365,630	NA	25	33,154,640
GSE53693	ribo_5hPF_1	141,503,942	106,667,995	72,011,863	28,29	12,12	26,047,445
GSE53693	RNA_5hPF_1	114,116,421	71,834,250	52,766,644	NA	25	71,834,225

Table 1: Summary statistics for Ribo-seq and RNA-seq data in HEK293 cells and *Danio rerio*.

A critical aspect of Ribo-seq analysis involves the analysis of aggregate profiles over start and stop codons. As shown in Section 2.3.6, some read lengths display a distinct bias towards one of the translated frames, together with a clear peak of distance from annotated start codons (usually around 12 nt for 29 nt footprints, see Discussion), thus revealing the P-site position within each ribosomal footprint (Figure 7 and Figure 17). We investigated this pattern in our data, generating aggregate profiles over annotated start and stop codon positions (Figure 20, Supplementary Figure 1).

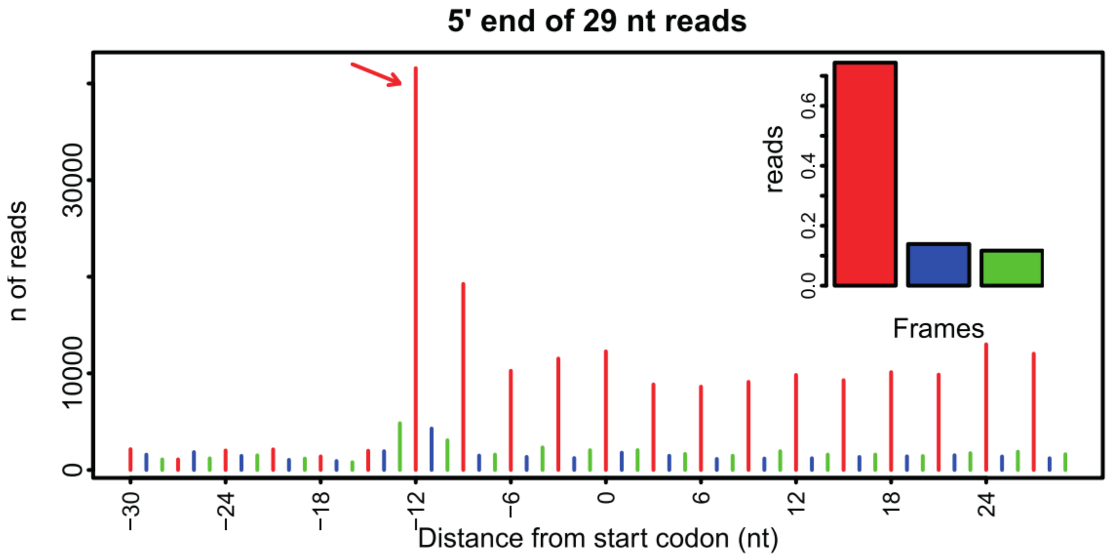


Figure 20: Metagene analysis of 29 nt Ribo-seq reads in HEK293. 5'ends of reads are aligned to annotated start codon position. Reads are colored based on the 3 possible coding frames. A clear peak at 12nt offset from the the start codon is visible. The barplot shows the fraction of reads mapping to the different frames.

Aggregate profiles were visually inspected, to infer millions of single-nucleotide P-sites positions which were used to determine the translated frames (Table 1, Supplementary Figure 1). When using P-sites positions to define the coding frame of each CCDS exon, ~90% or more of the exonic frames agreed with the annotation, suggesting high precision in the frame definition.

As this sub-codon resolution is caused by subsequent 3 nt steps during translation elongation (Figure 7), we decided to apply spectral analysis methods to confidently identify this pattern over the transcriptome. As seen in Figure 21, CDS regions displayed a clear peak of power at a frequency of 3nt after applying the raw Fourier transform (Section 2.3.5). This first observation confirmed the applicability of spectral analysis methods to the identification of 3nt periodicity in P-sites profiles, but also presented us with other challenges. Non CDS regions can exhibit a much noisier periodogram than others, or high coefficients at frequencies other than 3 nt.
The choice of a window function (Section 2.3.5) or of a cutoff over the spectral coefficient is necessary to distinguish regions were ribosomes are translating from non-translated regions. Unfortunately, as with many other score-based approaches, such cutoffs are strictly data-dependent, and can greatly vary for different datasets. A more general and statistically principled approach to detect the 3nt pattern in P-sites profiles was represented by the multitaper method (Section 2.3.5). The main advantage of the multitaper method is its statistical test for significance of frequency components, which clearly identified the presence of 3nt periodicity in CCDS profiles, despite a marked distortion of the frequency spectrum due to windowing.

Inspired by these early results, we decided to globally investigate the performance of the multitaper method in identifying translation.

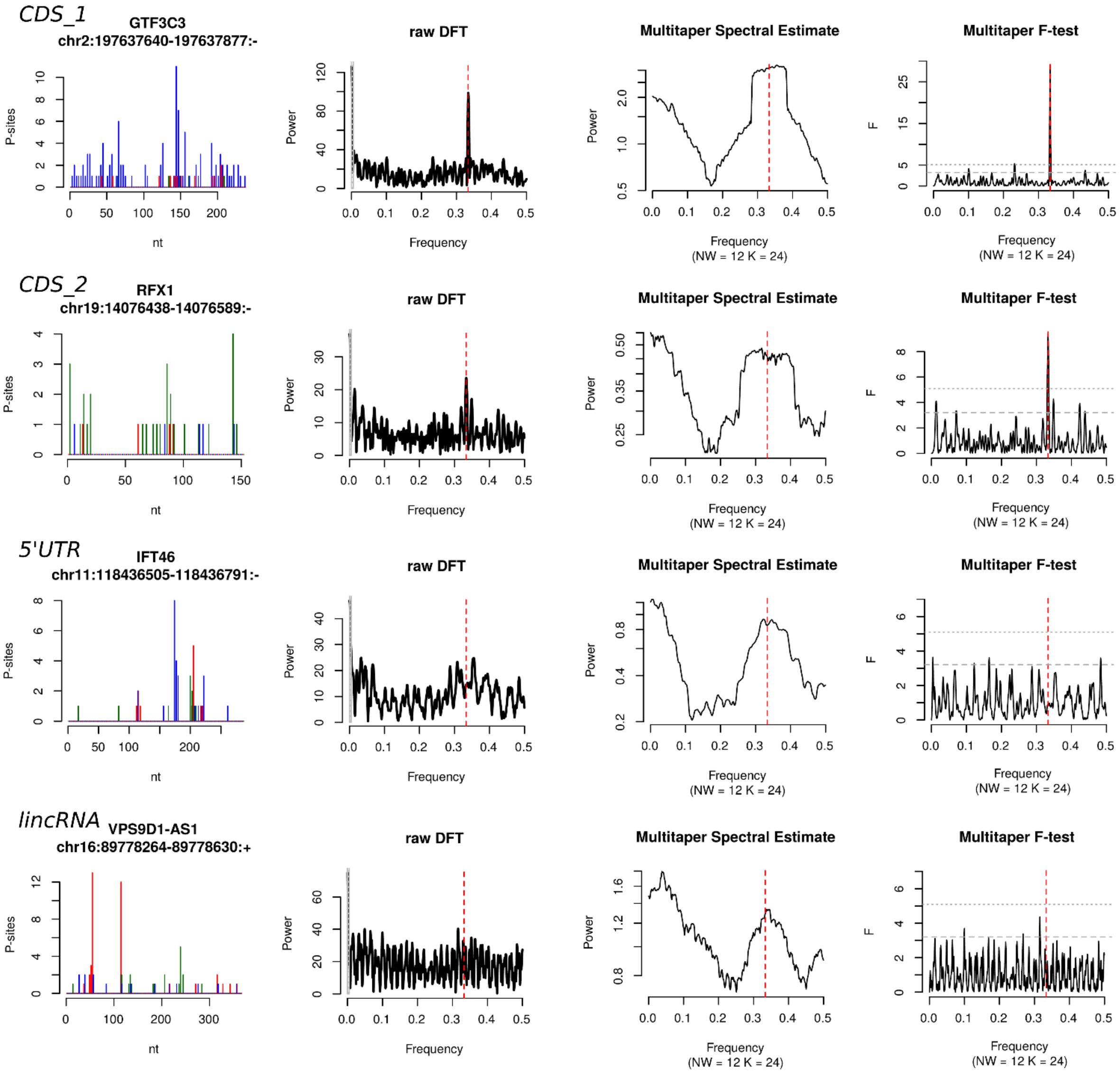


Figure 21: Spectral analysis of individual exonic P-sites profiles. P-sites profiles are shown on the left, the output of the raw DFT is shown in the middle left, the spectrum obtained by the multitaper in the middle right, while the F-value for each frequency bin is shown on the right. Dashed vertical lines around 3nt frequency. For the F-test plot, dashed horizontal lines represent p-values of 0.05 and 0.01.

3.1.2 On sensitivity and specificity

As discussed in Section 2.3.5, the multitaper analysis requires 2 parameters, the number of tapers to use and a time/frequency resolution parameter. To understand the influence of these two parameters we applied the multitaper method on P-sites profiles of CCDS exons, for different length and coverage values. To further test the validity of the multitaper-derived p-values, we ran the tests on randomly shuffled P-sites profiles from the same CCDS exons (Appendix B.4). As shown in Figure 22, little to no sensitivity improvements were observed by using more than 24 tapers, in all the length/coverage categories. Most importantly, at a p-value cutoff of 5%, exactly 5% of the shuffled P-sites profiles exhibited a significant 3nt component in all categories and tests (Figure 22), confirming the validity of the significance values reported by the multitaper test.

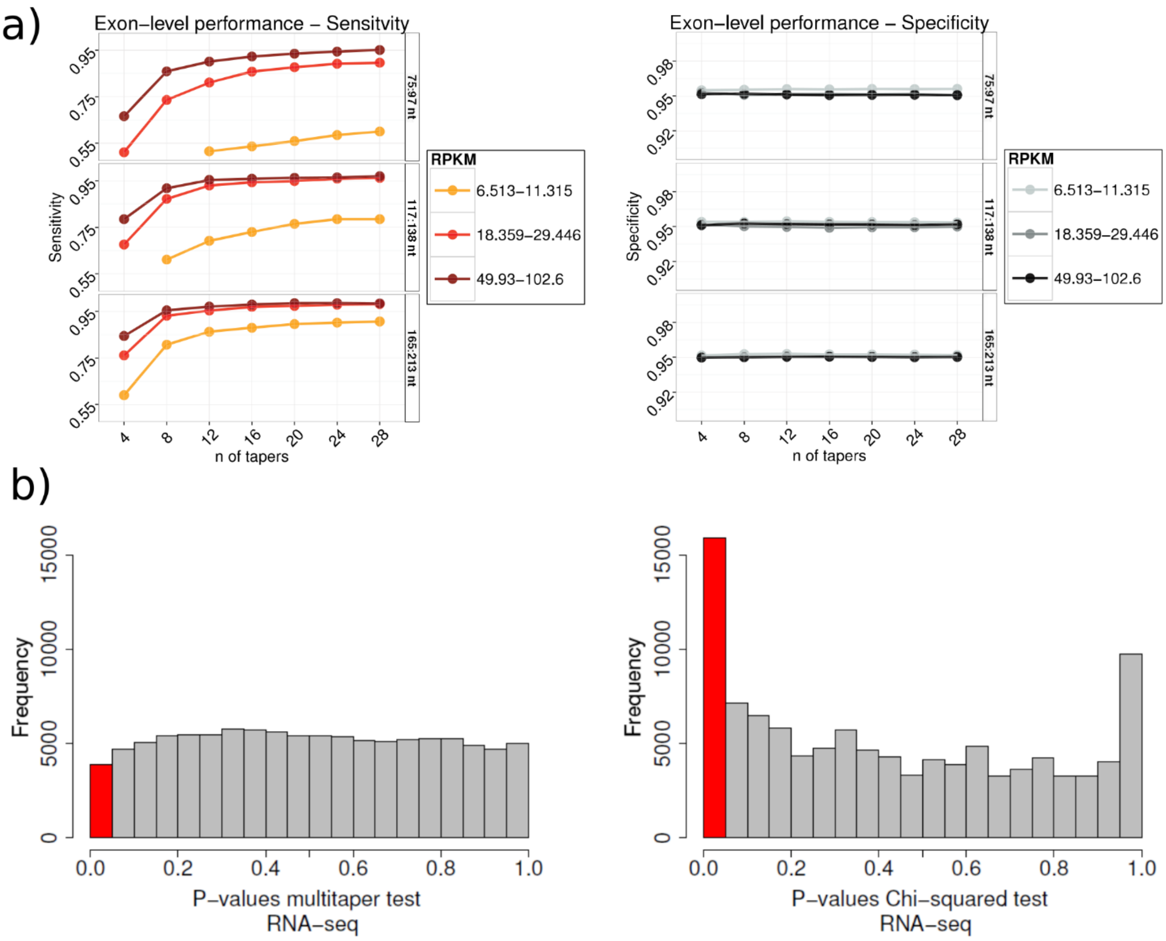


Figure 22: Sensitivity and specificity of the multitaper method. In a) a sensitivity analysis (left) is shown: using different tapers, the fraction of 3nt-periodic exons (p-value at 3nt <0.05) is shown for different CCDS exonic lengths, and for different coverage values. A similar plot is shown for specificity (right), where the y-axis indicates the fraction of simulated non-periodic CCDS exons (p-value at 3nt >0.05). Simulated profiles were obtained randomly shuffling the P-sites positions in each exon. As an additional proof for specificity, in b) the histogram of p-value for RNA-seq profiles in CCDS exons is shown, for the multitaper test (left) and Chi-squared test (right). The red bar highlights the fraction of p-values <0.05.

As shown in Figure 22, the multitaper test achieved very good sensitivity on CCDS exonic profiles of different length and coverage values, even for exons spanning from 75 to 97 nt of length. This analysis can also point out the level of resolution that we can achieve when trying to identify translation on small regions, such as uORFs or other short ORF categories (Section 2.3.6). The high sensitivity of the multitaper led us to investigate whether our False Positive Rate on real sequencing data would be misinterpreted when looking at our shuffled profiles. A uniform distribution of signal, as the one obtained by randomly shuffling P-sites positions, does not represent an ideal case to test the specificity of our approach. Sequencing data is far from uniform in any assay, where sequencing artefacts and protocol-specific biases result in non-uniformity of the coverage signal[120]. To further confirm the specificity of the multitaper method in detecting 3nt periodicity in Ribo-seq data, we decided to apply it to profiles derived from RNA-seq. For comparison, we used a Chi-squared test for frame preference, which tests against the assumption of uniformity of signal in the 3 frames, the same assumption behind the proposed ORFScore method[130] (Section 2.3.6). A strong skew in the p-values distribution for the Chi-squared test on RNA-seq profiles shows how the assumption of uniformity of signal is not adequate when dealing with sequencing data and would lead to a high number of false positive calls. On the other hand, the distribution of p-values for the multitaper test shows a desirable uniformity, showing again excellent specificity in detecting 3nt periodicity.

To further investigate the advantages of using a significance metric (the multitaper p-value) derived from spectral analysis, we compared the multitaper test with other Fourier transform coefficients, estimated with and without the use of different windows (see Section 2.3.5). We applied these different metrics on P-sites profiles and RNA-seq profiles (as a negative control) of CCDS exons of different length and coverage, and derived accuracy metrics to evaluate their performance in identifying translation (Figure 23). We calculated the Area Under the Curve (AUC, measuring overall performance) and sensitivity at 5% False Positive Rate for the different approaches. Once again, the p-values from the multitaper test could efficiently separate RNA-seq profiles from P-sites profiles, outperforming other metrics, including the spectral coefficients calculated by the multitaper method itself, which are distinct from the significance values (Appendix A).

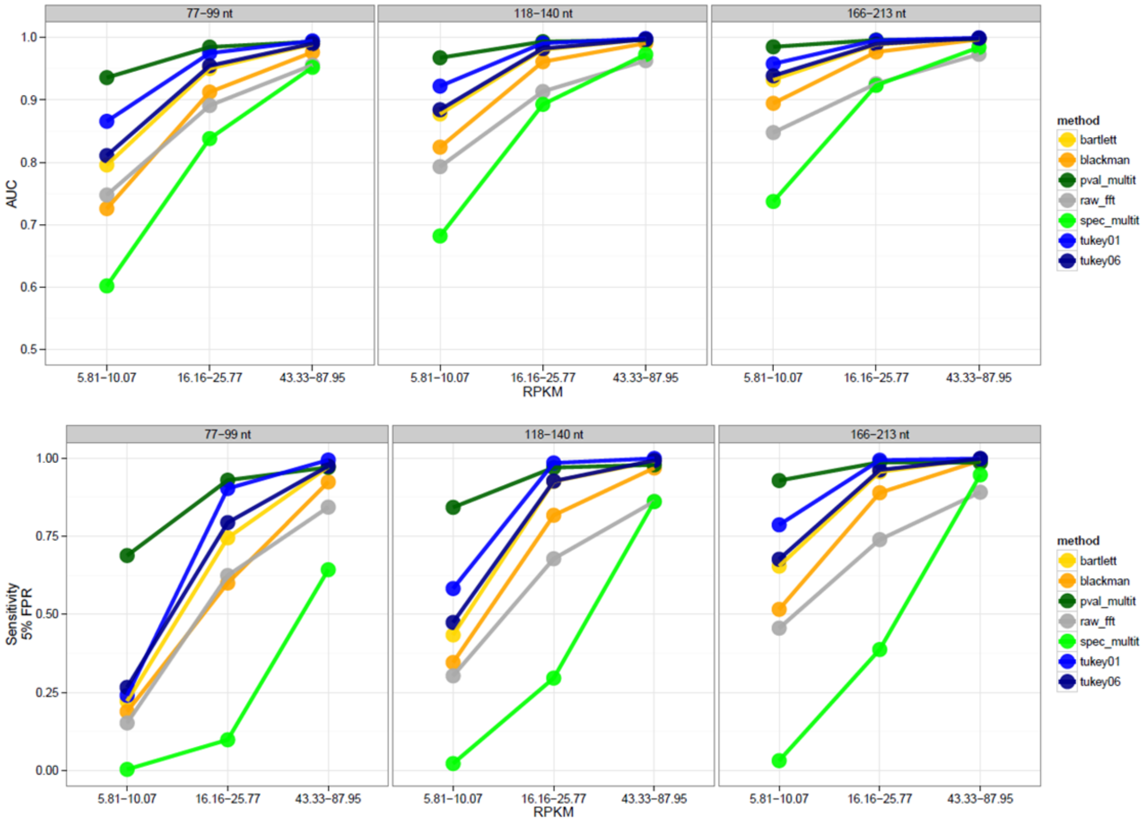


Figure 23: Comparison between the multitaper method and other windowing approaches. To estimate sensitivity and specificity, exonic RNA-seq profiles were treated as negative control, and exonic P-sites profiles as positive. At varying DFT coefficient (or p-value) cutoffs, the number of RNA-seq profiles and P-sites profiles retained was used to calculate performance metrics, using different approaches. In the top panel AUC values are shown, while in the bottom panel sensitivity values at 5% False Positive Rate are depicted. For all length categories, the p-value from the multitaper test (in dark green) outperforms other metrics. bartlett: Bartlett window; blackman: Blackman window; pval_multit: p-value at 3nt using the multitaper method (24 tapers); raw_fft: Raw DFT coefficient; spec_multit: Coefficient of the multitaper method (24 tapers); tukey: Tukey Window. For the Tukey window, two values of the alpha parameters were used (0.6 and 0.1). Different exonic lengths and RPKM values (x-axis) are shown.

Due to the excellent sensitivity and specificity of the multitaper in detecting 3nt periodicity, we decided to use it to identify the ensemble of translated ORFs in the transcriptome.

3.1.3 The RiboTaper strategy to identify translated ORFs

As shown in Figure 24, the first step in our analysis pipeline (named RiboTaper) is about parsing a genome fasta file and a GTF file (Section 2.3.1) to create sequence tracks and BED files for different

exonic regions, differentiating between coding exons (CDS regions), non-coding exons in coding genes (e.g. UTRs), and exons in non-coding genes. Importantly, we decided to use only transcripts annotated as part of the CCDS[95] annotation and part of the APPRIS[96] set of annotated transcripts, to limit the analysis only on well annotated transcript structures (Section 2.3.1).

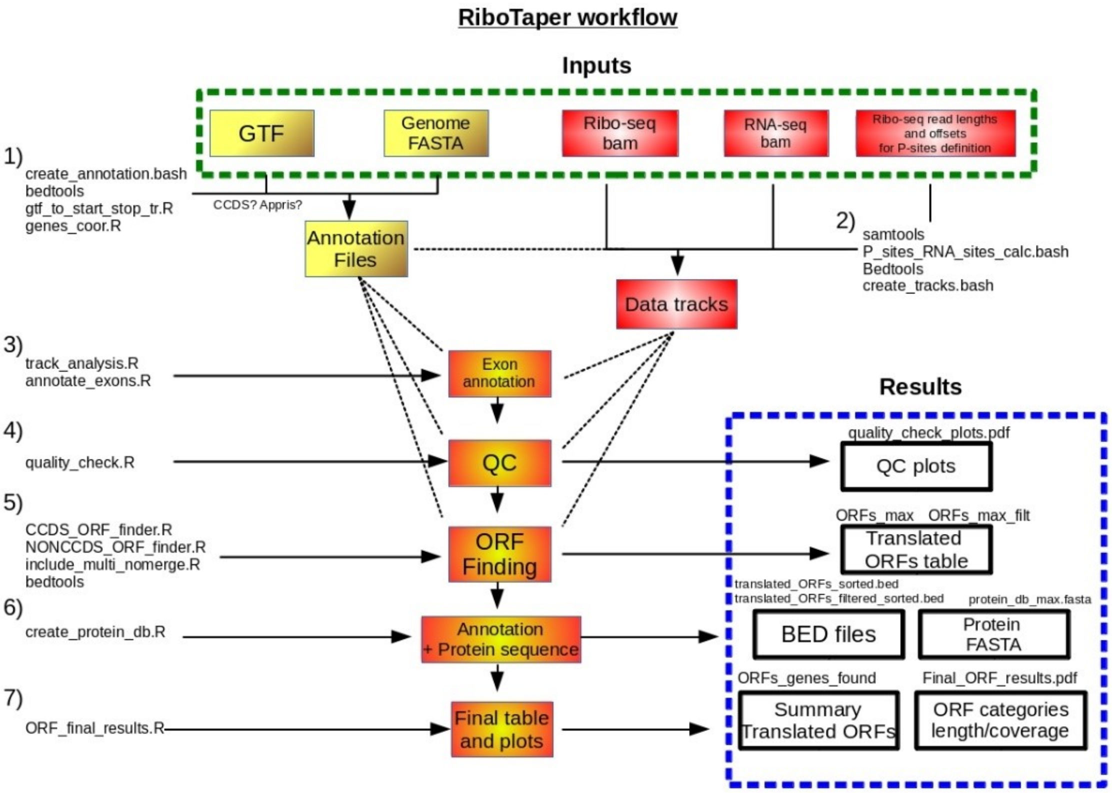


Figure 24: The RiboTaper workflow. From input files to output files, dependencies between steps are depicted, together with the analysis scripts involved.

Ribo-seq and RNA-seq alignment files are filtered to contain only the primary alignment per each read, to limit possible artefacts derived from multi-mapping reads (Section 2.3.2). The Ribo-seq and RNA-seq filtered alignments are then intersected with the annotation files to create data tracks needed for the next analysis steps. Of utmost importance is the definition of which Ribo-seq read lengths must be used to infer P-sites positions, together with their corresponding distance cutoffs from the 5' ends (Figure 20, Table 1, Supplementary Fig. 1). A separate program (*create_metaplots.bash*) is also provided to produce such aggregate profiles for different read lengths. The exonic data tracks are then analyzed with the multitaper method to evaluate sensitivity on different exonic length and coverage values, and the P-sites positions are compared to the annotated frames to measure the precision of the P-sites frame definition. Additional statistics, as the number of total reads, and frame precision on different genomic regions are also provided in this step. Subsequently, exonic tracks are merged according to the annotated transcript structures to create transcript tracks.

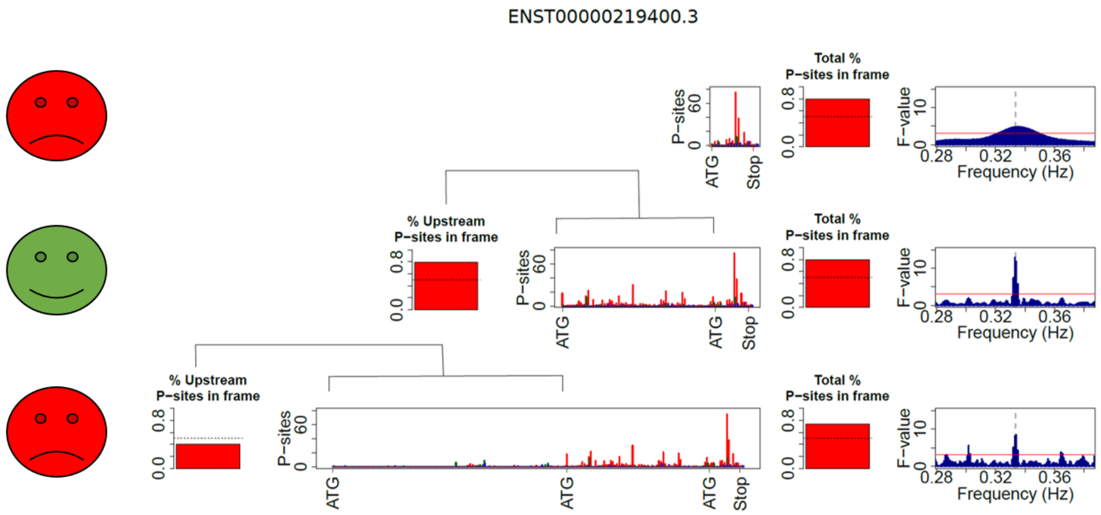


Figure 25: RiboTaper *de novo* ORF-finding strategy. For the shown transcript, the ORF structure in the middle is chosen, as its AUG preserves P-sites in frame when compared to the first ORF. The ORF in the bottom is discarded, as its AUG does not contain additional P-sites in frame (<50%). All the depicted ORFs show overall frame preference (>50%) and 3nt periodicity (on the right, red bar corresponding to p-value 0.05).

For each transcript, each pair of consecutive AUG-stop codon (ORF) is tested for its 3nt periodic pattern using the multitaper method, in all the three possible frames (p-value for multitaper test at 3nt periodicity <0.05). ORFs with less than 50% of in-frame P-sites are then excluded. In case of multiple possible start codons, we choose the most upstream in-frame AUG with more than 5 P-sites positions (>50% in-frame) between it and its closest neighbor AUG (Figure 25). In case of multiple transcript isoforms harboring the same ORF, the transcript with the highest number of RNA-seq reads was chosen.

ORFs are then annotated based on their transcript position and overlap with known CDS regions (Figure 26):

ORFs_ccds -> ORFs in CCDS genes, overlapping known CDS regions
non-CCDS coding ORFs -> ORFs in non-CCDS genes, overlapping known CDS regions
uORFs -> ORFs in CCDS genes, not overlapping with any CDS exon, upstream the annotated ORF
dORFs -> ORFs in CCDS genes, not overlapping with any CDS exon, downstream the annotated ORF
ncORFs -> ORFs in non-CCDS genes, not overlapping with any CDS exon

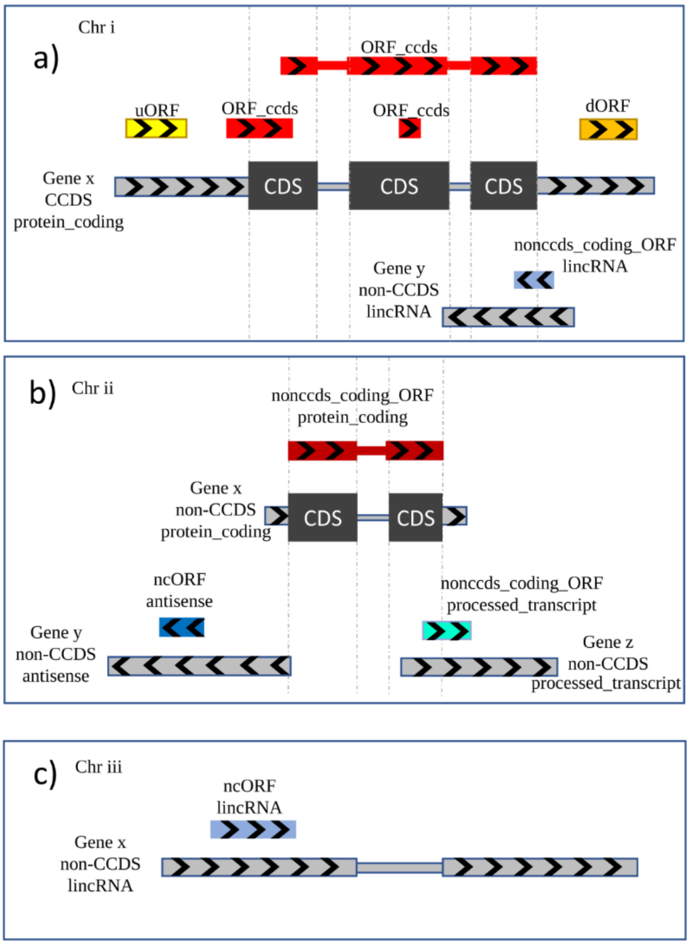


Figure 26: Schematics of RiboTaper ORFs annotation. In a) a "uORF" is defined as upstream the annotated start codon and non-overlapping any coding exon, while different "ORFs_ccds" are overlapping annotated coding exons. A "dORF" is defined as downstream the stop codon and not overlapping any coding exon. Shown also a lincRNA ORF overlapping a coding exon, therefore annotated as "nonccds_coding_ORF". In b) a "nonccds_coding_ORF" in a non-CCDS protein coding gene, defined as overlapping a coding exon. A "nonccds_coding_ORF" in a processed transcript gene is also present. A ncORF is defined as an ORF in a non-CCDS gene not overlapping any coding exon, here in an antisense gene. In c) a ncORF in a lincRNA gene.

Furthermore, to limit the effect due to multi-mapping alignments, a filtered set of ORFs was created including ORFs with <30% of the Ribo-seq coverage supported by multi-mapping reads only.

Note:

All the analyses in this manuscript are performed on the filtered set of RiboTaper-identified ORFs. Filtering was disabled only for the creation of the custom protein database.

Summary tables containing ORF positions, number of P-sites and RNA-seq reads per ORF are then created, together with BED files for the detected ORFs and in-silico translated protein sequence per each ORF. Using the RiboTaper method, we sought to detect translated ORFs in new and published Ribo-seq datasets.

3.2 Identification of actively translated ORFs in a human cell line

Contribution Statement:

Ribo-seq in HEK293 was performed by Neelanjan Mukherjee, Emanuel Wyler and Antje Hirsekorn, supervised by Markus Landthaler. Conservation analysis was performed by Benedikt Obermayer and Lorenzo Calviello. Mass-spec data analysis was performed by Henrik Zauber and Lorenzo Calviello, supervised by Matthias Selbach. Lorenzo Calviello ran the RiboTaper method and performed the rest of the data analysis, supervised by Uwe Ohler. Material appearing in this Section has been copied or adapted from our publication[145].

3.2.1 Known and novel ORFs across a wide expression range.

We ran RiboTaper on our HEK293 dataset, using the GENCODE 19 GTF with the CCDS and APPRIS tags.

We identified a total of ~21,000 ORFs spread over ~14,000 genes, over a wide range of expression levels. Such detected ORFs display 3nt periodicity, and multiple ORFs can be present in one single transcript (Figure 27).

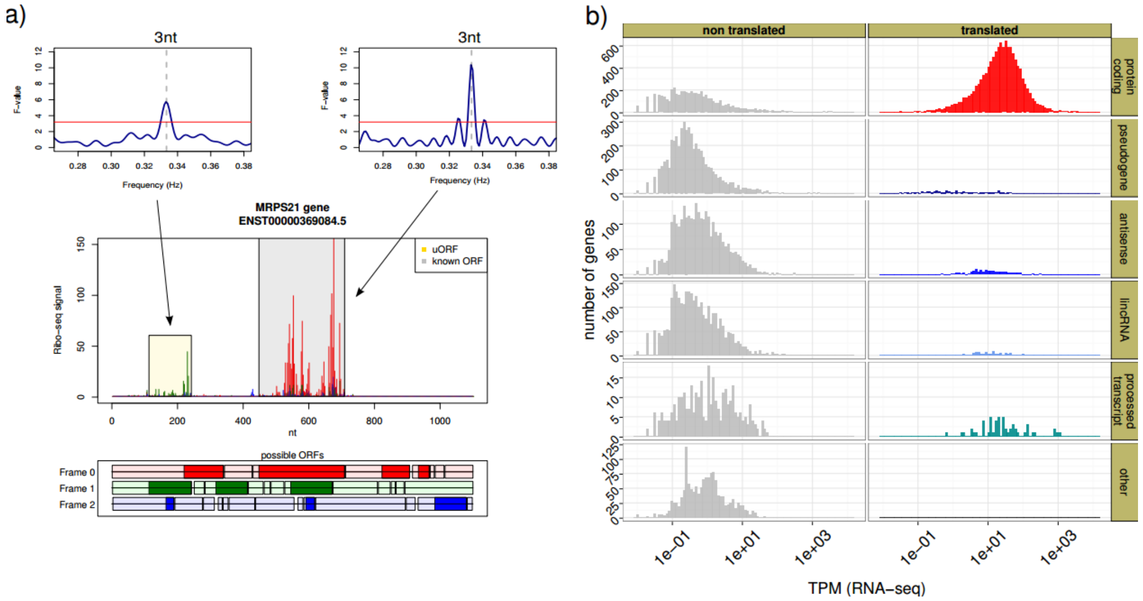


Figure 27: RiboTaper-detected ORFs in HEK293, across gene biotypes and expression values. In a) an example of two ORFs in a protein-coding transcript: 3nt periodicity (top) is capture by the multitaper test; the P-sites profile is shown in the middle, while the ensemble of all possible ORFs (dark colors) is shown at the bottom. In b) a gene-level summary of RiboTaper-detected ORFs: genes are divided based on containing a translated ORF or not; TPM values for RNA-seq, indicating steady-state gene expression, are depicted on the x-axis. The distribution of expression values for each gene biotype is shown. TPM values were calculated using RSEM.

Results were analyzed by aggregating ORFs by gene, as RiboTaper was not designed to resolve the mixture coming from different RNA isoforms per gene. As expected, the vast majority of detected ORFs belonged to transcripts from protein-coding genes, which showed a wide distribution across expression values, again confirming the excellent sensitivity of our approach. Few non-coding biotypes also contained translated ORFs, despite representing a minor fraction of the detected translatome (Figure 28).

Most of ORFs in protein-coding genes overlapped known CCDS coding regions, belonging to >11,000 CCDS protein-coding genes; 369 non-CCDS protein-coding genes were identified as harboring translated ORFs. We detected >600 genes with translated upstream ORFs (uORFs, Figure 27) and 54 genes with downstream ORFs (dORFs). We also identified ORFs in 504 non-coding genes (ncORFs), mainly belonging to pseudogene, antisense and long intergenic non-coding RNA (lincRNA) biotypes.

The detected ORFs categories showed different length and coverage profiles, with ORFs_ccds being the longest ORFs and the most covered by P-sites positions, and uORFs representing the shortest ORF category (median = 78 nucleotides). While the normalized coverage was similar for different ORFs in protein coding genes, the few detected dORFs displayed the lowest Ribo-seq signal. Antisense and lincRNA ncORFs showed a similar pattern with respect to both length and coverage values, while pseudogenes and processed transcripts ncORFs showed a similar pattern to protein coding genes, with more sustained coverage and longer ORFs.

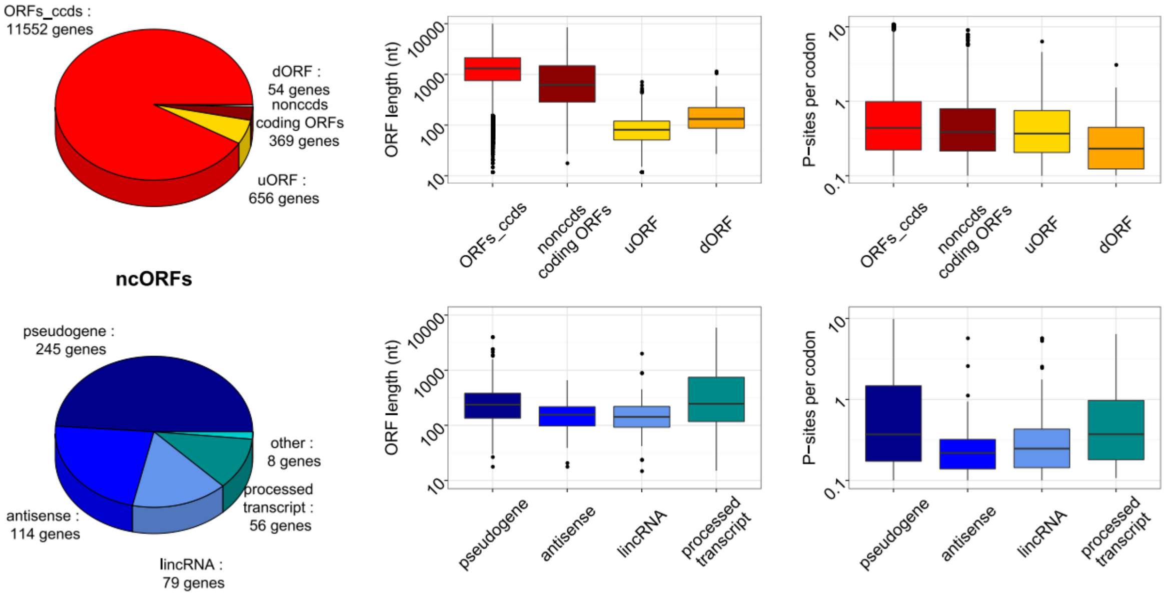


Figure 28: ORFs categories identified by RiboTaper. Shown at the top left corner is the number of protein-coding genes harboring different ORF categories. On the right, length and coverage values (expressed in P-sites per codon) are plotted for the different ORF categories. At the bottom left, statistics about ncORFs gene biotypes (bottom left) are shown, while at the bottom right length and coverage statistics

To validate our ORF detection strategy for our HEK293 data, we compared the genomic coordinates of our detected ORFs with AUG translation initiation sites defined by QTI-seq[84] (Section 2.2.3, Appendix B.5), also performed in HEK293 cells, or the annotated start codons (Figure 29). By plotting the distance between RiboTaper or QTI-seq start sites and the reference annotation, we could evaluate the agreement between the two sets. Compared with the reference, 149 upstream initiation sites were detected by both QTI-seq and RiboTaper, mostly corresponding to uORF start codons (Figure 29). 52 internal starts were identified by both QTI-seq and our method. However, approximately 1,000 QTI-seq AUG start codon candidates did not overlap with either annotated or RiboTaper-defined start codons (see Discussion). As a measure of between-lab reproducibility, we applied RiboTaper to CHX-treated Ribo-seq data from the same study, and we observed that more than 99% of RiboTaper-identified CCDS genes were also found in our data. Agreement dropped to 68% for lincRNAs/antisense genes with ncORFs and 47% for uORF-containing genes, possibly due to their relatively short length, but also low expression levels.

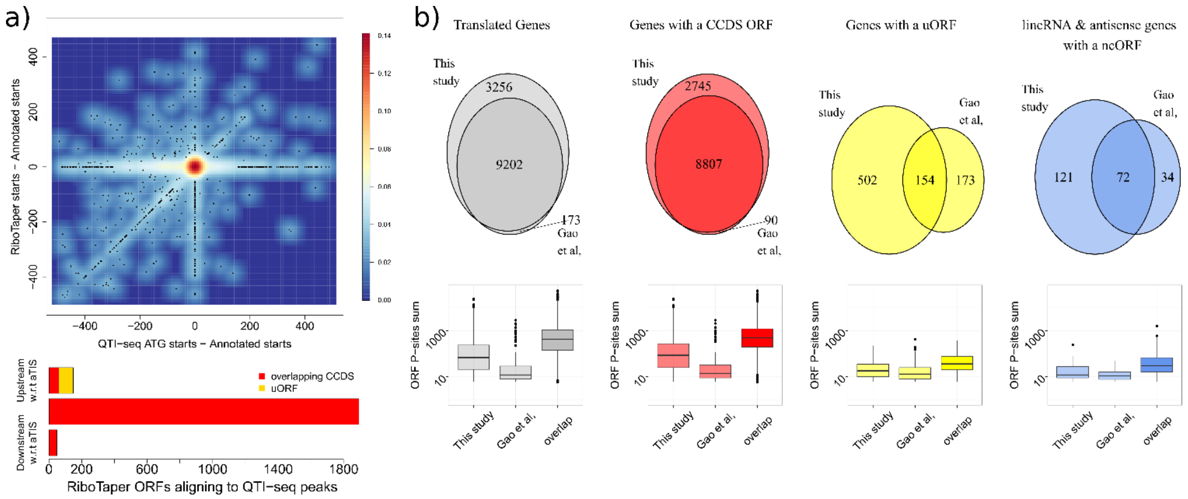


Figure 29: QTI-seq comparison ad between-samples reproducibility. a) Scatterplot (top) of the distance between reported QTI-seq AUG peaks and annotated start codons (x-axis) vs distance between RiboTaper ORFs starts and the annotation (y-axis). Each dot represents the start codon of one ORF. Barplot (bottom) of the number of start positions identified by both QTI-seq and RiboTaper with respect to the annotated translation initiation site (aTIS). b) Overlap (top) and coverage (bottom) of ORFs identified in the Gao et al, data set compared to our data set, split by ORF category.

As an additional confirmation of protein-coding-like behavior of the Ribo-seq signal in the detected ORFs, we calculated FLOSS scores[138] (Section 2.3.6) for all our ORF categories and compared them to known protein-coding regions. In all ORF categories, FLOSS scores confirmed the *bona fide* coding capacity of our detected ORFs (Figure 30).

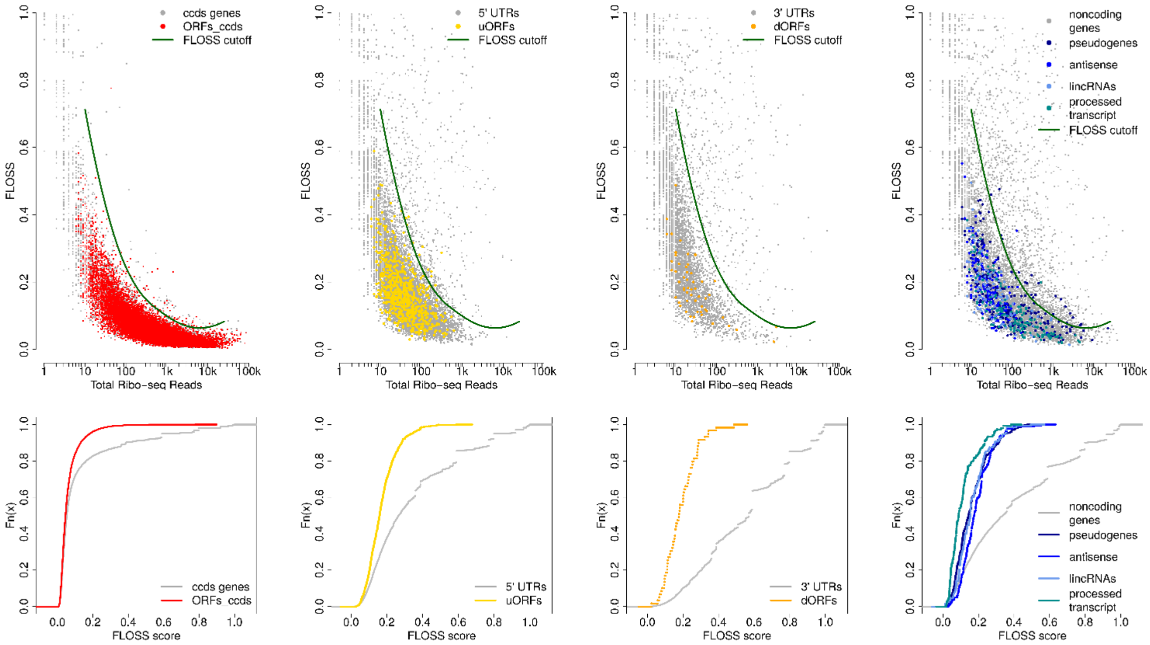


Figure 30: FLOSS scores for ORFs identified by RiboTaper. Shown are (top) FLOSS scores with their cumulative distributions for CCDS genes and ORFs ccds (left), 5'UTRs and uORFs (middle left), 3'UTRs and dORFs (middle right), non-coding genes and different ncORFs categories (right). Cumulative density function (CDF) plots for the same values are shown in the bottom. FLOSS values and cutoffs were calculated as in Ingolia et al, 2014. Low FLOSS scores indicate a protein-coding-like fragment length distribution.

Of 110 human genes with entries in the manually curated uORFdb[180] database, 63 were detected in our dataset. 12 of our predicted uORFs-containing genes mapped to these entries, which referred to 20 different studies that reported on the possible roles for these uORFs, as a regulatory translation event or via the encoded small peptide product[48].

Moreover, to demonstrate the general applicability of RiboTaper, we applied it to Ribo-seq data coming from experiments in the zebrafish embryo[130]. We identified thousands of coding ORFs and few dozens of ncORFs (Table 2). Among the identified ncORFs was the recently discovered ORF in the lincRNA *toddler*[181], which encodes a small polypeptide morphogen essential for zebrafish embryonic development (Supplementary Figure 2).

3.2.2 Distinct evolutionary conservation patterns in different ORF categories

As pointed out in Section 2.3.7, coding sequences are by far the most conserved element in the genome, as they define the amino acidic sequence, and thus the function, of proteins. Given the distinct features we observed for the different ORF categories and biotypes described in Section 3.2.1, we decided to investigate their evolutionary signatures.

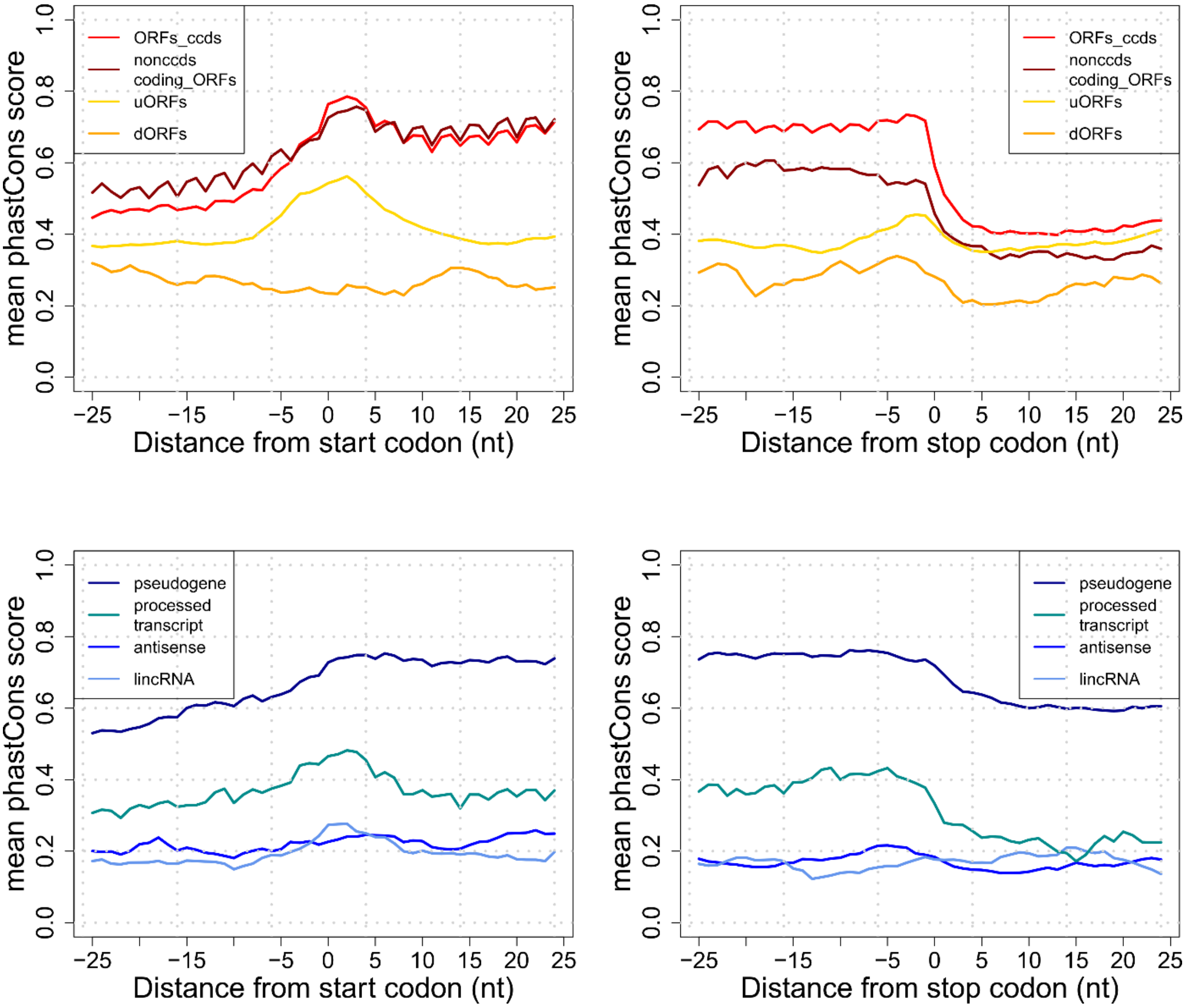


Figure 31: Nucleotide conservation at ORF boundaries. Shown are PhastCons values for ORFs in protein-coding genes (top) and for ncORFs (bottom).

When looking at the nucleotide-level conservation defined by PhastCons[158] (Section 2.3.7) in a 50nt window around start and stop codons, we observed distinct patterns for the different ORF categories/biotypes (Figure 31): ORFs overlapping known CDS regions displayed a peak of conservation around start and stop codons, together with high nucleotides conservation inside their coding sequence. uORFs, on the other hand, displayed a high conservation values around start and stop codons, but low conservation inside the putative CDS, thus showing evolutionary selection on the genomic positions rather than on the encoded protein product. ncORFs categories showed overall low levels of nucleotide conservation. High nucleotide conservation for pseudogenes ncORFs might represent a “mirror effect” (due to difficulties in the mapping) from their protein-coding parent genes. Although much weaker, an enrichment of nucleotide conservation at the start codon can also be observed for some ncORFs categories, like lincRNAs and processed transcripts. The latter also showed protein-coding-like high nucleotide conservation, in line with their protein-coding like features for length and Ribo-seq coverage (Section 3.2.1).

To investigate the presence of selection on the encoded protein sequence of the identified ORFs, we examined the coding potential of different ORF categories by means of hexamer frequencies (CPAT) [162], codon substitution frequencies (PhyloCSF) [159] and dN/dS ratio[157], [160] (Figure 32, Appendix B.8). To limit the influence of sequence length or nucleotide conservation in our estimation of coding potential, we included, as a control for each ORF category, a set of ORFs (only defined on their sequence) from the non-coding transcriptome (UTRs and non-coding transcripts) with matching length and nucleotide conservation[157].

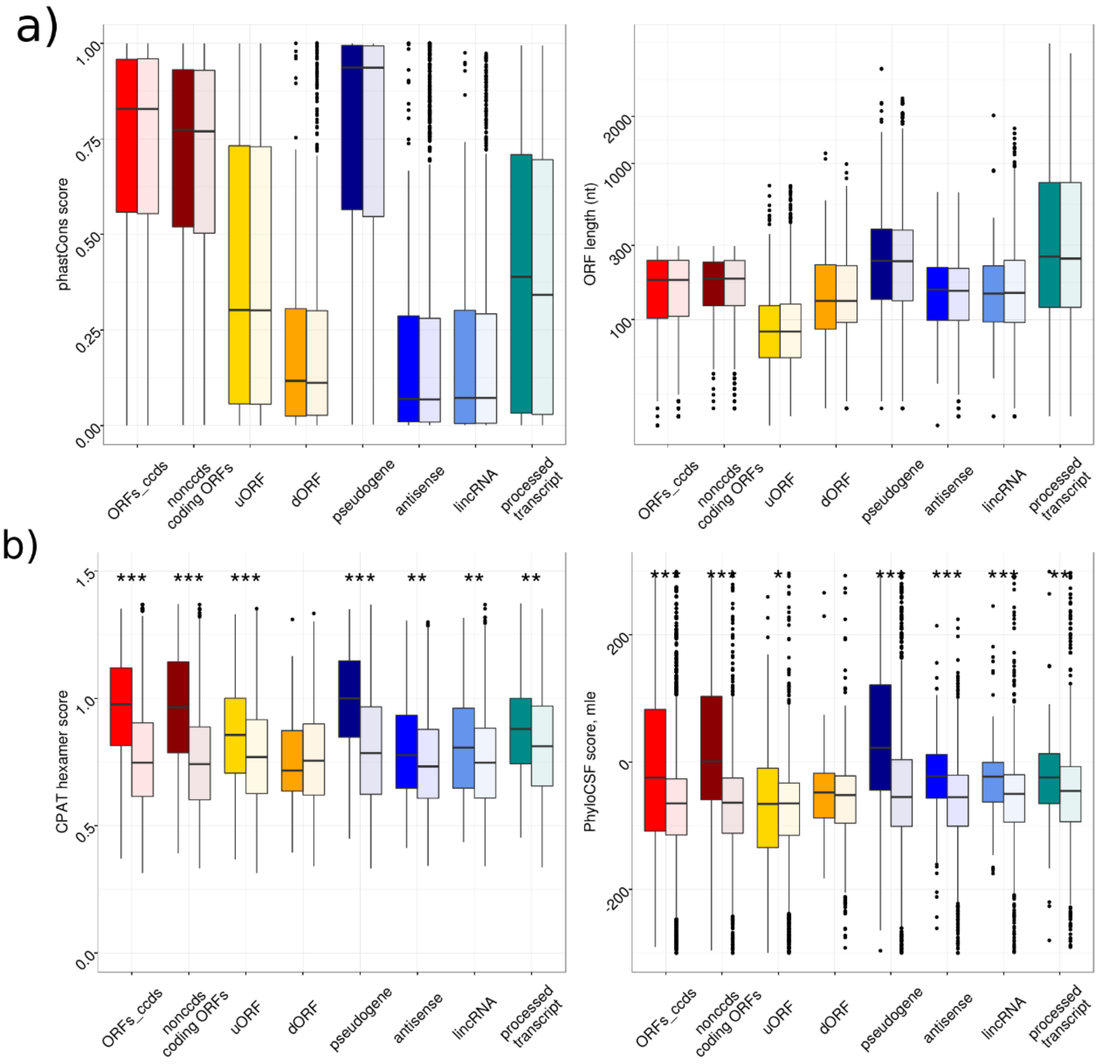


Figure 32: Coding potential of different ORF categories. In a) phastCons scores (left) and length (right) are shown for RiboTaper-identified ORFs (dark colors) and controls (dim colors). No significant difference in terms of length and nucleotide conservation (PhastCons) was found between detected ORFs and controls. In b) are shown the scores from CPAT (left) and PhyloCSF (right, used in the -mle mode). * = p-value below 0.05; **= p-value below 0.01; ***= p-value below 0.001, Wilcoxon-Mann-Whitney test. ORFs_ccds and nonccds coding ORFs were selected if shorter than 300 nucleotides, to match negative controls ORFs.

For ORFs overlapping known CDS regions, as well as for processed transcript ncORFs, nucleotide conservation was accompanied by high hexamer scores and a depletion of non-synonymous SNPs (dN/dS), indicating selection on the encoded protein sequence. In line with the nucleotide conservation analysis, uORFs showed low hexamer scores and no significant enrichment of synonymous substitutions (dN/dS). Additional ncORFs categories showed a positive trend for hexamer scores, but no depletion for non-synonymous SNPs, revealing a positive coding potential when compared to controls, but no detectable selection in the human population (Figure 33).

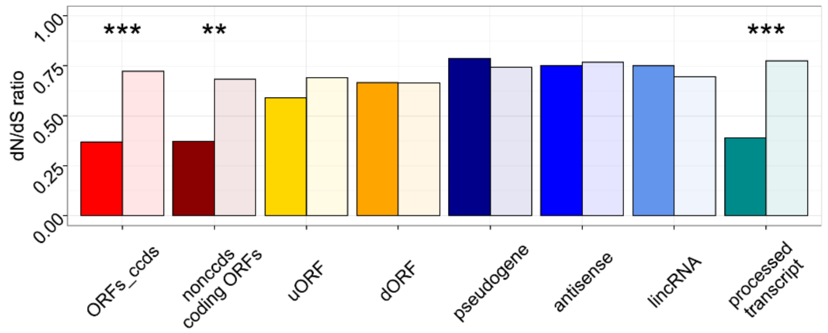


Figure 33: Positive selection for different ORFs in the human population. Shown is the dN/dS ratio for each ORF category and controls. Low values indicate positive selection of the encoded peptide sequence. Controls are shown in dim colors. **P < 0.01, ***P < 0.001 by chi-squared test, using as expected frequencies the values from the ORF control set (See Appendix B.8). CCDS ORFs and non-CCDS coding ORFs <300 nt long were used in this analysis, to match negative-control ORFs.

Taken together, ORFs detected by RiboTaper do not necessarily entail strong selection on the amino acidic sequence; however, selection on the genomic positions of some ORF categories suggests their possible regulatory roles.

3.2.3 The *de novo* identified translatome as a proxy for the cellular proteome.

The ORFs identified by RiboTaper covered annotation over more than 10,000 genes, and over a wide range of expression values (Figure 27), to an extent that it might be used as an effective proxy to define the ensemble of proteins present in a cell. To evaluate this, we created a custom database from our set of identified ORFs to match the spectra of a recent HEK293 tandem mass spectrometry dataset[182] (Appendix B.9).

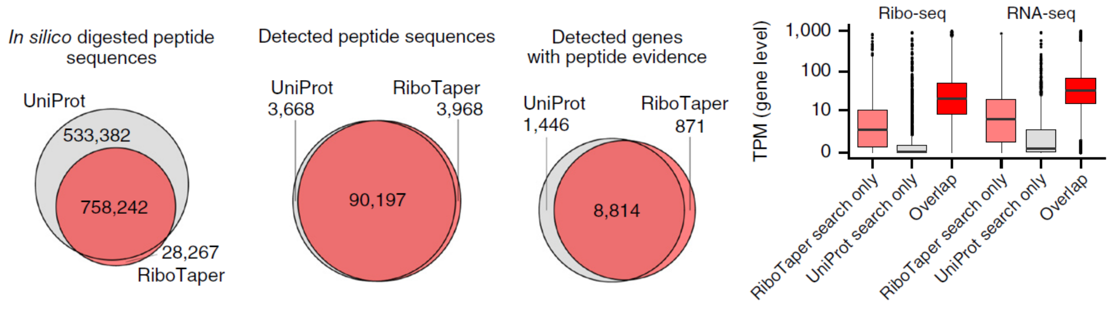


Figure 34: Proteome-wide detection of translated ORFs. Overlap between the protein databases derived from Uniprot or RiboTaper ORFs. All possible peptide sequences (left), detected peptide sequences and detected genes (middle) are shown. Gene expression levels (Ribo-seq and RNA-seq) for genes showing peptide support in the two search strategies (right, RiboTaper vs Uniprot).

Compared to the human entries in the Uniprot database (rel. October 2014), the RiboTaper dataset included ~59% of the possible tryptic peptides (Figure 34). Moreover, our set included around 2% of additional peptides not present in Uniprot. When matching spectra from MS/MS to peptide sequences, we confirmed >90,000 RiboTaper peptide sequences, belonging to >8,000 genes, similar to the results of the full Uniprot search. Over 3,900 peptide sequences were found only by our custom search but not using the Uniprot database (1% FDR, Figure 34, Appendix B.9). The RiboTaper-only peptides matched more spectra than UniProt-only peptides, despite being shorter and with lower matching scores (Supplementary Figure 3). On the other hand, our custom search missed >3,600 peptides matched by Uniprot sequences. However, we found little evidence for expression or translation for most of the genes encoding for Uniprot-only peptides, suggesting that those identifications may derive from erroneous spectral matching, or very stable peptides derived from genes which are no longer active.

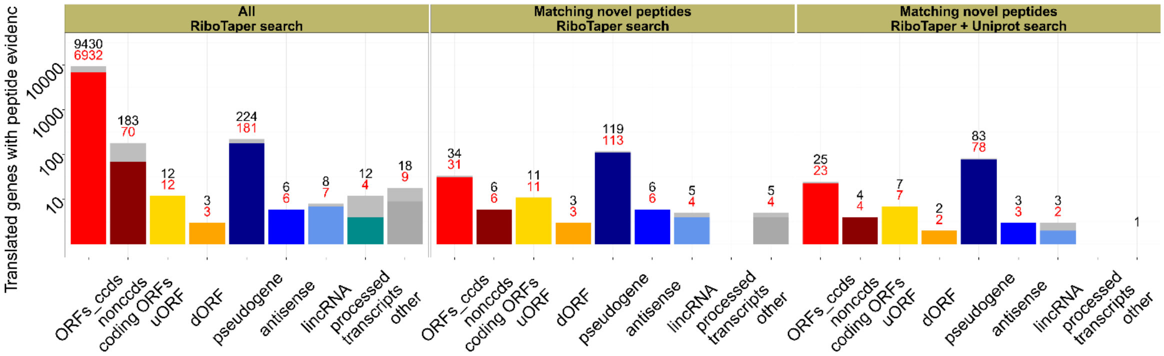


Figure 35: Translated ORFs with peptide evidence. The number of genes containing at least one RiboTaper identified ORF with peptide evidence, divided by ORF category (left). Genes containing at least one RiboTaper identified ORF with novel peptide evidence (not found in Uniprot, human entries, rel. October 2014) using the RiboTaper database (middle; 191 ORFs in 189 genes) or a database of the union of RiboTaper and Uniprot entries to exclude potential cross-matches (right; 157 novel peptides mapping to 129 ORFs in 127 genes). Red numbers indicate evidence from uniquely assigned peptides.

The RiboTaper ORFs with peptide support mapped mostly to known protein-coding genes, with very few exceptions (Figure 35). Pseudogene ncORFs represented the biggest class of ncORFs with unique peptide evidence, an observation which probably needs additional confirmation (See Discussion). Among the novel matches, we identified 2 peptides belonging to a uORF in the MIEF1 gene, previously reported as a novel ORF with high coding potential[183]. Additional ORFs with peptide evidence encompassed 3 dORFs and some ncORFs, located in conserved and non-conserved genomic regions of the genome. In total, 228 identified peptide sequences were not annotated in Uniprot. In most cases, the novel identified peptides mapped uniquely to their respective ORFs, even when merging the Uniprot and RiboTaper databases (Figure 35).

Taken together, these results show how Ribo-seq data can be used as an effective proxy to define the cellular proteome.

3.3 ORF detection with an improved protocol in *Arabidopsis thaliana*

Contribution Statement:

Ribo-seq in *Arabidopis thaliana* and western blot validations for novel peptides was performed by Polly Yingshan Hsu, supervised by Philip N. Benfey. Conservation analysis was performed by Fay-Wei Li and Carl J. Rothfels. Ribo-seq data analysis was performed by Lorenzo Calviello, supervised by Polly Yingshan Hsu, Uwe Ohler and Philip N. Benfey. Material appearing in this Section has been copied or adapted from our publication[184].

3.3.1 Analysis of an improved, high-resolution Ribo-seq protocol

A modified Ribo-seq protocol (Appendix B.11) has been developed by Polly Yingshan Hsu to investigate translation in roots and shoots of *Arabidopsis thaliana*. Inspired by early studies describing the importance of the composition of extraction buffer during the Ribo-seq protocol[81], four buffers at different ionic strengths were tested in their ability to produce high-quality ribosomal footprints. Despite showing very similar counts per gene and similar read lengths distributions, the four buffers produced footprints at different levels of sub-codon resolution (Figure 36).

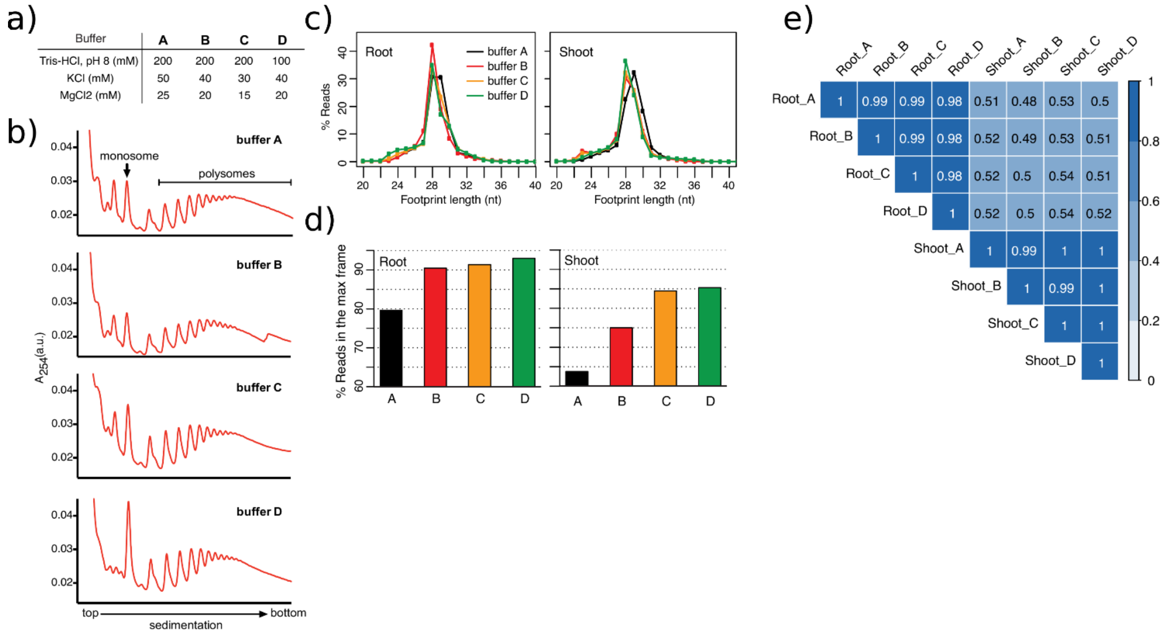


Figure 36: An improved Ribo-seq protocol. In a) the 4 tested conditions for the polysome/lysis buffer in the Ribo-seq protocol. In b) the polysome profiles using the 4 different buffers. In c) read length distributions for the different buffers, while in d) the percentage of reads in frame is shown. In e) correlation between the different buffers using Ribo-seq counts on protein-coding genes.

To evaluate the ability of this modified protocol in capturing translation at high resolution, Ribo-seq was performed again with optimized conditions (using buffer D) in roots and shoots, and results were compared to other published Ribo-seq datasets in *Arabidopsis thaliana*[185]–[187] (Appendix B.11).

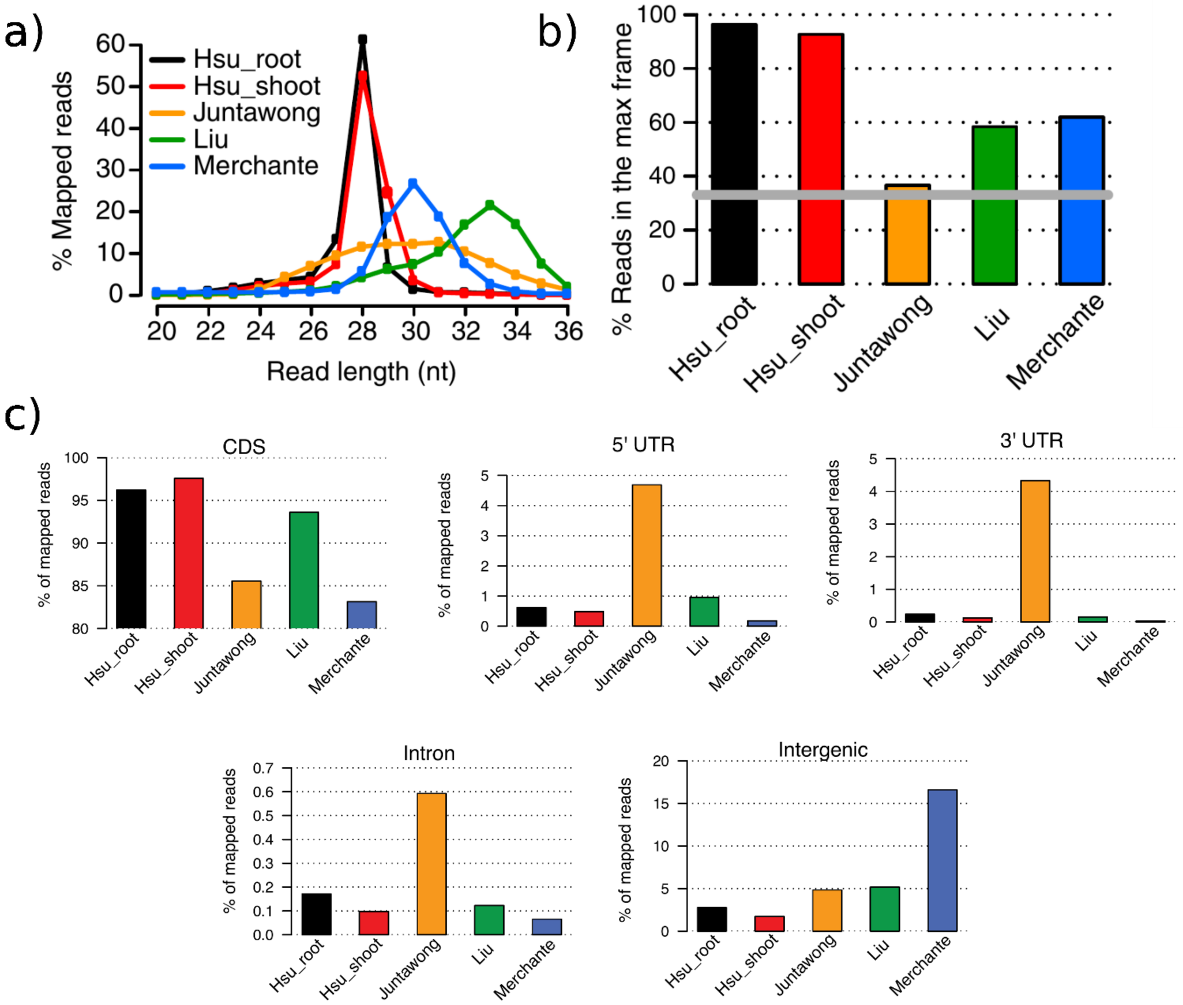


Figure 37: Comparisons of different Ribo-seq protocols. a) Length distribution of ribosome footprints among the current study (Hsu_root and Hsu_shoot), and three other published datasets. b) Percentage of Ribo-seq reads in the coding reading frame. Data were extracted from the meta-gene analysis using 28nt footprints in which most of the datasets display the best 3nt periodicity. The gray line marks 33%, which is the percentage of reads expected if there is no enrichment in any frame. In c) mapping statistics for the different datasets, across different genomic features.

The different datasets showed profound differences when looking at read length distribution and mapping statistics on different genomic regions (Figure 37): the Hsu dataset showed a very narrow read length distribution, peaking at 28nt, while the other datasets consisted of longer read lengths and broader read length distributions, especially for the Juntawong dataset. When looking at mapping statistics, we observed >95% of the Hsu libraries mapping to known CDS regions, similar to the Liu libraries, and a lower proportion (~85%) for the Merchante and Juntawong datasets. When looking at different genomic regions, the Juntawong data showed a substantial number of reads mapping to introns and UTR regions, higher than the other datasets. The Merchante libraries showed an enrichment of reads mapping to intergenic regions.

When looking at the sub-codon resolution of the different datasets, we again observed marked differences: most of the read lengths in the Hsu dataset showed a clear initiation peak followed by a substantial frame precision, with >95% reads mapping to the translated frame. The Merchante and Liu datasets showed moderate frame preference, with ~60% reads in frame, followed by the Juntawong dataset, which showed very little frame preference despite the preference of a clear initiation peak, at the exact 5’end of each read length (Figure 38).

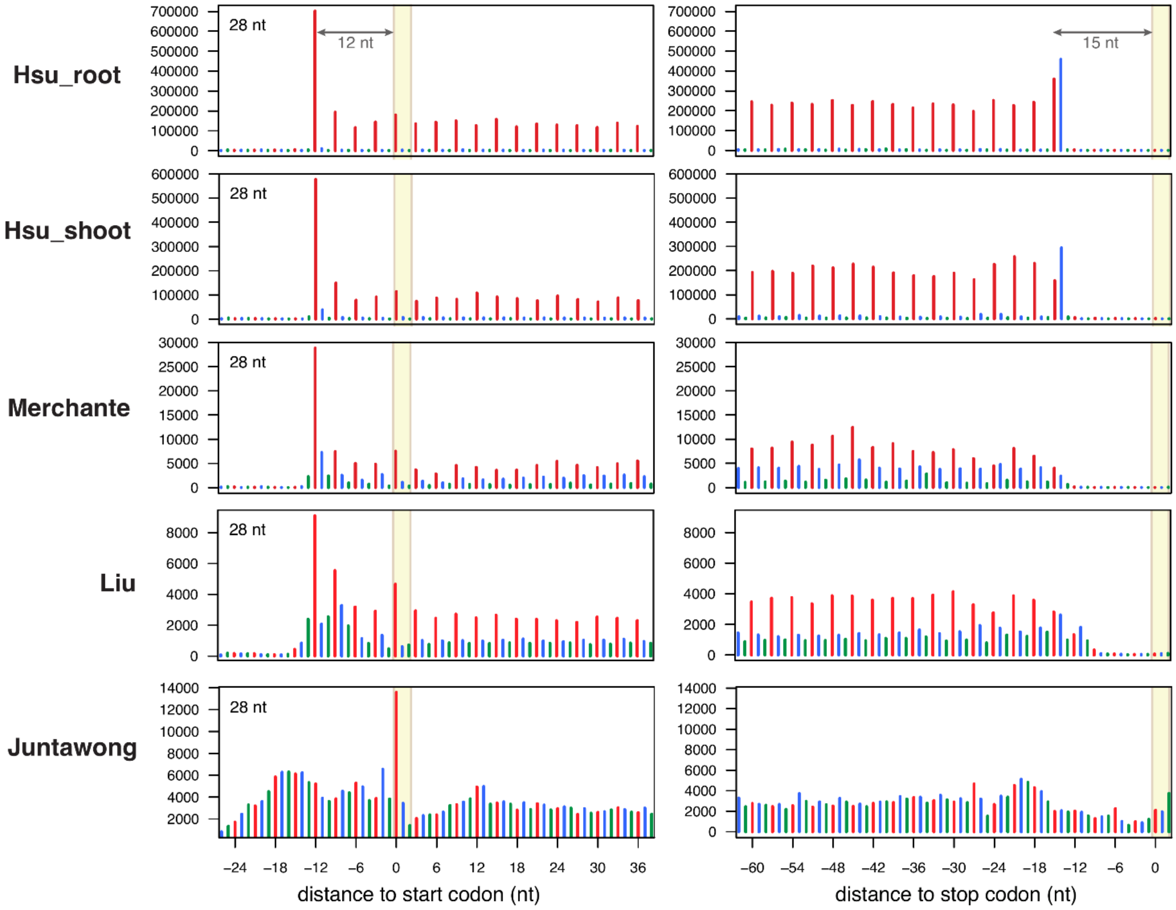


Figure 38: Sub-codon resolution of different Ribo-seq datasets. Aggregate profiles are shown around annotated start (left) and stop (right) codons, for the different datasets.

To investigate how such marked differences in the libraries would impact the detected translatome, we applied RiboTaper (Section 3.1). Using the Hsu data, we could confidently identify translation in >15,000 protein-coding genes, also at moderate sequencing depths (25M mapped reads). The Liu and Merchante datasets enabled the identification of 12,000 and 10,000 genes, while we could identify translated ORFs for less than 1,000 genes using the Juntawong dataset. Results were similar when checking the number of genes with translated ORFs against their expression values measured by RNA-seq (Figure 39): at an expression cutoff of 1 TPM (Section 2.3.3), we detected translation for >90% of expressed genes using the Hsu data, with percentages dropping for the other datasets. A similar trend is visible when looking at the detected ORF lengths compared to the annotation (Figure 39). The quality of the different datasets also influenced the detected ORF length. Using the Hsu data, we could capture (on average) >90% of the annotated ORFs lengths, even at low sequencing depths which compared favorably to the other datasets. Taken together, these results show how the new protocol compares favorably to other datasets and greatly improves ORF detection with RiboTaper.

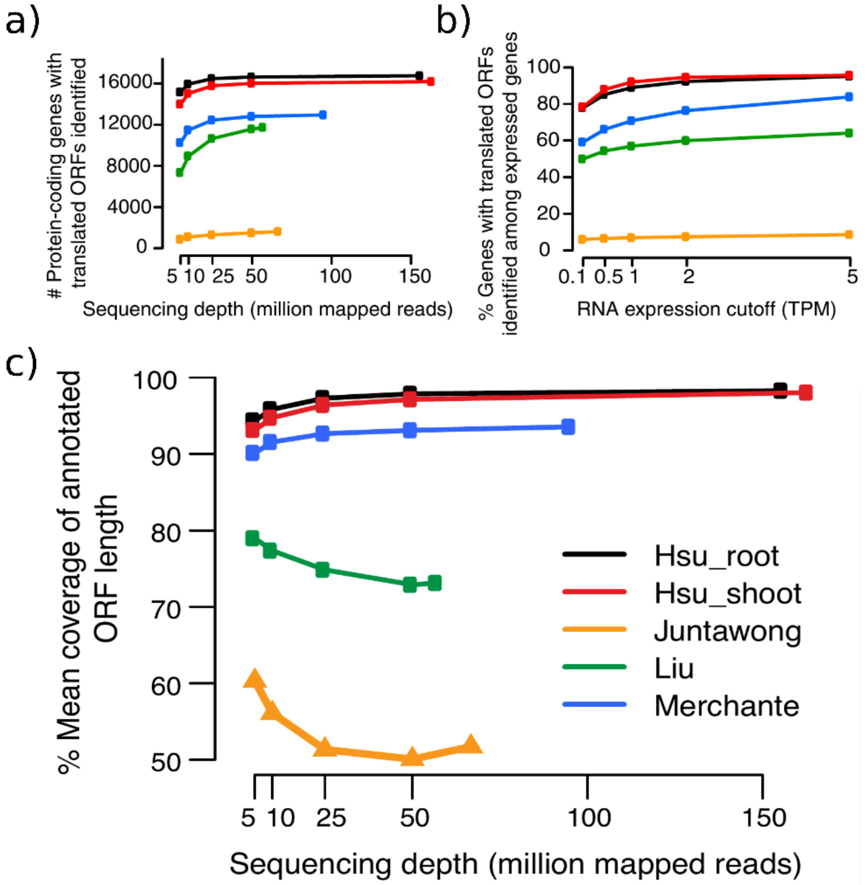


Figure 39: ORF detection with RiboTaper across datasets. In a) the number of protein-coding genes harboring translated ORF(s), as a function of sequencing depth, for the different datasets. In b) the percentage of genes harboring translated ORFs as a function of their steady-state RNA expression (25 Million reads were used for each dataset). In c) the average of annotated ORFs length (in percentages) captured by RiboTaper in the different datasets. Low numbers indicate short, truncated ORFs.

3.3.2 New, ultra-conserved ORFs in non-coding genes

More than 18,000 protein-coding genes harbored a translated ORF, including 187 uORFs and 10 dORFs. 44% of the detected uORFs overlapped with annotated CPuORFs, which are conserved uORFs thought to encode for functional peptides[188]. Moreover, we could detect ~100 ncORFs, divided in non-coding RNAs (ncRNAs), pseudogenes and transposable elements (Table 2).

	uORFs	ORFs_ccds	dORFs	ncRNAs	Pseudogenes	Transposable elements
Root	136	16,657	2	23	27	31
Shoot	87	16,107	8	14	14	40
Total	187	18,153	10	27	37	57

Table 2: RiboTaper-detected ORFs in *Arabidopsis thaliana*. Shown are the genes harboring translated ORFs, divided by ORF category/biotype. There are 27,416 protein-coding genes, 394 ncRNAs, 924 pseudogenes, and 3,903 transposable element genes annotated in TAIR10. The 89 known CPuORFs are annotated as protein-coding genes, and they were grouped with uORFs here.

Inspired by the high quality of our data, we decided to test whether these new coding elements produce stable peptides.

To investigate whether the detected novel ORFs code for stable small proteins *in planta*, we picked 4 candidates to be experimentally verified (Figure 40).

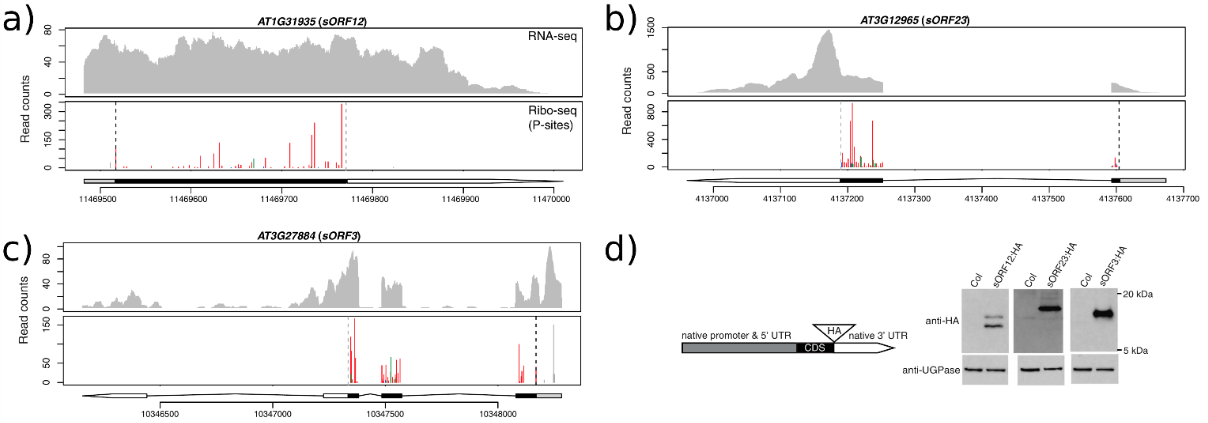


Figure 40: Experimental validation of ORF candidates. a)-c) RNA-seq and P-sites in ribosome footprints in root for three ORFs identified within annotated ncRNAs. The predicted CDS and 5' UTR are depicted as black and gray boxes, respectively. 3' UTR is represented by a white arrow. In d) a schematic diagram of HA-tagged constructs and western blot analysis of novel proteins produced by the three annotated ncRNAs in panels a)-c). Total protein in control plants (Col-0) and transgenic plants expressing individual HA-tagged proteins were isolated and analyzed with either anti-HA or anti-UGPase antibodies (loading control).

ORF candidates were cloned (including their UTRs) into a construct with a HA (hemagglutinin) tag at the end of their CDS, and subsequently transformed in *Arabidopsis*. Using western blots against the HA tag we could confirm the presence of 3 out of 4 candidates tested, and thus the high fidelity of our novel ORF candidates.

The experimental confirmation of stable peptides coming from the novel ORFs detected by RiboTaper led us to the investigation of whether these peptides may have an important function in *Arabidopsis*, or even in additional plant species. We surveyed 15 different plant genomes, including sequences from plant species very distant from *Arabidopsis* in the phylogenetic tree (Appendix B.14). Of the 19 single-exon ORFs in annotated ncRNA genes, 15 showed at least one homolog in other plant genomes, (Figure 41). We found a homologous sequence only in the genome of *Arabidopsis lyrata* for one candidate, while all the other ORFs had homologous genes throughout the *Brassicaceae* family. For six ORFs we found homologous sequences in other Eudicot or even Monocot species, with one ORF being conserved even in *Selaginella*, a plant genus diverged >400 million years ago from *Arabidopsis*[189].

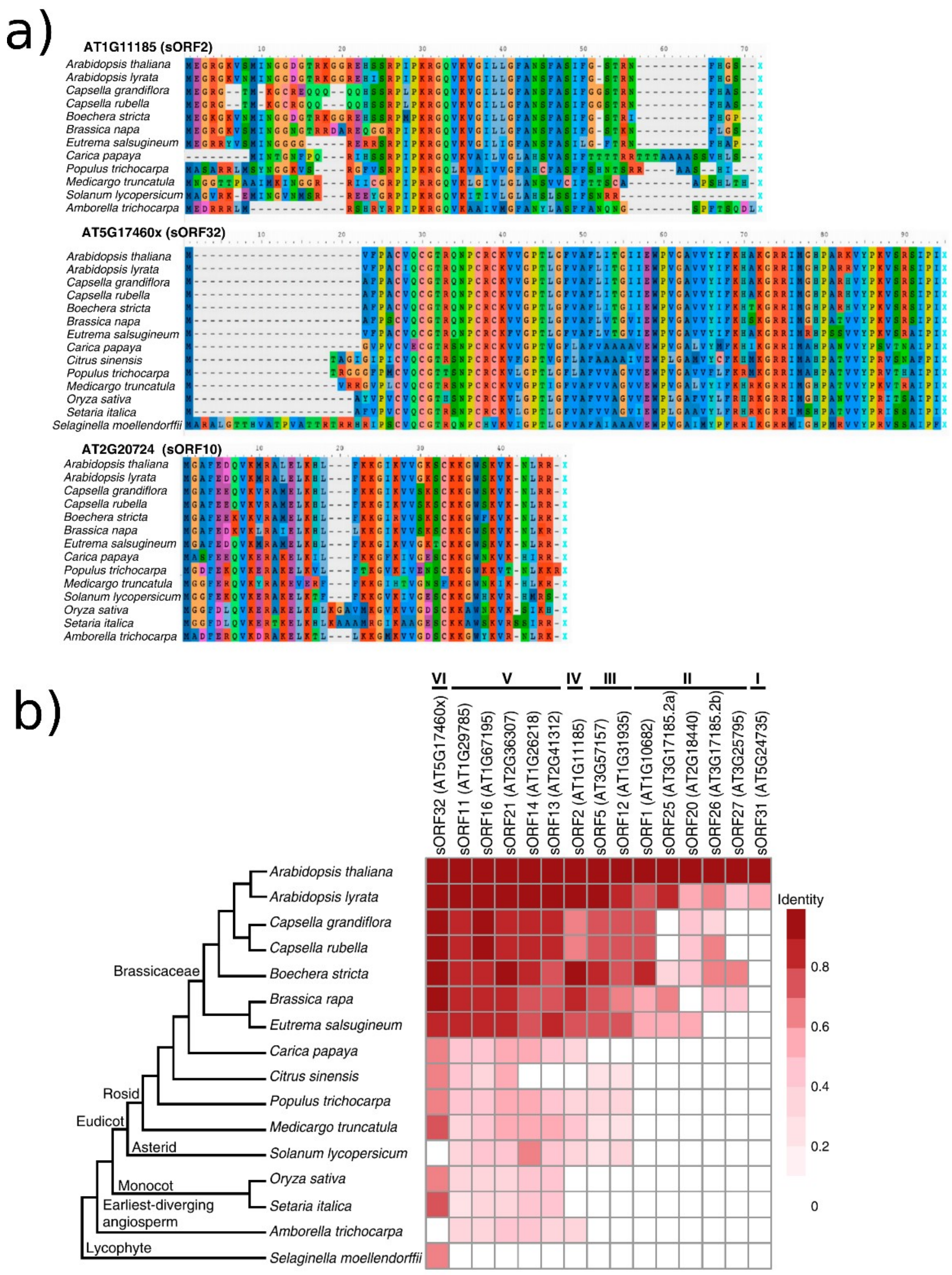


Figure 41: Evolutionary conservation of novel ORFs. In a) multiple species alignment of 3 novel ORFs identified in annotated ncRNAs. Amino acids with the same functional groups are shown in similar colors. In b) amino acid sequence identities between novel ORFs within annotated ncRNAs in *Arabidopsis thaliana* and their corresponding homologs in other 15 plant species. A phylogenetic tree showing evolutionary divergence is on the left. ORFs were grouped based on their homologs identified in other species (I to VI).

3.4 Annotating and quantifying the translated transcriptome

Contribution Statement:
Except for the Mass-spec database search, performed by Henrik Zauber and Matthias Selbach, all the analyses in this chapter were performed by Lorenzo Calviello, supervised by Uwe Ohler. All the results in this section are unpublished material.

3.4.1 The SaTAnn strategy

As shown in Section 3.1, we can use active ribosome elongation from Ribo-seq data to confidently identify translation. However, the complexity of the eukaryotic transcriptome poses a great challenge in understanding the full ensemble of synthesized proteins in a cell. RNA-seq techniques have uncovered the presence of multiple transcripts per gene, but the contribution of alternative splicing to protein diversity has remained elusive.

When attempting isoform-level quantification with Ribo-seq and RNA-seq, it is possible to observe many inconsistencies, with many transcripts exhibiting no expression but sustained level of ribosome occupancy, or vice versa (Figure 42). Many genomic loci will present a rich ensemble of possible transcript structures, thus resulting in difficulties in resolving the mixture of isoforms. Moreover, ribosomes do not map to 3'UTR regions, making the comparison between isoform-specific quantification estimates very challenging.

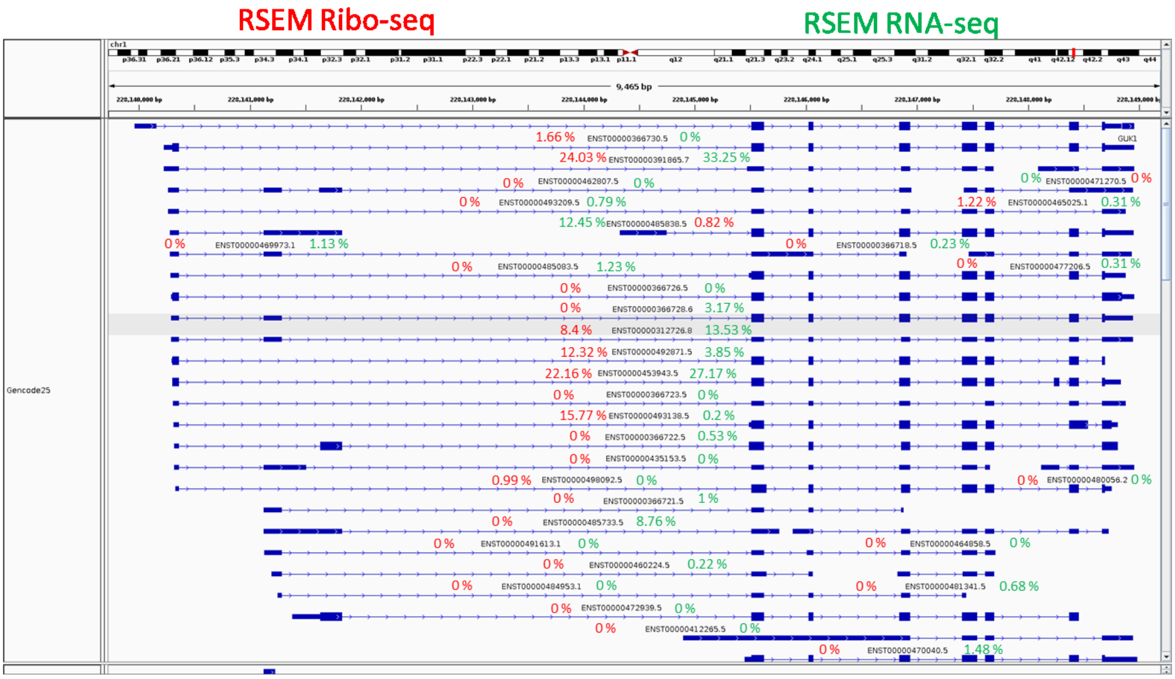

RSEM Ribo-seq
RSEM RNA-seq
chr1
p36.31 p36.21 p36.12 p35.3 p34.3 p34.1 p32.3 p32.1 p31.2 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 p11.1 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.2 q41 q42.12 q42.2 q43 q44
9,465 bp
228,140,000 bp
228,141,000 bp
228,142,000 bp
228,143,000 bp
228,144,000 bp
228,145,000 bp
228,146,000 bp
228,147,000 bp
228,148,000 bp
228,149,000 bp
Gencode25
GUK1
1.66 % ENST00000366730.5 0 %
24.03 % ENST00000391865.7 33.25 %
0 % ENST00000471270.5 0 %
0 % ENST00000462807.5 0 %
0 % ENST00000493209.5 0.79 %
1.22 % ENST00000465025.1 0.31 %
12.45 % ENST00000485838.5 0.82 %
0 % ENST00000469973.1 1.13 %
0 % ENST00000366718.5 0.23 %
0 % ENST00000485083.5 1.23 %
0 % ENST00000477206.5 0.31 %
0 % ENST00000366726.5 0 %
0 % ENST00000366728.6 3.17 %
8.4 % ENST00000312726.8 13.53 %
12.32 % ENST00000492871.5 3.85 %
22.16 % ENST00000453943.5 27.17 %
0 % ENST00000366723.5 0 %
15.77 % ENST00000493138.5 0.2 %
0 % ENST00000366722.5 0.53 %
0 % ENST00000435153.5 0 %
0.99 % ENST00000498092.5 0 %
0 % ENST00000480056.2 0 %
0 % ENST00000366721.5 1 %
0 % ENST00000485733.5 8.76 %
0 % ENST00000491613.1 0 %
0 % ENST00000464858.5 0 %
0 % ENST00000460224.5 0.22 %
0 % ENST00000484953.1 0 %
0 % ENST00000481341.5 0.68 %
0 % ENST00000472939.5 0 %
0 % ENST00000412265.5 0 %
0 % ENST00000470040.5 1.48 %

Figure 42: Example of transcript-level quantification on the GUK1 gene. RSEM transcript-level quantification on the *GUK1* gene. Shown are the percentages of total gene expression (using RNA-seq or Ribo-seq) for each isoform (only a subset of isoforms is shown).

We then decided to extend our ORF finding strategy to identify and quantify the ensemble of translated transcript isoforms in a cell, again using Ribo-seq data.

Our strategy, named Splice-aware Translatome Annotation (SaTAnn), is comprised of multiple steps, which consist in quantifying ORF usage in a subset of the annotated transcripts using Ribo-seq signal over their transcript positions (**Figure 43**).

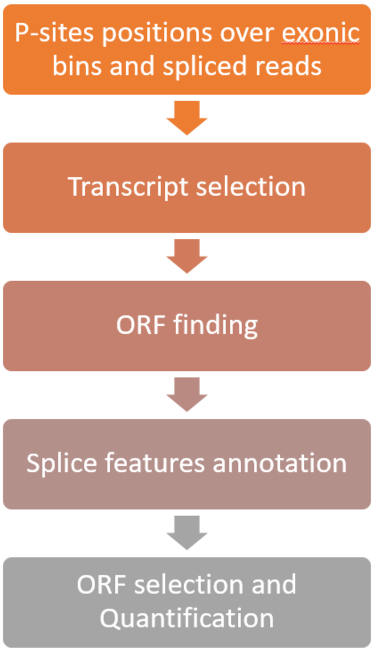

P-sites positions over exonic bins and spliced reads
Transcript selection
ORF finding
Splice features annotation
ORF selection and Quantification

Figure 43: The SaTAnn workflow. Multiple steps are depicted, from the assignment of Ribo-seq signal to transcript regions until ORF quantification.

A detailed explanation of the single steps is outlined below, together with our validation scheme which integrates multiple data sources from different technologies.

Transcript Filtering:

For each gene, transcripts are divided into different features (exonic regions and splice junctions), which can be unique or shared between different transcripts. A transcript feature can or cannot contain evidence from Ribo-seq. Spliced reads are used to extract evidence for splice junctions, while P-sites positions indicate evidence over exonic bins. Exonic bins are defined as in the DEXSeq strategy, by flattening the transcript structures and delineate common and shared regions[114].

Our strategy aims at selecting a small number of transcripts which contain all the features with Ribo-seq evidence, minimizing the number of structures with no evidence in their putative CDS. As transcripts are selected, we keep track of the explained features, which will be used to further select or filter transcripts. As we parse the list of annotated transcript structures, the following rules are employed for each transcript Tx_i:

1) Tx_i contains a new feature with evidence:

   Tx_i is selected and each previously selected Tx_j is re-analyzed:

   i) If all the features with evidence of Tx_j are in Tx_i, transcript j is filtered,

   ii) If Tx_j contains the same features with evidence of Tx_i, but more *internal features* with no Ribo-seq evidence, it is filtered.

2) Tx_i does not contain a new feature with evidence:

   Tx_i is not immediately selected, but it competes with each previously selected structure Tx_j.

   If all the features with evidence of Tx_j are in Tx_i, but Tx_i has less *internal features* with no Ribo-seq evidence, Tx_i is selected and Tx_j is filtered out.

Internal features are defined as the features contained between the first and the last feature with at least 1 Ribo-seq read, which should represent the translated exons. As Ribo-seq maps on 5'UTRs and within the coding regions, this approach cannot distinguish between 3'UTR isoforms (see Discussion).

As they lack Ribo-seq evidence, we hypothesized that the discarded transcript structures do not represent mature mRNA structures in HEK293. To validate our approach, we calculated expression values of selected and non-selected transcripts using deep, paired-end, strand-specific RNA-seq data, using RSEM[98] (Section 2.3.3) to calculate transcript-specific expression levels.

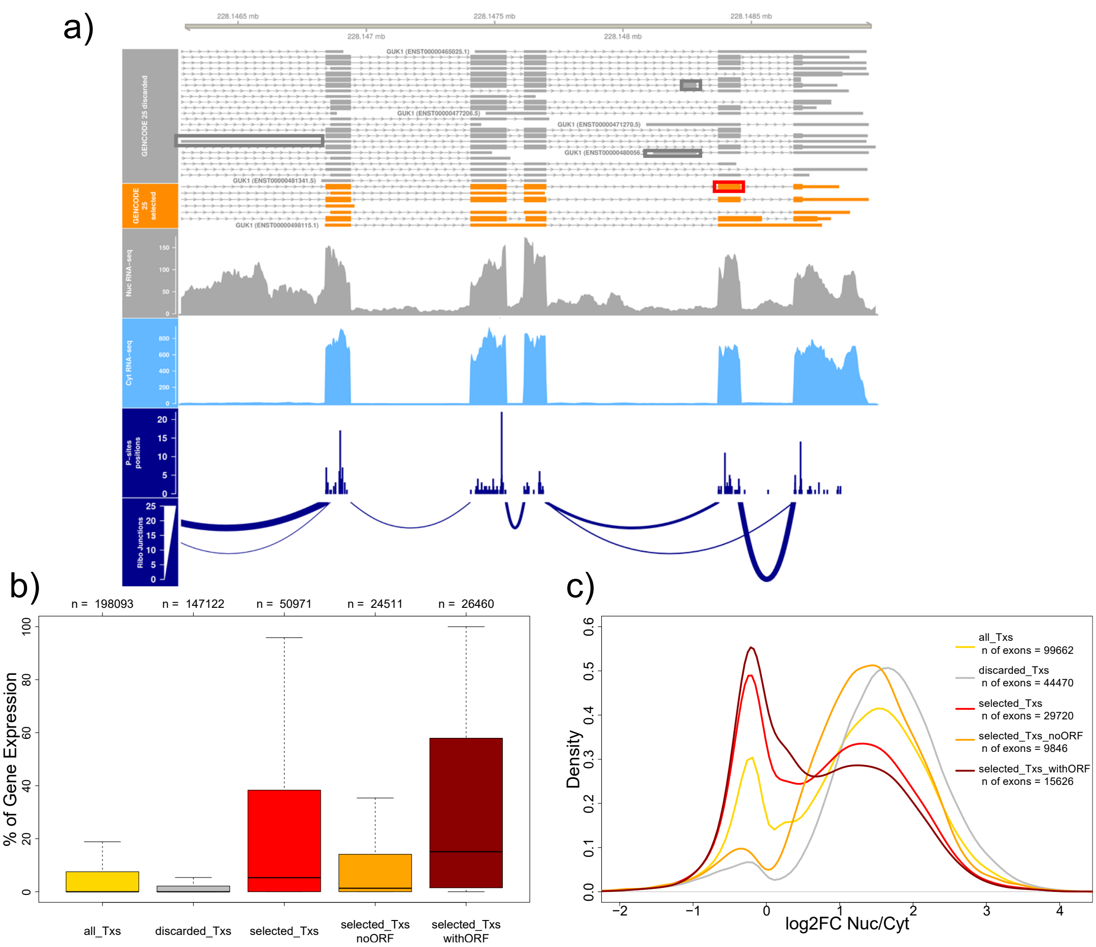


Figure 44: Transcript filtering. In a) an example gene (*GUK1*) with depicted discarded (top, in gray) and selected (orange) transcripts structures, together with tracks of nuclear (gray) and cytoplasmic (light blue) RNA-seq. Shown are also P-sites positions and junctions from Ribo-seq (dark blue). In gray boxes examples of exonic bins of discarded structures, while in orange an exonic bin of a selected structure. In b) isoform-specific expression values from RNA-seq (in % of total gene expression) for different transcript (txs) categories (all txs, discarded txs, selected txs, selected txs without a translated ORF, selected txs with a translated ORF). Outliers are not shown. In c) the distribution of nuclear vs. cytoplasmic localization of exonic bins of different transcripts categories. Negative numbers indicate cytoplasmic localization.

As shown in Figure 44, roughly 2/3 of the annotated transcript structures were discarded by our strategy, and showed little to no expression in RNA-seq data (median RNA-seq counts = 0), while the small subset of selected transcripts (n ~ 51,000) showed appreciable expression levels, confirming the validity of our selection strategy. An additional separation can be made between selected transcripts whether they harbor a translated ORF or not (discussed in next section), as translated transcripts show more sustained level of steady-state expression.

As translation is a cytoplasmic process, the selected transcript structures should represent *bona fide* cytoplasmic transcripts. Conversely, discarded transcript structures can represent RNA-processing intermediates which can be detected in the nucleus, but not in the cytoplasm. To test this hypothesis, we performed a differential exon usage analysis with DEXSeq (Section 2.3.3), using RNA-seq data from nuclear and cytoplasmic extracts, in HEK293 cells[70]. As shown in Figure 44, the differentially expressed exons show a bimodal distribution in their localization pattern. One distribution peaks at a moderate enrichment in the cytoplasm (log2FC<0), while many exons show a marked enrichment in the nucleus (log2FC>0). Such skewed pattern is expected, as the nuclear RNA-seq represents a mixture of signal coming from nascent pre-mRNA, splicing intermediates and mature RNAs, while cytoplasmic RNA-seq exhibits coverage only mature RNAs. Exons uniquely belonging to discarded transcript structures show a clear nuclear localization, while selected transcripts are more enriched in the cytoplasm. Again, a further separation can be made between selected transcripts, where non-translated selected transcript structures also show an enrichment in the nuclear fraction. Taken together, our analysis shows how it is possible to select for mature cytoplasmic transcript structures using Ribo-seq only.

ORF Finding:

As in the RiboTaper strategy (Section 3.1), the nucleotide sequence is used to determine the ORF position. The multitaper method is then employed to ensure the P-sites 3nt periodicity (thus active translation elongation) over the ORF.

ORF selection and Quantification:

We select ORFs using the same rules used for transcript filtering, this time using ORF bins and splice junctions derived from the ORF structures. ORF quantification is subsequently performed, using the length-normalized Ribo-seq coverage Cov on the ORF features.

$$Cov = \frac{\#reads}{length} \qquad (10)$$

P-sites positions are used for exonic regions, while spliced reads for exon-exon junctions. For splice junctions, the length is set to 60, according to the possible nucleotide space covered by a spliced read of ~30nt.

A feature F can be unique to one ORF or shared between multiple ORFs (Figure 44). For unique features Fu we can calculate the average coverage $AvCovUn$, using the coverage Cov_{Fu} on each of the $\#Fu$ unique features (Equation 11).

$$AvCovUn = \frac{\sum_{Fu}^{\#Fu} Cov_{Fu}}{\#F_u} \quad (11)$$

The same can be applied to all features $Fall$ mapping to the ORF (Equation 12)

$$AvCovAll = \frac{\sum_{Fall}^{\#F_{all}} Cov_{Fall}}{\#F_{all}} \qquad (12)$$

A scaling factor C_{ORF} is calculated, for each ORF, using the ratio between $AvCovUn$ and $AvCovAll$ (Equation 13). Such scaling factor represents the portion of signal that can be attributed to one ORF.

$$C_{ORF} = \frac{AvCovUn}{AvCovAll} \quad (13)$$

The maximum value for C_{ORF} is set to 1. When no unique region is present in one ORF (all regions are shared with other ORFs), the coverage Cov_{Fadj} on each feature $Fadj$ attributed to that ORF is calculated subtracting the expected signal coming from other ORF_{Fadj} mapping to the feature, using the scaling factors calculated in the Equations 10-13. In such cases, the calculation of the adjusted coverage for each feature $Fadj$ is as follows (Equations 14-16):

$$Cov_{Fadj} = \frac{\#reads_{Fadj}}{length_{Fadj}} - \frac{\#reads_{Fadj}}{length_{Fadj}} * \sum_{ORF_{Fadj}}^{\#ORFadj} C_{ORF_{Fadj}} \quad (14)$$

$$AvCovAdj = \frac{\sum_{F_{adj}}^{\#F_{adj}} Cov_{Fadj}}{\#F_{adj}} \qquad (15)$$

$$C_{ORF} = \frac{AvCovAdj}{AvCovAll} \quad (16)$$

If no unique region is present in any detected ORF in the gene (all regions are shared among ORFs), the scaling factor is calculated assuming uniform Ribo-seq coverage on each ORF. Coverage Cov_{Fsh} is simply divided by the number of ORF_{Fsh} mapping to the feature Fsh. (Equation 17-19).

$$Cov_{Fsh} = \frac{\#reads_{Fsh}}{length_{Fsh}} / \#ORF_{Fsh} \quad (17)$$

$$AvCovSh = \frac{\sum_{F_{sh}}^{\#F_{sh}} Cov_{Fsh}}{\#Fsh} \quad (18)$$

$$C_{ORF} = \frac{AvCovSh}{AvCovAll} \quad (19)$$

After the calculation of C_{ORF}, the adjusted number of P-sites for each ORF (P_{ORF}) is calculated using the raw number of P-sites mapping to the ORF multiplied by the scaling factor, to obtain ORF-specific quantification estimates (Equation 20).

$$P_{ORF} = Psites * C_{ORF} \quad (20)$$

For each ORF of length L_{ORF}, the scaled numbers of P-sites P_{ORF} is normalized over the entire set of detected ORFs $ORFN$, to obtain TPM-like values (see Equation 2), named P-sites per Nucleotide per Million (*P_sites_pNpM*), using this formula (Equation 21).:

$$\text{P\_sites\_pNpM}_{ORF} = \frac{P_{ORF}}{L_{ORF}} * \frac{10^6}{\sum_{ORF}^{\#ORF} \frac{P_{ORF}}{L_{ORF}}} \quad (21)$$

Moreover, we calculated the contribution of each ORF to the overall translation output of a single gene. Such metric, named *Iso_P_sites* (or percentage of gene translation), is calculated dividing P_{ORF} by the sum of P_{ORF} of all ORFs ($\#ORFg$) detected in a gene (Equation 22).

$$Iso\_P\_sites_{ORF} = \frac{P_{ORF}}{\sum_{ORF}^{\#ORFg} P_{ORF}} \quad (22)$$

Normalization by length is here not applied, as this metric wants to quantify the amount of translation per gene coming from each ORF. To filter out lowly translated ORFs, we retained ORFs until reaching 99% of the overall gene translation.

To check the validity of our quantification estimates, we ran our analysis on Ribo-seq from HEK293 (same dataset as in Section 3.2). We obtained ~27,000 ORFs, divided in ~14,000 genes. More than half (55%) of the detected genes harbored only one translated transcript, with the distribution of translated transcripts per gene exhibiting a power-law-like behavior (Figure 45). When looking at the quantification for the detected ORFs we observed a large amount (~30%) of lowly translated ORFs, harboring only between 0 and 5% of their host gene translation. This scenario outlined how more than 75% of genes have only 1 translated isoform representing >90% of gene translation.

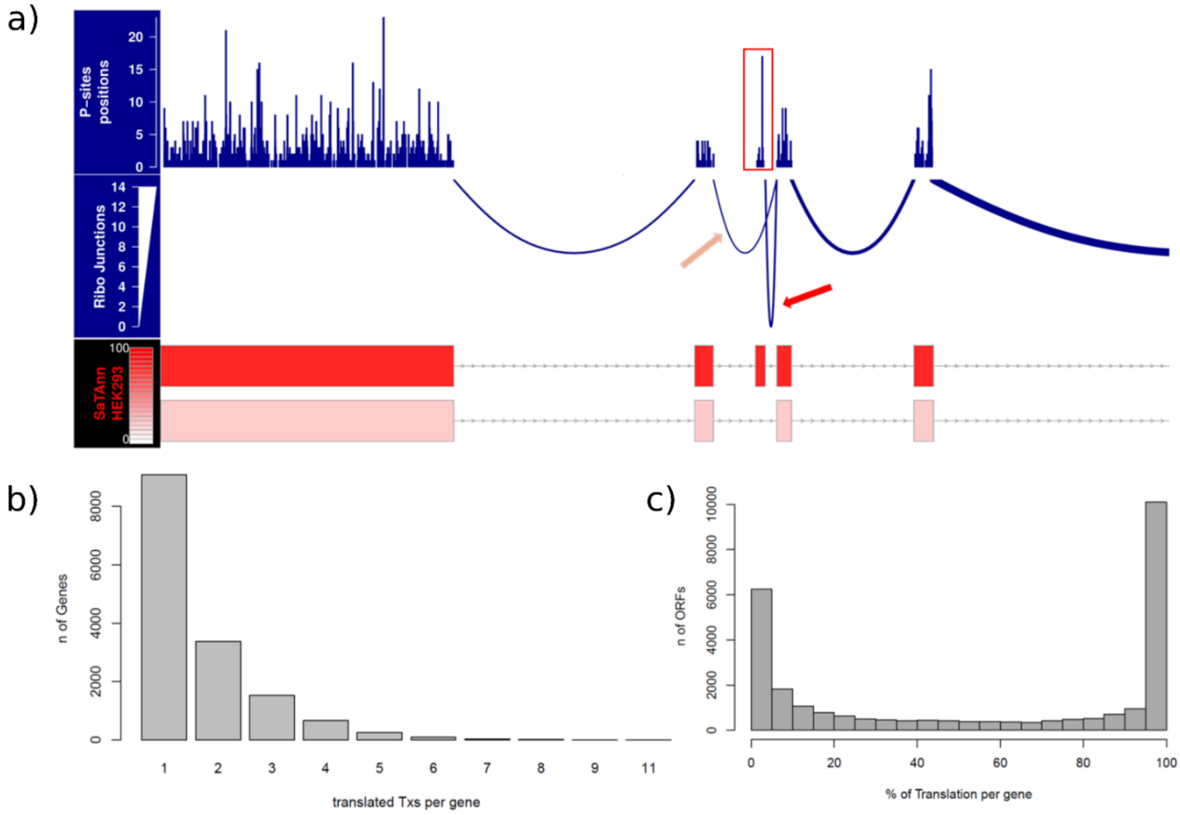


Figure 45: ORF-specific quantification of translation. In a) the quantification strategy is outlined: P-sites positions and splice junctions which are unique to an ORF structure (dark red or light red) are used to scale the total Ribo-seq coverage, allowing for the calculation of % of gene translation (used to color ORF structures of the lower track). In b) the distribution of translated transcripts per gene. In c) the percentage of gene translation across all the detected ORFs.

3.4.2 Validating translation quantification

To validate our quantification estimates, we compared them with a deep polysome profiling dataset from the same cell line[62] (Appendix B.15).

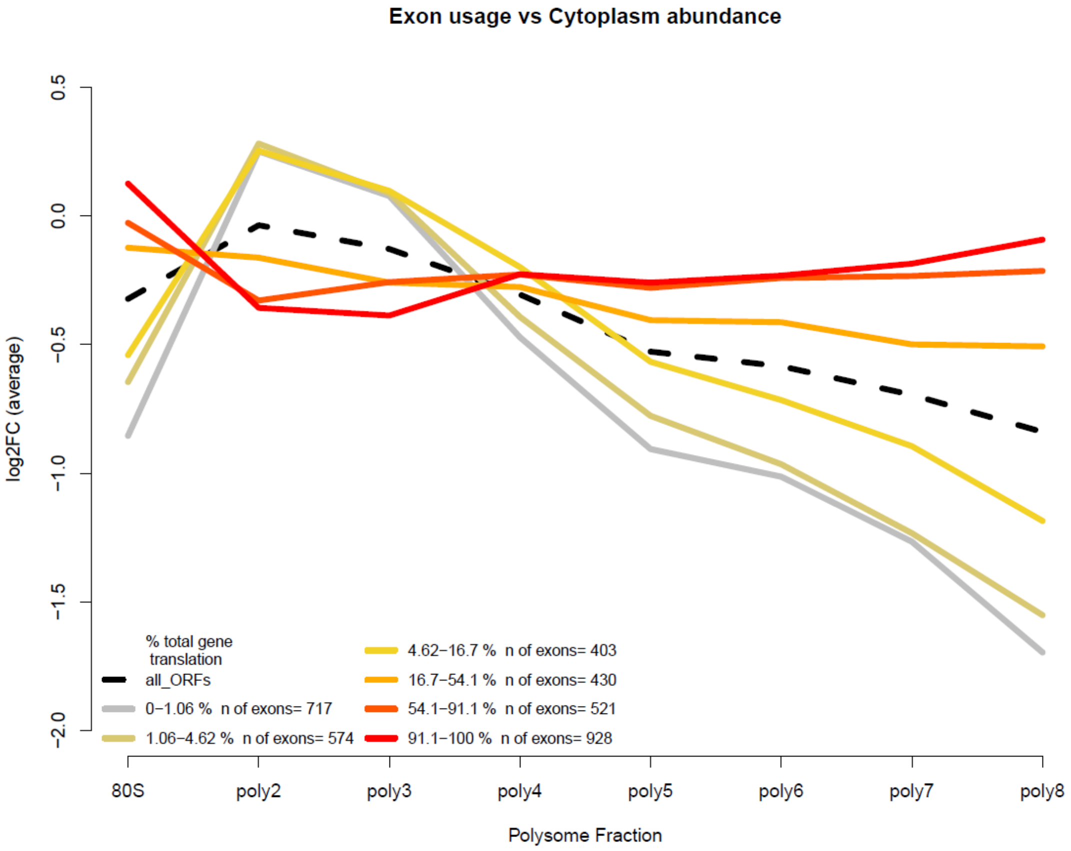


Figure 46: Polysome Profiling comparison. On the y-axis, the average log2 fold change over cytoplasmic abundance. On the x-axis, the different polysome fractions. Different lines indicate ORFs grouped by different translation levels (Iso_P_sites). Lowly translated ORFs do not migrate to heavy polysome fractions.

When looking at differential exon usage (again using DEXSeq) across the polysome fractions, we observed how our quantitative estimates correspond to distinct polysome profiles (Figure 46). Exons belonging to lowly translated ORFs migrate to low polysomes and are depleted in heavier polysomal fractions. Conversely, highly translated ORFs migrate also to the heavy polysomes.

Despite the fundamental differences between polysome profiling and Ribo-seq in representing the translated transcriptome, the two techniques agreed in detecting quantitative differences in the translation of multiple isoforms per gene.

Intuitively, our quantification of translation should reflect the rate at which proteins are synthesized. As in Section 3.2.3, we decided to match a deep proteomics dataset from the same cell line to our set of identified ORFs. We observed a good correlation between our quantitative estimates and protein abundance (Figure 47). The correlation values are slightly higher when compared to transcript-level translation quantification using RSEM supplied with Ribo-seq, again validating our ORF detection and quantification strategy.

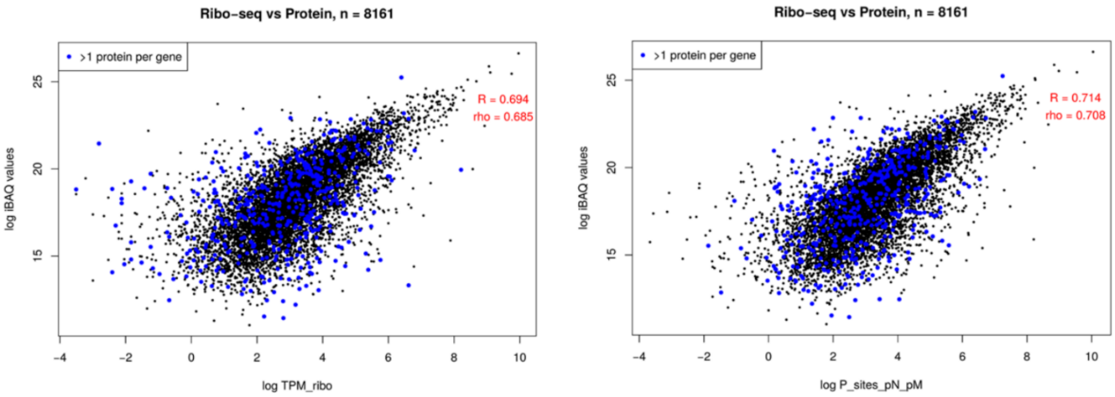


Figure 47: Proteome-wide correlations with translation estimates. Steady-state protein abundance, represented by the iBAQ values (obtained from label-free quantification) are correlated with either TPM values from RSEM (left), or our P_sites_pNpM values (right). Shown in red are the coefficients from Pearson (R) and Spearman (rho) correlations. Proteins from genes with more than one detected protein (in the MS/MS) are shown in blue. When multiple transcripts coded for the same protein, their TPM values were summed up.

Splice Features Annotation:

For each ORF, the genomic position of its features was annotated with respect to the longest annotated CDS, to identify alternative splicing events such as alternative acceptor and donor sites.

For genes with multiple translated transcripts, we built aggregate profiles over alternative splice sites, and we observed Ribo-seq coverage spanning alternative exons (Figure 48). As a control, we built the same profiles over region where the splice sites are shared between the isoforms. This analysis validated our ORF detection strategy, and confirmed the presence of a detectable mixture of signals coming from the translation of alternative RNA isoforms from a single gene.

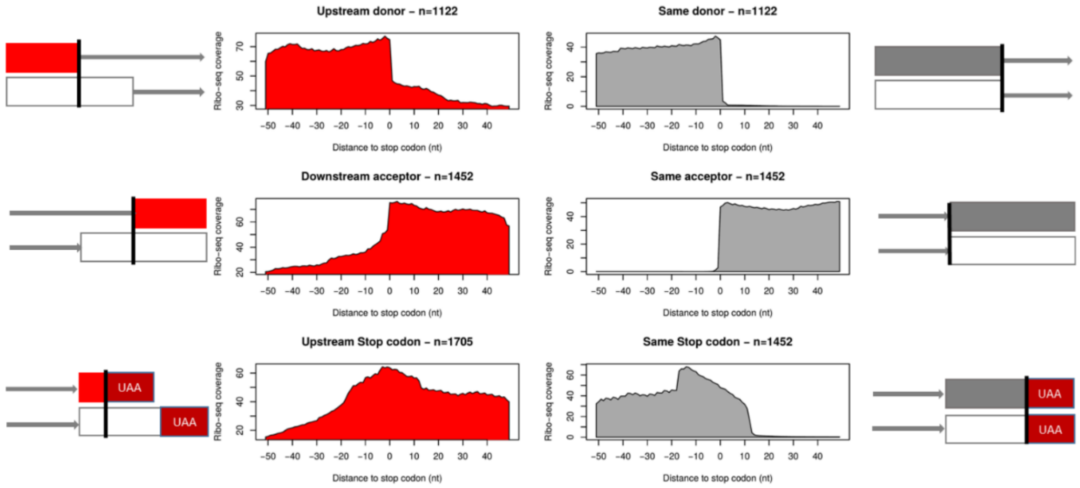


Figure 48: Alternative splicing events in Ribo-seq. Events (shown in red in the gene models) such as upstream donors (top), downstream acceptors (middle) or upstream stop codons (down) are annotated with respect to the longest isoform (shown in white in the gene model snippet). Aggregate profiles of Ribo-seq coverage are shown, delineating the presence of multiple translated RNAs. On the right, aggregate profiles are built over canonical events (in dark grey), as a control.

Taken together, multiple data sources validated the SaTAnn strategy to detect translation on multiple RNA isoforms, and allowed us to focus on interesting events at the interface between protein synthesis and RNA metabolism.

3.4.3 Translation on degraded RNA isoforms

The presence of numerous lowly translated transcripts (Figure 45) implies the presence of inefficient translation and/or very low steady-state abundance of the transcript. As translation and RNA degradation are intertwined in different RNA surveillance pathways (Section 2.1.7), we decided to investigate the different isoforms susceptibility to RNA decay mechanisms as a cause of low Ribo-seq signal. As a baseline, we investigated whether lowly translated transcripts were annotated as non-coding RNA isoforms. When focusing on genes with multiple translated transcript biotypes (Section 2.1.2), we observed how non-coding biotypes, such as nonsense-mediated decay and retained intron transcripts, were enriched in lowly translated transcripts (Supplementary Figure 4). This indicated that low levels of translation could be related to the transcript instability.

To test this hypothesis, we investigated the position of the detected ORFs on different transcripts. An important factor which can trigger NMD is the presence of a premature stop codon (PTC, Section 2.1.7). In one of the NMD modes of action, a stop codon can be recognized as premature when a downstream Exon Junction Complex (EJC) is not displaced during translation. To investigate the putative action of NMD on PTC-containing transcripts, we divided transcripts based on the presence of a splice site downstream of the detected stop codon. In theory, the presence of an EJC on the downstream splice site should trigger the recruitment of the NMD machinery, and thus cause RNA decay. We used data from a recent study in HEK293[57] (Appendix B.15), to map the cleavage events on NMD target transcripts.

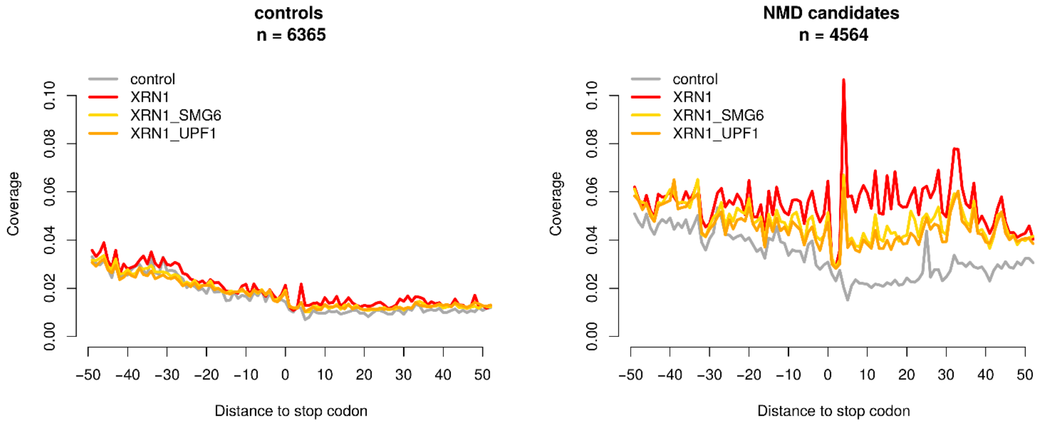


Figure 49: Degradation pattern over NMD target candidates. On the left, profiles of 5’ endonucleolytic cleavages are built around control ORFs. On the right, a higher coverage in the 5’ cleavages can be observed, accompanied by an enrichment when depleting the exonuclease XRN1 and a reduction when co-depleting members of the NMD pathway, like SMG6 or UPF1.

When mapping the cleavage sites over the stop codons of PTC and non-PTC containing transcripts from the same genes, we observed a clear difference (Figure 49): transcripts where the EJC is displaced (no PTC) showed background-like signal, while transcripts harboring a PTC, enriched in non-coding transcripts biotypes, showed a degradation profile around their stop codon, as expected. The degradation signal was less pronounced when also SMG6 or UPF1 were knocked-down, confirming the involvement of key elements of the NMD pathway in mediating translation-dependent isoform-specific RNA degradation.

One of the candidates for isoform-specific translation-dependent degradation is represented in Figure 50, where a lowly translated ORF in the *Diablo* gene shows a PTC and displays a higher XRN1-dependent degradation profile.

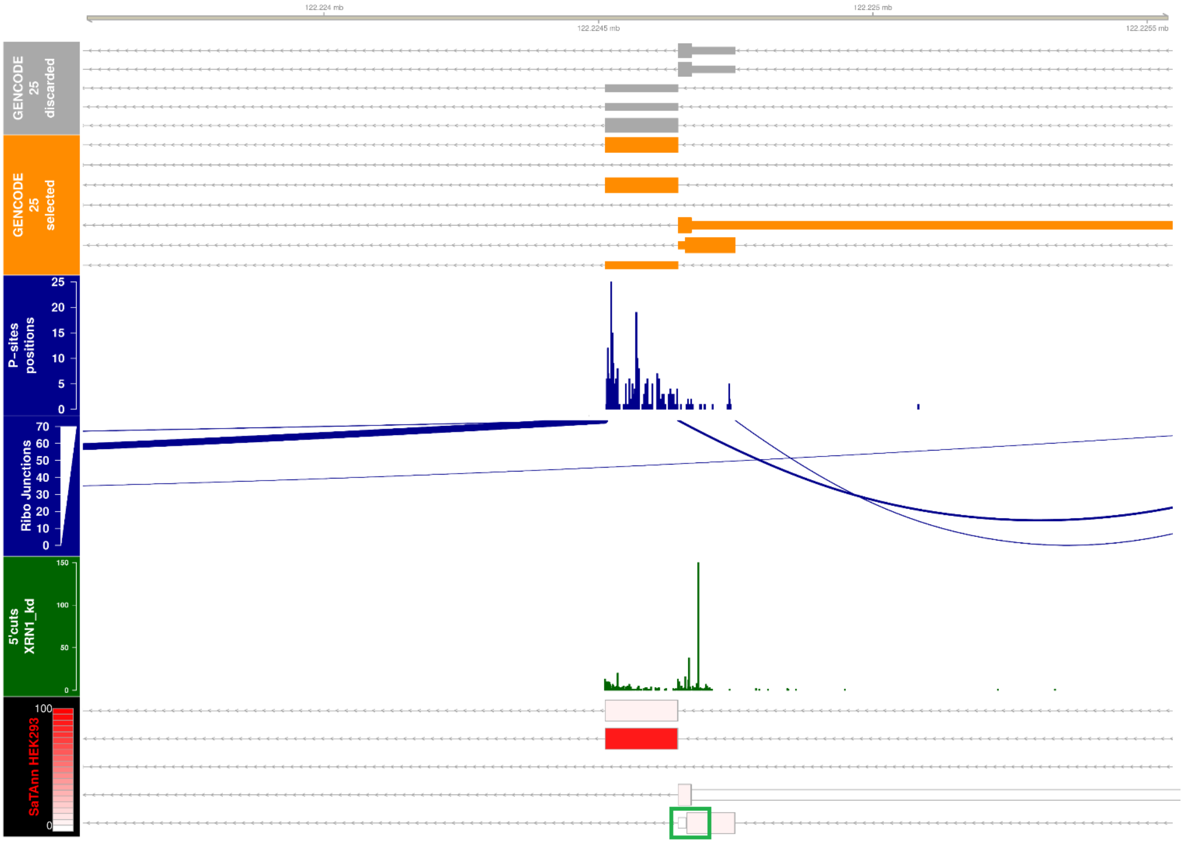


Figure 50: Example of isoform-specific NMD action. Depicted is a section of the *Diablo* gene. Shown are the discarded (gray) and selected (orange) transcript structures. P-sites positions and Ribo-seq junctions (in blue) show the presence of translation on alternative spliced isoforms. Shown is also the cut profile after XRN1 depletion (green), together with the SaTAnn-derived ORF structures below. Translation quantification (percentage of gene translation) shown in red scale. In the green box, an isoform with a recognized PTC is shown. Plot generated using Gviz.

To further explore the dependency of NMD action to the PTC location and the transcript type, we plotted the number of endonucleolytic cuts at the stop codon as a function of PTC distance to the exon-exon junction. As shown in Figure 51, for all the surveyed ORFs (including uORFs), we observed an increase in degradation with increasing distance from the exon-exon junction. As expected[57], ORFs in snoRNA-host genes showed the highest degradation profile, while other categories exhibited a lower amount of degradation, showing a wide dynamic range of isoform-specific RNA degradation.

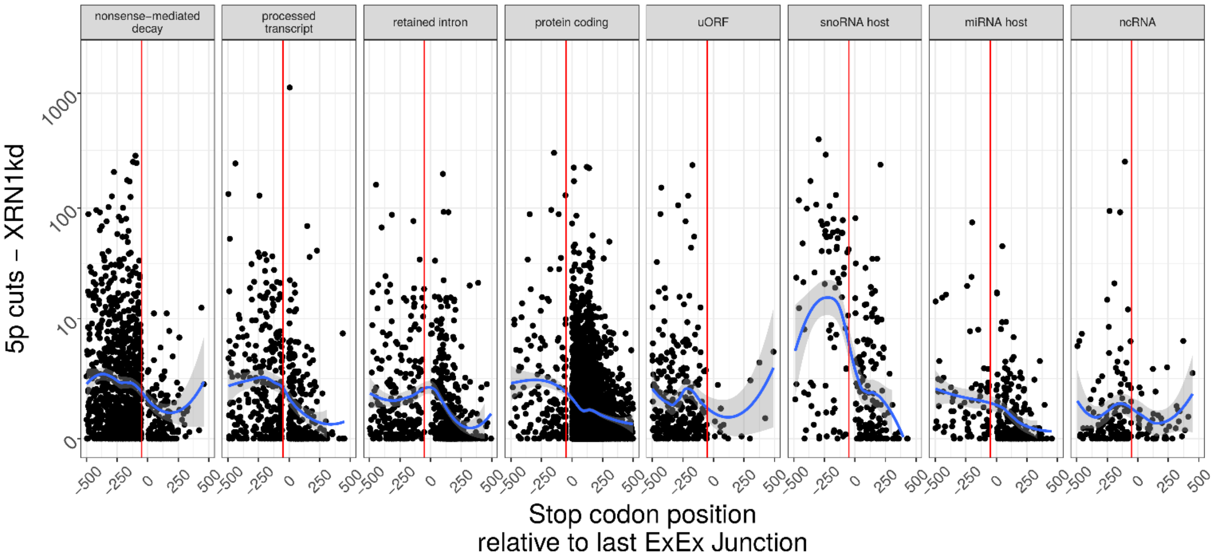


Figure 51: NMD action in different ORF categories. Each dot represents the number of endonucleolytic cuts (y-axis) in a window of 50nt around a stop codon; on the x-axis, the distance (in transcript coordinates) between the stop codon and the last exon-exon junction. Positive values (on the x-axis) indicate a “canonical” ORF, while negative values indicate PTC candidates. In red, a distance of 50nt is shown. ORFs are divided based on the host transcript biotype (first 4 panels), their position in the transcript (Panel 5), or based on their gene host biotype (Panels 6-8). The blue lines represent a local polynomial fitting (“loess”) of degree 2 with default parameters. Gray shadings represent 95% confidence intervals.

4 Discussion

Ribo-seq allows us to have an unprecedented look at a crucial step of gene regulation, where the genetic information is ultimately transferred to protein. At the same time, the translational status of a transcript influences its stability and its interaction with other cellular processes like intracellular trafficking, promoting the ribosome as a central hub for post-transcriptional gene regulation. Understanding the impressive wealth of information present in Ribo-seq data allows researchers to focus on such global aspects of gene expression regulation, uncovering the amount of control of translation[131]. At the same time, Ribo-seq provides a detailed description of the mechanisms of translation itself, thanks to its single-nucleotide resolution[190].

Such resolution provides a new angle for the computational analysis of Ribo-seq data, given the peculiar features of Ribo-seq data over translated regions (Section 2.3.6). The analysis of such features over the transcriptome allows us to identify the actively translated regions, resulting in an escalation of computational approaches which go beyond the collection of count-based statistics, switching to analysis strategies inspired by concepts coming from signal processing theory (Section 2.3.6). However, one of the problems in detecting high-confidence translation using Ribo-seq is the lack of a proper negative control. As the protocol does not entail the sequencing of an “input-like” pool of RNAs, we are left with only “positive” signal. As most of the Ribo-seq signal comes from expressed protein-coding genes, researchers have immediately shown great excitement about the specificity of the technique, with great efforts in the community to confirm translation of other non-coding regions of the transcriptome, such as uORF and small ORF translation, in all the systems where Ribo-seq has been performed.

However, the presence of ribosomal coverage on virtually any long transcript raised some criticism about the active translation of those regions, but also about the functional relevance of such translation events. Regarding the first aspect, many of the different proposed strategies leveraged on Ribo-seq coverage features observed in known coding regions, scoring ORFs transcriptome-wide and selecting high-scoring candidates[134], [138], [140]. It is unclear whether such approaches can determine the ensemble of translated regions with high specificity and sensitivity in different datasets. In fact, many of the proposed methods were applied to extremely deep Ribo-seq datasets, sometimes requiring multiple protocol variations[140], challenging the feasibility of such approaches at average data depths. A proper evaluation of the different available methods, using both simulations and real experimental data, will highlight the strength and pitfalls of the available ORF finding methods.

Thanks to the sub-codon resolution observed in Ribo-seq, we proposed a novel analysis strategy, which uses the multitaper method to identify translation on the basis of ribosomal elongation. The codon-by-codon movement of the elongating ribosomes is a universal feature of translation, which can be observed independent of the kinetics of initiation or termination, and allows for detection of translation over a wide range of expression levels (Figure 27). With its ability to combine statistical testing (Section 2.3.5 and Appendix A) with the spectral representation of a discrete signal, the multitaper proved to be a suitable tool to detect such pattern in noisy data such as NGS data. The high sensitivity of the multitaper method comes with its excellent specificity, as shown by simulations and by its performance on RNA-seq data (Section 3.1.1). The uniform distribution of multitaper-derived p-values on RNA-seq (Figure 22) showed the high specificity of the multitaper in identifying 3nt periodicity in Ribo-seq only, especially when compared with a simpler test for frame preference, which is the underlying engine of many early ORF-finding approaches[130], [142], [143]. Such extensive testing allowed us to confidently proceed to identify translation genome-wide.

We decided to focus only on AUG-starting ORF, as the validity of non-AUGs as efficient start codons is a matter of debate[36], despite the presence of few well-documented cases [191]. The choice of using only features of ribosome elongation limits the need of additional Ribo-seq variants. However, translation initiation mapping can prove useful when specifically interested in 5’UTR translation given the widespread usage of non-AUG start codons[37]. Distinguishing between ribosome initiation/elongation and other translation events leading to high ribosome occupancy (such as abortive translation initiation events) can be extremely challenging when looking at 5’UTRs, especially when analyzing conditions such as stress, tumor onset, or differentiation, which have been reported to exhibit high ribosomal coverage in 5’UTRs, most likely as a result of low levels of canonical, cap-dependent mechanisms of translation[35], [183], [192]. The integration of Ribo-seq with CLIP data for different translation factors[193] might represent a successful strategy to understand the different mechanisms that shape the dynamics of translation initiation.

Another limitation of our approach is its poor performance in detecting frameshifting events and multiple overlapping ORFs, which may require specialized analysis approaches[151]. Despite the limited presence of multi-frame translation in the human genome, such limitation can be crucial when studying translation of viral transcriptomes. Similarly, dedicated analyses might be used for other organisms where non-canonical mechanisms of translation elongation and termination represents vital means of correct protein expression[194].

In addition, our excitement is challenged by the high variability of the Ribo-seq technique, calling for a need of standardization of both the experimental method and the analysis strategy[89]. The RiboTaper strategy proved successful when applied to data with precise sub-codon resolution, while many published Ribo-seq datasets did not exhibit such high-resolution feature. Using *Arabidopsis thaliana* as a model, we investigated the performance of our computational approach when applied to Ribo-seq datasets from different labs, including a new dataset exhibiting an extremely accurate frame precision. As shown in Section 3.3.1, the datasets exhibiting the best sub-codon resolution also showed a better mapping over expected coding regions, and thus enabled accurate identification of coding genes also at moderate sequencing depths (Figure 39). Such differences in the protocol can thus have a big impact on our conclusion, depending on the level of resolution required by our research question. This aspect becomes crucial when analyzing events like translational pausing or determining codon-specific translation rates, which are known to have an impact in cellular homeostasis and proteome integrity[41]. The scientific community should dedicate more efforts in describing the technical limitation one might encounter when interpreting Ribo-seq results, and understand the possible biases caused by translation inhibitors[88], nuclease digestion[87], or other steps in the protocol[89].

Regarding the functional interpretation of novel ORFs, what is clear from dozens of different studies is that Ribo-seq challenged our “naïve” view of the genome where *1 gene -> 1 protein*, as hundreds of novel elements were shown to undergo active translation of one, or more, products. Our work (and dozens of other studies) highlighted the presence of novel translated elements, both in UTRs of coding genes and in non-coding genes. The presence of uORF translation was expected and known for many single cases, despite the difficulty in determining the actual usage of these element along the entire transcriptome[180]. As uORFs are very short, their active translation might be missed by many analysis methods, also considering that 0-aminoacid ORFs (a start-stop codon pair) do exists and can exhibit high ribosome occupancy[137]. The low coding potential of uORFs (Section 3.2.2) suggests that translation from 5’UTR has roles beyond protein synthesis. In line with this hypothesis, we detected a clear enrichment of nucleotide conservation at the boundaries of uORFs (Figure 31). This again confirms a role of such elements which is linked to their precise genomic location. Such observations are in line with our understanding of uORF as translational repressors of the main

CDS[47], despite the presence of multiple, and yet unresolved, molecular mechanisms proposed[46], [195]. As it allows us to zoom in the entire process of scanning and initiation, TCP-seq data might prove crucial in our understanding of the regulatory importance of 5'UTR features such as uORFs and other sequence and structural elements. However, the detection of few peptides uniquely mapping to uORFs (as in other studies[192]) raises the question whether some of these elements might be coding for important protein products, an observation which can be extended to other non-coding section of the transcriptome.

Translation of non-coding RNAs displays a similar pattern to 5'UTR translation[137]. For dozens of long non-coding transcripts, we could detect multiple translated short ORFs, as they displayed the 3nt periodicity typical of active translation. However, most ncORFs are lowly translated (Figure 28). The majority of novel candidate CDS did not show purifying selection at the codon level (Figure 32), and the same conclusion can be drawn from the analysis of genomic variation in the human population. As discussed before for uORFs, the overall low levels of translation and evolutionary conservation will surely mask the presence of few important regions. Whether multiple ORFs on a transcript might represent a way to regulate each other translation levels (as in uORFs) remains to be elucidated. More importantly, the presence of such elements must be connected to the precise transcript structures and thus other aspects of RNA metabolism (Section 3.4), as discussed below.

Translation in 3'UTR is a very rare event, which has been mostly linked with stop codon readthroughs[196] or alterations in ribosomal recycling[197]. We observed few dozens of genes with translated dORFs (Section 3.2.1), but their translation levels are very low (Figure 28). Despite the detection of a couple of peptides uniquely mapping to dORFs, the putative biological function of such elements remains unclear.

However, despite the analogy between our findings and the ones from similar studies[140], [141], [144], our conclusion might be affected by the choice of organism and protocol. Using an improved Ribo-seq protocol, the few novel CDS detected in *Arabidopsis thaliana* display homology even with distant plant species (Section 3.3.2), suggesting high coding potential (Figure 41). Such surprising results may also arise from a poorer annotation of coding loci in plant genomes. Another, perhaps more interesting, hypothesis is that the improved ability of the new protocol in isolating actively translating ribosomes enables the detection of high-fidelity coding loci. The improvement might depend on the choice of polysomal buffer conditions (Section 3.3.1), which will influence the RNAse footprinting step. Such observation is in line with the difference in the footprints obtained by different polysomal fraction, which can show marked differences: very recently, it has been shown how footprints obtained from the monosome fraction shows a clear enrichment of translated uORFs and NMD-sensitive transcripts[198], thus revealing a different subset of the translated transcriptome.

The understanding of such variability in the protocol is crucial, as the Ribo-seq method represents a powerful link between transcriptomics and proteomics techniques. As shown in Sections 3.2.3 and 3.3.2, such link allows us to improve gene annotation and protein detection. However, a comparison between the two scenarios depicted by Ribo-seq and shotgun proteomics must carefully consider fundamental differences between the two approaches.

The excellent sensitivity of Ribo-seq enables us to identify translation even for lowly expressed genes and small ORFs, which excited the research community in the recent years with many publications focused on the detection and characterization of small peptides. In our hands, we could show how such sensitivity enables us to further quantify the synthesis, for each gene, of multiple protein isoforms coming from distinct transcripts, showing again the tremendous potential of sequencing the ribosomal footprints to quantify the translational status of the entire transcriptome.

On the other hand, proteomics methods rely on the precise detection of millions of fragmented peptide ions (Section 2.2.4). However, shotgun proteomics methods might suffer technological limitations when trying to detect and quantify entire proteomes, providing evidence only for a subset of the synthesized proteins (~9,000 in this study). Such limitation may result from the inefficient detection of the entire spectrum of tryptic peptides. Different biochemical properties of trypsinized peptides can influence the likelihood of their detection, calling for the need of additional care when comparing the two techniques. Another, perhaps more important, aspect to consider is the presence of nearly ~200 post-translational modification (PTMs) which can occur in multiple protein residues, resulting in an exponential increase of the possible obtained *m/z* spectra in any experiment. A bit more than a dozen PTMs are common in most organisms, and only a limited number of them can be allowed when matching experimental and theoretical *m/z* spectra. It has been recently shown how, depending on the experimental conditions, different PTMs can cause 20-50% of false peptide identification, producing modified spectra which can perfectly match to a different peptide sequence[169]. In a similar fashion, the choice of sequence database has a heavy impact on the obtained results, a well-documented and discussed phenomenon in the proteomics community[170].

In our hands, we could show how the sensitivity of our approach resulted in excellent coverage of the detectable proteome in a deep MS/MS dataset in a human cell line, again confirming the validity of our ORF detection strategy. The detection of most novel peptides was confirmed when merging our set of translated ORFs with the entire catalog of annotated proteins (Figure 34). This strategy, consisting of multiple searches against our database and the Uniprot catalog, might also prove successful in the annotation of multiple tissues and different species, for which several mass spectrometry datasets are publicly available[176].Additional techniques, such as N-terminal COFRADIC, aim at isolating the N-termini of the synthesized proteins, aiding the quantification of different start codons usage[199], [200]. Unfortunately, the low throughput of shotgun-based proteomics techniques, limits our ability in deriving performance metrics for the identification of translation events together with their protein products.

However, a more cautious attitude must be taken before undermining the potential of proteomics methods, as equating ribosomal density to the production of a stable functional protein product might represent an over-simplistic approach, which ignores the presence of poorly understood phenomena, such as co-translational protein folding and degradation. A more careful analysis of the output of the two techniques together will shed light on the long standing question about the importance of splicing in determining proteome complexity[201]–[203].

We observed good proteome-wide correlation between protein steady-state abundance estimates and our translation quantification estimates (Section 3.4.2, Figure 47). However, many factors must be carefully taken into account when performing such a correlative analysis[204]. Ribo-seq reflects the protein synthesis rates of different ORFs, which would ideally correspond to the estimates given by proteomics techniques such as pSILAC[94]. The integration of Ribo-seq with proteomics techniques will help our understanding of proteome dynamics of synthesis and decay. Fine tuning of protein turnover represents a powerful tool for gene expression control[32], and additional mechanisms of co-translation protein degradation and folding can be tackled when integrating and improving our understanding of the translatome and the proteome.

As it represents one of the final steps of the RNA life cycle, Ribo-seq data characterizes the ultimate result of a mixture of biological processes, from RNA synthesis to selective degradation via several pathways (Section 2.1.7). Translation is a tightly regulated process, thanks to the presence of multiple mechanisms of translational control (Section 2.1.6), and to quantify the extent of such regulation, the overall RNA abundance must be considered. Quantitative modeling of translation regulation enables us to better understand gene expression, and Ribo-seq (when coupled to RNA-seq) can be used as a proxy to quantify translational control (Section 2.3.6). Expanding on this concept, the integration of Ribo-seq with data representing different layers of RNA regulation enables us to appreciate functional differences between different gene products[205]. The relevance of such strategy comes with a better understanding of the functional role of many lincRNA genes, which seem to not code for conserved proteins, and whose diverse functions are currently being investigated by the research community[205].

Such scenario becomes more complex when considering the plethora of alternatively spliced-transcripts, as different products from the same gene can undergo drastically different processing steps[61], [62]. In the SaTAnn approach (Section 3.4.1), we aimed at quantifying translation on the detectable RNA isoforms, uncovering the presence of multiple translated ORFs per gene in >40% of the detected genes (Figure 45). Given the quantitative nature of our Ribo-seq data (thanks to the use of UMIs), we decided to quantify translation at each candidate ORF. With SaTAnn, our goal is to determine the impact of transcript heterogeneity on the translated regions, which might reflect important differences at the level of the cellular proteome[206]. Our transcript selection strategy drastically reduced the number of possible structures to analyze, and thus enabled us to exploit signal on unique exons and splice junctions to estimate translation in an isoform-specific fashion. Such strategy differs from already well-established methods for transcript quantifications[98], [106], and one possible future development for our method can be the incorporation of an EM-like strategy to refine quantification estimates.

The comparison with polysome profiling (Figure 46) and proteomics data (Figure 47) showed the consistency of our quantification strategy, showing how different techniques can detect quantitative differences in the translation of different coding sequences. Nevertheless, while Ribo-seq offers a detailed picture of ribosomal movement across the translated region, polysome profiling adds information about entire transcript structures, representing a superior alternative when studying the impact of different 3’UTR isoforms on the translational output.

Despite such exiting scenario, we observed that thousands of translated RNA isoforms showed very low translation levels. A deeper look at the features of such low translated transcripts enabled us to uncover the presence of several NMD candidates (Figure 49), where the recognition of a PTC by the translating ribosomes is able to trigger RNA degradation[55]. Such information becomes crucial when investigating translation of long non-coding RNAs, which can exhibit high ribosomal coverage, but also high degradation levels, as shown for the snoRNA host genes (Figure 51). Isoform-level regulation of cytoplasmic RNA metabolism can thus be inferred using Ribo-seq data alone, enabling us to switch from detection to the functional investigation of the cytoplasmic transcriptome, one of the most fundamental topics in RNA biology.

Zooming in the sub-cellular translation status of different transcripts[20], [59] will further increase our understanding of the eukaryotic cytoplasm, where the coordinated action of heterogeneous ribosomes[52] in different organelles[207], [208] shapes protein function and cellular homeostasis.

5 References

[1] P. L. Luisi, F. Ferri, and P. Stano, “Approaches to semi-synthetic minimal cells: A review,” *Naturwissenschaften*, vol. 93, no. 1. pp. 1–13, 2006.

[2] S. D. Domagal-Goldman, K. E. Wright, K. Adamala, L. Arina de la Rubia, J. Bond, L. R. Dartnell, A. D. Goldman, K. Lynch, M.-E. Naud, I. G. Paulino-Lima, K. Singer, M. Walter-Antonio, X. C. Abrevaya, R. Anderson, G. Arney, D. Atri, A. Azúa-Bustos, J. S. Bowman, W. J. Brazelton, G. A. Brennecka, R. Carns, A. Chopra, J. Colangelo-Lillis, C. J. Crockett, J. DeMarines, E. A. Frank, C. Frantz, E. de la Fuente, D. Galante, J. Glass, D. Gleeson, C. R. Glein, C. Goldblatt, R. Horak, L. Horodyskyj, B. Kaçar, A. Kereszturi, E. Knowles, P. Mayeur, S. McGlynn, Y. Miguel, M. Montgomery, C. Neish, L. Noack, S. Rugheimer, E. E. Stüeken, P. Tamez-Hidalgo, S. I. Walker, and T. Wong, “The Astrobiology Primer v2.0,” *Astrobiology*, vol. 16, no. 8, pp. 561–653, Aug. 2016.

[3] O. T. Avery, C. M. Macleod, and M. McCarty, “STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III.,” *J. Exp. Med.*, vol. 79, no. 2, pp. 137–58, Feb. 1944.

[4] J. D. WATSON and F. H. C. CRICK, “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid,” *Nature*, vol. 171, no. 4356, pp. 737–738, Apr. 1953.

[5] G. Gamow, *Possible mathematical relation between deoxyribonucleic acid and proteins.* København : I kommission hos Munksgaard, 1954.

[6] F. H. C. CRICK, L. BARNETT, S. BRENNER, and R. J. WATTS-TOBIN, “General Nature of the Genetic Code for Proteins,” *Nature*, vol. 192, no. 4809, pp. 1227–1232, Dec. 1961.

[7] J. H. MATTHAEI, O. W. JONES, R. G. MARTIN, and M. W. NIRENBERG, “Characteristics and composition of RNA coding units.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 48, no. 4, pp. 666–77, Apr. 1962.

[8] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir, “Structure of a Ribonucleic Acid,” *Science (80-.).*, vol. 147, no. 3664, 1965.

[9] F. H. CRICK, “On protein synthesis.,” *Symp. Soc. Exp. Biol.*, vol. 12, pp. 138–63, 1958.

[10] F. CRICK, “Central Dogma of Molecular Biology,” *Nature*, vol. 227, no. 5258, pp. 561–

563, Aug. 1970.

[11] T. J. Treangen and S. L. Salzberg, “Repetitive DNA and next-generation sequencing: computational challenges and solutions,” *Nat. Rev. Genet.*, Nov. 2011.

[12] J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigó, and T. J. Hubbard, “GENCODE: the reference human genome annotation for The ENCODE Project.,” *Genome Res.*, vol. 22, no. 9, pp. 1760–74, Sep. 2012.

[13] J. Merkin, C. Russell, P. Chen, and C. B. Burge, “Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues,” *Science (80-.).*, vol. 338, no. 6114, 2012.

[14] X. Wang, J. Hou, C. Quedenau, and W. Chen, “Pervasive isoform-specific translational regulation via alternative transcription start sites in mammals,” *Mol Syst Biol*, vol. 12, no. 7, pp. 1–16, 2016.

[15] F. Carrillo Oesterreich, L. Herzel, K. Straube, K. Hujer, J. Howard, and K. M. Neugebauer, “Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II,” *Cell*, vol. 165, no. 2, pp. 372–381, Apr. 2016.

[16] L. Weill, E. Belloc, F.-A. Bava, and R. Méndez, “Translational control by changes in poly(A) tail length: recycling mRNAs,” *Nat. Struct. Mol. Biol.*, vol. 19, no. 6, pp. 577–585, Jun. 2012.

[17] B. Tian and J. L. Manley, “Alternative polyadenylation of mRNA precursors,” *Nat. Rev. Mol. Cell Biol.*, vol. 18, no. 1, pp. 18–30, Sep. 2016.

[18] M. Rabani, J. Z. Levin, L. Fan, X. Adiconis, R. Raychowdhury, M. Garber, A. Gnirke, C. Nusbaum, N. Hacohen, N. Friedman, I. Amit, and A. Regev, “Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells,” *Nat. Biotechnol.*, vol. 29, no. 5, pp. 436–442, May 2011.

[19] K. Bahar Halpern, I. Caspi, D. Lemze, M. Levy, S. Landen, E. Elinav, I. Ulitsky, and S. Itzkovitz, “Nuclear Retention of mRNA in Mammalian Tissues,” *Cell Rep.*, vol. 13, no. 12, pp. 2653–2662, Dec. 2015.

[20] D. W. Reid and C. V. Nicchitta, “Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling,” *J. Biol. Chem.*, vol. 287, no. 8, pp. 5518–5527, 2012.

[21] K. C. Martin and A. Ephrussi, “mRNA Localization: Gene Expression in the Spatial Dimension,” *Cell*, vol. 136, no. 4. pp. 719–730, 2009.

[22] P. Lasko, “mRNA Localization and Translational Control in Drosophila Oogenesis,” *Cold Spring Harb. Perspect. Biol.*, vol. 4, no. 10, pp. a012294–a012294, 2012.

[23] S. Gerstberger, M. Hafner, and T. Tuschl, “A census of human RNA-binding proteins,” *Nat. Rev. Genet.*, vol. 15, no. 12, pp. 829–845, 2014.

[24] R. Parker and U. Sheth, “P Bodies and the Control of mRNA Translation and Degradation,” *Molecular Cell*, vol. 25, no. 5. pp. 635–646, 2007.

[25] A. Castello, B. Fischer, M. W. Hentze, and T. Preiss, “RNA-binding proteins in Mendelian disease,” *Trends in Genetics*, vol. 29, no. 5. pp. 318–327, 2013.

[26] E. Szostak and F. Gebauer, “Translational control by 3’-UTR-binding proteins,” *Brief. Funct. Genomics*, vol. 12, no. 1, pp. 58–65, 2013.

[27] M. Kozak, “Initiation of translation in prokaryotes and eukaryotes,” *Gene*, vol. 234, no. 2, pp. 187–208, 1999.

[28] J. A. Steitz, “Nucleotide sequences of the ribosomal binding sites of bacteriophage R17 RNA.,” *Cold Spring Harb. Symp. Quant. Biol.*, vol. 34, pp. 621–630, 1969.

[29] K. Rooijers, F. Loayza-Puch, L. G. Nijtmans, and R. Agami, “Ribosome profiling reveals features of normal and disease-associated mitochondrial translation,” *Nat. Commun.*, vol. 4, no. 1, p. 2886, 2013.

[30] L. F. Lareau, D. H. Hite, G. J. Hogan, and P. O. Brown, “Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments,” *Elife*, vol. 2014, no. 3, 2014.

[31] N. T. Ingolia, S. Ghaemmaghami, J. R. S. Newman, and J. S. Weissman, “Genome-Wide Analysis in Vivo of Resolution Using Ribosome Profiling,” *Science (80-.).*, vol. 324, pp. 218–23, 2009.

[32] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach, “Global quantification of mammalian gene expression control,” *Nature*, vol. 473, no. 7347, pp. 337–342, 2011.

[33] M. Chekulaeva, H. Mathys, J. T. Zipprich, J. Attig, M. Colic, R. Parker, and W. Filipowicz, “miRNA repression involves GW182-mediated recruitment of CCR4–NOT through conserved W-containing motifs,” *Nat. Struct. Mol. Biol.*, vol. 18, no. 11, pp. 1218–1226, 2011.

[34] R. J. Jackson, C. U. T. Hellen, and T. V Pestova, “The mechanism of eukaryotic translation initiation and principles of its regulation.,” *Nat. Rev. Mol. Cell Biol.*, vol. 11,

no. 2, pp. 113–127, 2010.

[35] S. R. Starck, J. C. Tsai, K. Chen, M. Shodiya, L. Wang, K. Yahiro, M. Martins-Green, N. Shastri, and P. Walter, “Translation from the 5’ untranslated region shapes the integrated stress response,” *Science (80-.).*, vol. 351, no. 6272, p. aad3867-aad3867, 2016.

[36] A. M. Michel, D. E. Andreev, and P. V Baranov, “Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning.,” *BMC Bioinformatics*, vol. 15, no. 1, p. 380, 2014.

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

[38] I. Tzani, I. P. Ivanov, D. E. Andreev, R. I. Dmitriev, K. A. Dean, P. V Baranov, J. F. Atkins, and G. Loughran, “Systematic analysis of the PTEN 5’ leader identifies a major AUU initiated proteoform.,” *Open Biol.*, vol. 6, no. 5, pp. 731–745, 2016.

[39] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, “Predicting subcellular localization of proteins based on their N-terminal amino acid sequence.,” *J. Mol. Biol.*, vol. 300, no. 4, pp. 1005–1016, 2000.

[40] D. Tarrant and T. Von Der Haar, “Synonymous codons, ribosome speed, and eukaryotic gene expression regulation,” *Cellular and Molecular Life Sciences*, vol. 71, no. 21. pp. 4195–4206, 2014.

[41] D. D. Nedialkova and S. A. Leidel, “Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity,” *Cell*, vol. 161, no. 7, pp. 1606–1618, 2015.

[42] A. T. Belew, A. Meskauskas, and S. Musalgaonkar, “Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway,” *Nature*, vol. 512, no. 7514, pp. 265–269, 2014.

[43] I. Jungreis, M. F. Lin, R. Spokony, C. S. Chan, N. Negre, A. Victorsen, K. P. White, and M. Kellis, “Evidence of abundant stop codon readthrough in Drosophila and other metazoa,” *Genome Res.*, vol. 21, no. 12, pp. 2096–2113, 2011.

[44] A. Böck, K. Forchhammer, J. Heider, and C. Baron, “Selenoprotein synthesis: an expansion of the genetic code,” *Trends in Biochemical Sciences*, vol. 16, no. C. pp. 463–467, 1991.

[45] T. Schneider-Poetsch, J. Ju, D. E. Eyler, Y. Dang, S. Bhat, W. C. Merrick, R. Green, B. Shen, and J. O. Liu, “Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin.,” *Nat. Chem. Biol.*, vol. 6, no. 3, pp. 209–217, 2010.

[46] S. E. Calvo, D. J. Pagliarini, and V. K. Mootha, “Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 18, pp. 7507–12, 2009.

[47] T. G. Johnstone, A. A. Bazzini, and A. J. Giraldez, “Upstream ORFs are prevalent translational repressors in vertebrates,” *EMBO J.*, vol. 35, no. 7, pp. 1–18, 2016.

[48] J. Rohwedel, S. Kügler, T. Engebrecht, W. Purschke, P. K. Müller, and C. Kruse, “Evidence for posttranscriptional regulation of the multi K homology domain protein vigilin by a small peptide encoded in the 5’ leader sequence,” *Cell. Mol. Life Sci.*, vol. 60, no. 8, pp. 1705–1715, 2003.

[49] N. Thakor and M. Holcik, “IRES-mediated translation of cellular messenger RNA operates in eIF2α-independent manner during stress,” *Nucleic Acids Res.*, vol. 40, no. 2, pp. 541–552, 2012.

[50] J. S. Kieft, “Viral IRES RNA structures and ribosome interactions,” *Trends in Biochemical Sciences*, vol. 33, no. 6. pp. 274–283, 2008.

[51] N. Slavov, S. Semrau, E. Airoldi, B. Budnik, and A. van Oudenaarden, “Differential Stoichiometry among Core Ribosomal Proteins,” *Cell Rep.*, vol. 13, no. 5, pp. 865–873, 2015.

[52] S. Xue and M. Barna, “Specialized ribosomes: a new frontier in gene regulation and organismal biology,” *Nat. Rev. Mol. Cell Biol.*, vol. 13, no. 6, pp. 355–369, 2012.

[53] A. A. Bazzini, M. T. Lee, and A. J. Giraldez, “Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish.,” *Science*, vol. 336, no. 6078, pp. 233–7, 2012.

[54] O. Larsson and R. Nadon, “Re-analysis of genome wide data on mammalian microRNA-mediated suppression of gene expression,” *Translation*, vol. 1, no. 1, p. e24557, 2013.

[55] S. Lykke-Andersen and T. H. Jensen, “Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes.,” *Nat. Rev. Mol. Cell Biol.*, vol. 16, no. 11, pp. 665–677, 2015.

[56] Y.-F. F. Chang, J. S. Imam, and M. F. Wilkinson, “The nonsense-mediated decay RNA surveillance pathway,” *Annu. Rev. Biochem.*, vol. 76, no. February, pp. 51–74, 2007.

[57] S. Lykke-Andersen, Y. Chen, B. R. Ardal, B. Lilje, J. Waage, A. Sandelin, and T. H. Jensen, “Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes.,” *Genes Dev.*, vol. 28, no. 22, pp. 2498–517, Nov. 2014.

[58] E. D. Karousis, S. Nasif, and O. Mühlemann, “Nonsense-mediated mRNA decay: novel mechanistic insights and biological impact,” *Wiley Interdiscip. Rev. RNA*, vol. 7, no. 5,

pp. 661–682, 2016.

[59] C. H. Jan, C. C. Williams, and J. S. Weissman, “Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling,” *Science*, vol. 346, no. 6210, pp. 1257521–1257521, 2014.

[60] R. J. Weatheritt, T. Sterne-Weiler, and B. J. Blencowe, “The ribosome-engaged landscape of alternative splicing.,” *Nat. Struct. Mol. Biol.*, no. November, pp. 1–9, 2016.

[61] T. Sterne-Weiler, R. T. Martinez-Nunez, J. M. Howard, I. Cvitovik, S. Katzman, M. A. Tariq, N. Pourmand, and J. R. Sanford, “Frac-seq reveals isoform-specific recruitment to polyribosomes,” *Genome Res.*, vol. 23, no. 10, pp. 1615–1623, 2013.

[62] S. N. Floor and J. A. Doudna, “Tunable protein synthesis by transcript isoforms in human cells,” *Elife*, vol. 5, no. JANUARY2016, 2016.

[63] J. J. L. Wong, W. Ritchie, O. A. Ebner, M. Selbach, J. W. H. Wong, Y. Huang, D. Gao, N. Pinello, M. Gonzalez, K. Baidya, A. Thoeng, T. L. Khoo, C. G. Bailey, J. Holst, and J. E. J. Rasko, “Orchestrated intron retention regulates normal granulocyte differentiation,” *Cell*, vol. 154, no. 3, pp. 583–595, 2013.

[64] K. V Voelkerding, S. a Dames, and J. D. Durtschi, “Next-generation sequencing: from basic research to diagnostics.,” *Clin. Chem.*, vol. 55, no. 4, pp. 641–58, 2009.

[65] K. D. Hansen, S. E. Brenner, and S. Dudoit, “Biases in Illumina transcriptome sequencing caused by random hexamer priming,” *Nucleic Acids Res.*, vol. 38, no. 12, pp. e131–e131, Jul. 2010.

[66] X. Victoria, N. Blades, J. Ding, R. Sultana, and G. Parmigiani, “Estimation of sequencing error rates in short reads,” *BMC Bioinformatics*, vol. 13, no. 1, p. 185, 2012.

[67] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nat. Methods*, vol. 5, no. 7, pp. 621–628, Jul. 2008.

[68] R. Sooknanan, J. Pease, and K. Doyle, “Novel methods for rRNA removal and directional, ligation-free RNA-seq library preparation,” *Nat. Methods*, vol. 7, no. 10, Oct. 2010.

[69] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. Hayes, and C. M. Perou, “Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling,” *BMC Genomics*, vol. 15, no. 1, p. 419, 2014.

[70] M. Sultan, V. Amstislavskiy, T. Risch, M. Schuette, S. Dökel, M. Ralser, D. Balzereit, H. Lehrach, and M.-L. Yaspo, “Influence of RNA extraction methods and library selection schemes on RNA-seq data,” *BMC Genomics*, vol. 15, no. 1, p. 675, 2014.

[71] J. M. Taliaferro, M. Vidaki, R. Oliveira, S. Olson, L. Zhan, T. Saxena, E. T. Wang, B. R. Graveley, F. B. Gertler, M. S. Swanson, and C. B. Burge, “Distal Alternative Last Exons Localize mRNAs to Neural Projections,” *Mol. Cell*, vol. 61, no. 6, pp. 821–833, 2016.

[72] M. De Hoon and Y. Hayashizaki, “Deep cap analysis gene expression (CAGE): Genome-wide identification of promoters, quantification of their expression, and network inference,” *BioTechniques*, vol. 44, no. 5. pp. 627–632, 2008.

[73] G. Martin, A. R. Gruber, W. Keller, and M. Zavolan, “Genome-wide Analysis of Pre-mRNA 3’ End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3’ UTR Length,” *Cell Rep.*, vol. 1, no. 6, pp. 753–763, 2012.

[74] H. Chang, J. Lim, M. Ha, and V. N. Kim, “TAIL-seq: Genome-wide determination of poly(A) tail length and 3’ end modifications,” *Mol. Cell*, vol. 53, no. 6, pp. 1044–1052, 2014.

[75] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A. C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl, “Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP,” *Cell*, vol. 141, no. 1, pp. 129–141, 2010.

[76] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo, “Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP).,” *Nat. Methods*, vol. 13, no. November 2015, pp. 1–9, 2016.

[77] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, “Genome-wide measurement of RNA secondary structure in yeast.,” *Nature*, vol. 467, no. 7311, pp. 103–7, 2010.

[78] L. Windhager, T. Bonfert, K. Burger, Z. Ruzsics, S. Krebs, S. Kaufmann, G. Malterer, A. L’Hernault, M. Schilhabel, S. Schreiber, P. Rosenstiel, R. Zimmer, D. Eick, C. C. Friedel, and L. Dölken, “Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution,” *Genome Res.*, vol. 22, no. 10, pp. 2031–2042, 2012.

[79] P. Preker, J. Nielsen, S. Kammler, S. Lykke-Andersen, M. S. Christensen, C. K. Mapendano, M. H. Schierup, and T. H. Jensen, “RNA Exosome Depletion Reveals Transcription Upstream of Active Human Promoters,” *Science (80-.).*, vol. 322, no. 5909,

pp. 1851–1854, 2008.

[80] T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale, “Counting absolute numbers of molecules using unique molecular identifiers,” *Nat. Methods*, vol. 9, no. 1, pp. 72–74, 2011.

[81] N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, and J. S. Weissman, “The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments.,” *Nat. Protoc.*, vol. 7, no. 8, pp. 1534–50, 2012.

[82] N. T. Ingolia, “Ribosome Footprint Profiling of Translation throughout the Genome,” *Cell*, vol. 165, no. 1, pp. 22–33, 2016.

[83] S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, and S.-B. Qian, “Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 37, pp. E2424–E2432, Sep. 2012.

[84] X. Gao, J. Wan, B. Liu, M. Ma, B. Shen, and S. B. Qian, “Quantitative profiling of initiating ribosomes in vivo,” *Nat Methods*, vol. 12, no. 2, pp. 147–153, 2015.

[85] S. K. Archer, N. E. Shirokikh, T. H. Beilharz, and T. Preiss, “Dynamics of ribosome scanning and recycling revealed by translation complex profiling.,” *Nature*, vol. 535, no. 7613, pp. 570–4, 2016.

[86] N. Hornstein, D. Torres, S. Das Sharma, G. Tang, P. Canoll, and P. A. Sims, “Ligation-free ribosome profiling of cell type-specific translation in the brain,” *Genome Biol.*, pp. 1–15, 2016.

[87] M. V. Gerashchenko and V. N. Gladyshev, “Ribonuclease selection for ribosome profiling,” *Nucleic Acids Res.*, p. gkw822, 2016.

[88] M. V Gerashchenko and V. N. Gladyshev, “Translation inhibitors cause abnormalities in ribosome profiling experiments.,” *Nucleic Acids Res.*, vol. 42, no. 17, pp. 1–7, 2014.

[89] P. B. F. O’Connor, D. E. Andreev, and P. V Baranov, “Comparative survey of the relative impact of mRNA features on local ribosome profiling read density.,” *Nat. Commun.*, vol. 7, p. 12915, 2016.

[90] M. Larance and A. I. Lamond, “Multidimensional proteomics for cell biology.,” *Nat. Rev. Mol. Cell Biol.*, vol. 16, no. 5, pp. 269–280, 2015.

[91] P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin, “Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents,” *Mol. Cell. Proteomics*, vol. 3, no. 12, pp. 1154–1169, 2004.

[92] S. Wiese, K. A. Reidegeld, H. E. Meyer, and B. Warscheid, “Protein labeling by iTRAQ:

A new tool for quantitative mass spectrometry in proteome research," *Proteomics*, vol. 7, no. 3, pp. 340–350, 2007.

[93] S.-E. Ong, B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey, and M. Mann, "Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.," *Mol. Cell. Proteomics*, vol. 1, no. 5, pp. 376–86, May 2002.

[94] B. Schwanhäusser, M. Gossen, G. Dittmar, and M. Selbach, "Global analysis of cellular protein translation by pulsed SILAC," *Proteomics*, vol. 9, no. 1, pp. 205–209, Jan. 2009.

[95] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M.-M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman, "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes.," *Genome Res.*, vol. 19, no. 7, pp. 1316–23, Jul. 2009.

[96] J. M. Rodriguez, P. Maietta, I. Ezkurdia, A. Pietrelli, J.-J. Wesselink, G. Lopez, A. Valencia, and M. L. Tress, "APPRIS: annotation of principal and alternative splice isoforms.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D110-7, Jan. 2013.

[97] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733–D745, Jan. 2016.

[98] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.," *BMC Bioinformatics*, vol. 12, no. 1, p. 323, 2011.

[99] M. G. . Grabherr, N. Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson,

Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W., and and A. R. Friedman, “Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data,” *Nat. Biotechnol.*, vol. 29, no. 7, pp. 644–652, 2013.

[100] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, pp. 10–12, 2011.

[101] H. Li and R. Durbin, “Fast and accurate short read alignment with Burrows-Wheeler transform,” *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[102] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, “STAR: ultrafast universal RNA-seq aligner,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013.

[103] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome,” *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.

[104] C. Trapnell, L. Pachter, and S. L. Salzberg, “TopHat: discovering splice junctions with RNA-Seq,” *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, May 2009.

[105] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *Nat. Biotechnol.*, vol. 34, no. 5, pp. 525–527, Apr. 2016.

[106] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” *Nat. Methods*, vol. 14, no. 4, pp. 417–419, Mar. 2017.

[107] M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey, “Software for Computing and Annotating Genomic Ranges,” *PLoS Comput. Biol.*, vol. 9, no. 8, p. e1003118, Aug. 2013.

[108] A. R. Quinlan and I. M. Hall, “BEDTools: a flexible suite of utilities for comparing genomic features,” *Bioinformatics*, vol. 26, no. 6, pp. 841–842, Mar. 2010.

[109] L. Jiang, F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, T. R. Gingeras, and B. Oliver, “Synthetic spike-in standards for RNA-seq experiments,” *Genome Res.*, vol. 21, no. 9, pp. 1543–1551, Sep. 2011.

[110] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, “RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays,” *Genome Res.*, vol. 18, no. 9, pp. 1509–1517, Jul. 2008.

[111] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for

differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Jan. 2010.

[112] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol.*, vol. 11, no. 10, p. R106, 2010.

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

[114] S. Anders, A. Reyes, and W. Huber, "Detecting differential usage of exons from RNA-seq data," *Genome Res.*, vol. 22, no. 10, pp. 2008–2017, Oct. 2012.

[115] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter, "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nat. Biotechnol.*, vol. 28, no. 5, pp. 511–515, 2010.

[116] M. Teng, M. I. Love, C. A. Davis, S. Djebali, A. Dobin, B. R. Graveley, S. Li, C. E. Mason, S. Olson, D. Pervouchine, C. A. Sloan, X. Wei, L. Zhan, and R. A. Irizarry, "A benchmark for RNA-seq quantification pipelines," *Genome Biol.*, vol. 17, no. 1, p. 74, Dec. 2016.

[117] A. Kanitz, F. Gypas, A. J. Gruber, A. R. Gruber, G. Martin, and M. Zavolan, "Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data," *Genome Biol.*, vol. 16, no. 1, p. 150, Dec. 2015.

[118] H. Feng, X. Zhang, C. Zhang, U. H. Koszinowski, and R. Zimmer, "mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA-sequencing data," *Nat. Commun.*, vol. 6, p. 7816, Aug. 2015.

[119] L. Wang, J. Nie, H. Sicotte, Y. Li, J. E. Eckel-Passow, S. Dasari, P. T. Vedell, P. Barman, L. Wang, R. Weinshiboum, J. Jen, H. Huang, M. Kohli, and J.-P. A. Kocher, "Measure transcript integrity using RNA-seq data," *BMC Bioinformatics*, vol. 17, no. 1, p. 58, Dec. 2016.

[120] N. F. Lahens, I. Kavakli, R. Zhang, K. Hayer, M. B. Black, H. Dueck, A. Pizarro, J. Kim, R. Irizarry, R. S. Thomas, G. R. Grant, and J. B. Hogenesch, "IVT-seq reveals extreme bias in RNA sequencing," *Genome Biol.*, vol. 15, no. 6, p. R86, 2014.

[121] W. H. Majoros, N. Lebeck, U. Ohler, and S. Li, "Improved transcript isoform discovery using ORF graphs," *Bioinformatics*, vol. 30, no. 14, pp. 1958–1964, Jul. 2014.

[122] D. L. Corcoran, S. Georgiev, N. Mukherjee, E. Gottwein, R. L. Skalsky, J. D. Keene, and U. Ohler, "PARalyzer: definition of RNA binding sites from PAR-CLIP short-read

sequence data.," *Genome Biol.*, vol. 12, no. 8, p. R79, Aug. 2011.

[123] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[124] D. J. Thomson and D. J., "Spectrum Estimation and Harmonic Analysis," *Proc. IEEE, Vol. 70, p. 1055-1096*, vol. 70, pp. 1055–1096, 1982.

[125] D. Slepian, "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty-V: The Discrete Case," *Bell Syst. Tech. J.*, vol. 57, no. 5, pp. 1371–1430, May 1978.

[126] G. A. Prieto, R. L. Parker, D. J. Thomson, F. L. Vernon, and R. L. Graham, "Reducing the bias of multitaper spectrum estimates," *Geophys. J. Int.*, vol. 171, no. 3, pp. 1269–1281, Oct. 2007.

[127] D. J. Thomson, C. G. Maclennan, and L. J. Lanzerotti, "Propagation of solar oscillations through the interplanetary medium," *Nature*, vol. 376, no. 6536, pp. 139–144, Jul. 1995.

[128] B. Babadi and E. N. Brown, "A Review of Multitaper Spectral Analysis," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1555–1564, May 2014.

[129] K. J. Rahim, W. S. Burr, and D. J. Thomson, "Appendix A: Multitaper R Package in 'Applications of Multitaper Spectral Analysis to Nonstationary Data,' PhD diss., Queen's University,." pp. 149–183, 2014.

[130] A. a. Bazzini, T. G. Johnstone, R. Christiano, S. D. MacKowiak, B. Obermayer, E. S. Fleming, C. E. Vejnar, M. T. Lee, N. Rajewsky, T. C. Walther, and A. J. Giraldez, "Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation," *EMBO J.*, vol. 33, no. 9, pp. 981–993, 2014.

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

[132] O. Larsson, N. Sonenberg, and R. Nadon, "anota: analysis of differential translation in genome-wide studies," *Bioinformatics*, vol. 27, no. 10, pp. 1440–1441, May 2011.

[133] O. Larsson, N. Sonenberg, and R. Nadon, "Identification of differential translation in genome wide studies.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 50, pp. 21487–92, Dec. 2010.

[134] J. Ruiz-Orera, X. Messeguer, J. A. Subirana, and M. M. Alba, "Long non-coding RNAs as a source of new peptides," *Elife*, vol. 3, p. e03523, Sep. 2014.

[135] J. L. Aspden, Y. C. Eyre-Walker, R. J. Phillips, U. Amin, M. A. S. Mumtaz, M. Brocard, and J.-P. Couso, "Extensive translation of small Open Reading Frames revealed by Poly-

Ribo-Seq,” *Elife*, vol. 3, Aug. 2014.

[136] M. Guttman, P. Russell, N. T. Ingolia, J. S. Weissman, E. S. Lander, S. X. Huang, M. Ma, B. Shen, S. B. Qian, H. Hengel, et al., R. G. E. R. G. and G. S. G. (Genome N. P. C. Group), et al., W. U. G. S. Center, B. Institute, C. H. O. R. Institute, and et al., “Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins,” *Cell*, vol. 154, no. 1, pp. 240–251, Jul. 2013.

[137] G.-L. Chew, A. Pauli, J. L. Rinn, A. Regev, A. F. Schier, and E. Valen, “Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs,” *Development*, vol. 140, no. 13, pp. 2828–2834, Jul. 2013.

[138] N. T. Ingolia, G. A. Brar, N. Stern-Ginossar, M. S. Harris, G. J. S. Talhouarne, S. E. Jackson, M. R. Wills, J. S. Weissman, N. A. Hosken, F. Kern, et al., R. G. E. R. G. and G. S. G. (Genome N. P. C. Group), and et al., “Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes,” *Cell Rep.*, vol. 8, no. 5, pp. 1365–1379, Sep. 2014.

[139] J. N. Hutchinson, A. W. Ensminger, C. M. Clemson, C. R. Lynch, J. B. Lawrence, and A. Chess, “A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains.,” *BMC Genomics*, vol. 8, p. 39, 2007.

[140] A. P. Fields, E. H. Rodriguez, M. Jovanovic, N. Stern-Ginossar, B. J. Haas, P. Mertins, R. Raychowdhury, N. Hacohen, S. A. Carr, N. T. Ingolia, A. Regev, and J. S. Weissman, “A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation,” *Mol. Cell*, vol. 60, no. 5, pp. 816–827, Dec. 2015.

[141] A. Raj, S. H. Wang, H. Shim, A. Harpak, Y. I. Li, B. Engelmann, M. Stephens, Y. Gilad, and J. K. Pritchard, “Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling,” *Elife*, vol. 5, pp. 1586–1591, May 2016.

[142] J. E. Smith, J. R. Alvarez-Dominguez, N. Kline, N. J. Huynh, S. Geisler, W. Hu, J. Coller, K. E. Baker, W. Huber, L. M. Steinmetz, and et al., “Translation of Small Open Reading Frames within Unannotated RNA Transcripts in Saccharomyces cerevisiae,” *Cell Rep.*, vol. 7, no. 6, pp. 1858–1866, Jun. 2014.

[143] C. D. S. Duncan and J. Mata, “The translational landscape of fission-yeast meiosis and sporulation.,” *Nat. Struct. Mol. Biol.*, vol. 21, no. 7, pp. 641–7, 2014.

[144] Z. Ji, R. Song, A. Regev, and K. Struhl, “Many lncRNAs, 5’UTRs, and pseudogenes are translated and some are likely to express functional proteins,” *Elife*, vol. 4, p. e08890, Dec. 2015.

[145] L. Calviello, N. Mukherjee, E. Wyler, H. Zauber, A. Hirsekorn, M. Selbach, M. Landthaler, B. Obermayer, and U. Ohler, “Detecting actively translated open reading frames in ribosome profiling data,” *Nat. Methods*, vol. 13, no. December, pp. 1–9, 2015.

[146] S. Y. Chun, C. M. Rodriguez, P. K. Todd, and R. E. Mills, “SPECtre: a spectral coherence--based classifier of actively translated transcripts from ribosome profiling sequence data,” *BMC Bioinformatics*, vol. 17, no. 1, p. 482, Dec. 2016.

[147] B. Malone, I. Atanassov, F. Aeschimann, X. Li, H. Großhans, and C. Dieterich, “Bayesian prediction of RNA translation from ribosome profiling,” *Nucleic Acids Res.*, vol. 26, no. 6, p. gkw1350, Jan. 2017.

[148] J. Crappe, E. Ndah, A. Koch, S. Steyaert, D. Gawron, S. De Keulenaer, E. De Meester, T. De Meyer, W. Van Criekinge, P. Van Damme, and G. Menschaert, “PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration,” *Nucleic Acids Res.*, vol. 43, no. 5, pp. e29–e29, Mar. 2015.

[149] E. de Klerk, I. F. A. C. Fokkema, K. A. M. H. Thiadens, J. J. Goeman, M. Palmblad, J. T. den Dunnen, M. von Lindern, and P. A. C. ’t Hoen, “Assessing the translational landscape of myogenic differentiation by ribosome profiling,” *Nucleic Acids Res.*, vol. 43, no. 9, pp. 4408–4428, May 2015.

[150] A. Zupanic, C. Meplan, S. N. Grellscheid, J. C. Mathers, T. B. L. Kirkwood, J. E. Hesketh, and D. P. Shanley, “Detecting translational regulation by change point analysis of ribosome profiling data sets.,” *RNA*, vol. 20, no. 10, pp. 1507–18, Oct. 2014.

[151] A. M. Michel, K. R. Choudhury, A. E. Firth, N. T. Ingolia, J. F. Atkins, and P. V Baranov, “Observation of dually decoded regions of the human genome using ribosome profiling data.,” *Genome Res.*, vol. 22, no. 11, pp. 2219–29, Nov. 2012.

[152] C. G. Artieri and H. B. Fraser, “Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation.,” *Genome Res.*, vol. 24, no. 12, pp. 2011–21, Dec. 2014.

[153] C. E. Gamble, C. E. Brule, K. M. Dean, S. Fields, E. J. Grayhack, J. M. Barral, M. S. Sachs, Y. Liu, I. Furman, Y. Pilpel, and J. Coller, “Adjacent Codons Act in Concert to Modulate Translation Efficiency in Yeast,” *Cell*, vol. 166, no. 3, pp. 679–690, Jul. 2016.

[154] S. Zhang, H. Hu, J. Zhou, X. He, T. Jiang, and J. Zeng, “ROSE: a deep learning based framework for predicting ribosome stalling,” *bioRxiv*, 2016.

[155] A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, H. F. Curators, B. D. G. Project, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M. V. Han, M. L.

Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, and M. Kellis, "Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures," *Nature*, vol. 450, no. 7167, pp. 219–232, Nov. 2007.

[156] B. P. Lewis, C. B. Burge, D. P. Bartel, S. M. Cohen, B. Bartel, D. P. Bartel, H. R. Horvitz, G. Ruvkun, N. Rajewsky, P. Rorsman, and M. Stoffel, "Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets," *Cell*, vol. 120, no. 1, pp. 15–20, Jan. 2005.

[157] S. D. Mackowiak, H. Zauber, C. Bielow, D. Thiel, K. Kutz, L. Calviello, G. Mastrobuoni, N. Rajewsky, S. Kempa, M. Selbach, and B. Obermayer, "Extensive identification and analysis of conserved small ORFs in animals," *Genome Biol.*, vol. 16, no. 1, p. 179, Dec. 2015.

[158] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.," *Genome Res.*, vol. 15, no. 8, pp. 1034–50, Aug. 2005.

[159] M. F. Lin, I. Jungreis, and M. Kellis, "PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions," *Bioinformatics*, vol. 27, no. 13, pp. i275–i282, Jul. 2011.

[160] A. Nekrutenko, K. D. Makova, and W.-H. Li, "The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study.," *Genome Res.*, vol. 12, no. 1, pp. 198–202, Jan. 2002.

[161] N. Goldman and Z. Yang, "A codon-based model of nucleotide substitution for protein-coding DNA sequences.," *Mol. Biol. Evol.*, vol. 11, no. 5, pp. 725–36, Sep. 1994.

[162] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li, "CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model," *Nucleic Acids Res.*, vol. 41, no. 6, pp. e74–e74, Apr. 2013.

[163] A. I. Nesvizhskii, "A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics," *Journal of Proteomics*, vol. 73, no. 11. pp. 2092–2123, 2010.

[164] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, and M. Mann,

“Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment,” *J. Proteome Res.*, vol. 10, no. 4, pp. 1794–1805, Apr. 2011.

[165] J. Cox and M. Mann, “MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification,” *Nat. Biotechnol.*, vol. 26, no. 12, pp. 1367–1372, Dec. 2008.

[166] S. Kim, P. A. Pevzner, R. Burke, D. Agus, and P. Mallick, “MS-GF+ makes progress towards a universal database search tool for proteomics,” *Nat. Commun.*, vol. 5, p. 5277, Oct. 2014.

[167] J. E. Elias and S. P. Gygi, “Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.,” *Nat. Methods*, vol. 4, no. 3, pp. 207–14, Mar. 2007.

[168] M. Vaudel, J. M. Burkhart, R. P. Zahedi, E. Oveland, F. S. Berven, A. Sickmann, L. Martens, and H. Barsnes, “PeptideShaker enables reanalysis of MS-derived proteomics data sets,” *Nat. Biotechnol.*, vol. 33, no. 1, pp. 22–24, 2015.

[169] B. Bogdanow, H. Zauber, and M. Selbach, “Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides,” *Mol. Cell. Proteomics*, vol. 15, no. 8, pp. 2791–2801, Aug. 2016.

[170] G. M. Knudsen, R. J. Chalkley, C. Wick, M. Stanford, and A. Zulich, “The Effect of Using an Inappropriate Protein Database for Proteomic Data Analysis,” *PLoS One*, vol. 6, no. 6, p. e20873, Jun. 2011.

[171] J. M. Chick, D. Kolippakkam, D. P. Nusinow, B. Zhai, R. Rad, E. L. Huttlin, and S. P. Gygi, “A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides,” *Nat Biotech*, vol. 33, no. 7, pp. 743–749, 2015.

[172] M. Vaudel, K. Verheggen, A. Csordas, H. Raeder, F. S. Berven, L. Martens, J. A. Vizcaíno, and H. Barsnes, “Exploring the potential of public proteomics data,” *Proteomics*, vol. 16, no. 2, pp. 214–225, Jan. 2016.

[173] J. D. Venable, M.-Q. Dong, J. Wohlschlegel, A. Dillin, and J. R. Yates, “Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra,” *Nat. Methods*, vol. 1, no. 1, pp. 39–45, Oct. 2004.

[174] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, “Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.,” *Mol. Cell. Proteomics*, vol. 11, no. 6, p. O111.016717, Jun. 2012.

[175] A. I. Nesvizhskii, “Proteogenomics: concepts, applications and computational strategies,” *Nat. Methods*, vol. 11, no. 11, pp. 1114–1125, 2014.

[176] L. Martens and J. A. Vizcaíno, “A Golden Age for Working with Public Proteomics Data,” *Trends Biochem. Sci.*, 2017.

[177] L. Florens, M. J. Carozza, S. K. Swanson, M. Fournier, M. K. Coleman, J. L. Workman, and M. P. Washburn, “Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors,” *Methods*, vol. 40, no. 4, pp. 303–311, 2006.

[178] C. Schmidt, C. Lenz, M. Grote, R. Luhrmann, and H. Urlaub, “Determination of Protein Stoichiometry within Protein Complexes Using Absolute Quantification and Multiple Reaction Monitoring,” *Anal. Chem.*, vol. 82, no. 7, pp. 2784–2796, 2010.

[179] A. N. Kettenbach, J. Rush, and S. A. Gerber, “Absolute quantification of protein and post-translational modification abundance with stable isotope–labeled synthetic peptides,” *Nat. Protoc.*, vol. 6, no. 2, pp. 175–186, 2011.

[180] K. Wethmar, A. Barbosa-Silva, M. A. Andrade-Navarro, and A. Leutz, “uORFdb--a comprehensive literature database on eukaryotic uORF biology.,” *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D60-7, Jan. 2014.

[181] A. Pauli, M. L. Norris, E. Valen, G.-L. Chew, J. a Gagnon, S. Zimmerman, A. Mitchell, J. Ma, J. Dubrulle, D. Reyon, S. Q. Tsai, J. K. Joung, A. Saghatelian, and A. F. Schier, “Toddler: an embryonic signal that promotes cell movement via Apelin receptors.,” *Science*, vol. 343, no. 6172, p. 1248636, 2014.

[182] M. Eravci, C. Sommer, and M. Selbach, “IPG strip-based peptide fractionation for shotgun proteomics.,” *Methods Mol. Biol.*, vol. 1156, pp. 67–77, Jan. 2014.

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

[184] P. Y. Hsu, L. Calviello, H.-Y. L. Wu, F.-W. Li, C. J. Rothfels, U. Ohler, and P. N. Benfey, “Super-Resolution Ribosome Profiling Reveals Novel Translation Events in Arabidopsis,” *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 45, p. In Revision, 2016.

[185] C. Merchante, J. Brumos, J. Yun, Q. Hu, K. R. Spencer, P. Enríquez, B. M. Binder, S. Heber, A. N. Stepanova, and J. M. Alonso, “Gene-Specific Translation Regulation Mediated by the Hormone-Signaling Molecule EIN2,” *Cell*, vol. 163, no. 3, pp. 684–697, 2015.

[186] M.-J. Liu, S.-H. S.-H. Wu, J.-F. Wu, W.-D. Lin, Y.-C. Wu, T.-Y. Tsai, H.-L. Tsai, and

S.-H. S.-H. Wu, “Translational landscape of photomorphogenic Arabidopsis.,” *Plant Cell*, vol. 25, no. 10, pp. 3699–710, 2013.

[187] P. Juntawong, T. Girke, J. Bazin, and J. Bailey-Serres, “Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 1, pp. E203-12, 2014.

[188] R. a. Jorgensen and A. E. Dorantes-Acosta, “Conserved Peptide Upstream Open Reading Frames are Associated with Regulatory Genes in Angiosperms,” *Front. Plant Sci.*, vol. 3, no. August, pp. 1–11, 2012.

[189] J. T. Clarke, R. C. M. Warnock, and P. C. J. Donoghue, “Establishing a time-scale for plant evolution,” *New Phytol.*, vol. 192, no. 1, pp. 266–301, 2011.

[190] D. E. Andreev, P. B. F. O’Connor, G. Loughran, S. E. Dmitriev, P. V. Baranov, and I. N. Shatsky, “Insights into the mechanisms of eukaryotic translation gained with ribosome profiling,” *Nucleic Acids Res.*, vol. 45, no. 2, pp. 513–526, Jan. 2017.

[191] I. P. Ivanov, A. E. Firth, A. M. Michel, J. F. Atkins, and P. V. Baranov, “Identification of evolutionarily conserved non-AUG-initiated N-terminal extensions in human coding sequences,” *Nucleic Acids Res.*, vol. 39, no. 10, pp. 4220–4234, 2011.

[192] A. Sendoel, J. G. Dunn, E. H. Rodriguez, S. Naik, N. C. Gomez, B. Hurwitz, J. Levorse, B. D. Dill, D. Schramek, H. Molina, J. S. Weissman, and E. Fuchs, “Translation from unconventional 5′ start sites drives tumour initiation,” *Nature*, vol. 541, no. 7638, pp. 494–499, Jan. 2017.

[193] A. S. Y. Lee, P. J. Kranzusch, and J. H. D. Cate, “eIF3 targets cell-proliferation messenger RNAs for translational activation or repression,” *Nature*, vol. 522, no. 7554, pp. 111–114, Apr. 2015.

[194] E. C. Swart, V. Serra, G. Petroni, M. Nowacki, A. Liang, Y. Zhou, J. S. Khurana, A. D. Goldman, M. Nowacki, K. Schotanus, and et al., “Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination,” *Cell*, vol. 166, no. 3, pp. 691–702, Jul. 2016.

[195] C. Barbosa, I. Peixeiro, and L. Romão, “Gene Expression Regulation by Upstream Open Reading Frames and Human Disease,” *PLoS Genet.*, vol. 9, no. 8, pp. 1–12, 2013.

[196] J. G. Dunn, C. K. Foo, N. G. Belletier, E. R. Gavis, and J. S. Weissman, “Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster,” *Elife*, vol. 2013, no. 2, 2013.

[197] N. R. Guydosh, R. Green, K. Maier, D. Schulz, S. Dümcke, B. Zacher, A. Mayer, J. Sydow, L. Marcinowski, L. Dölken, and et al., “Dom34 Rescues Ribosomes in 3′

Untranslated Regions,” *Cell*, vol. 156, no. 5, pp. 950–962, Feb. 2014.

[198] E. E. Heyer, M. J. Moore, D. Inzé, L. De Veylder, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, Y. Pilpel, and J. Coller, “Redefining the Translational Status of 80S Monosomes,” *Cell*, vol. 164, no. 4, pp. 757–769, Feb. 2016.

[199] P. Willems, E. Ndah, V. Jonckheere, S. Stael, A. Sticker, L. Martens, F. Van Breusegem, K. Gevaert, and P. Van Damme, “N-terminal proteomics assisted profiling of the unexplored translation initiation landscape in Arabidopsis thaliana.,” *Mol. Cell. Proteomics*, p. mcp.M116.066662, Apr. 2017.

[200] D. Gawron, E. Ndah, K. Gevaert, and P. Van Damme, “Positional proteomics reveals differences in N-terminal proteoform stability,” *Mol. Syst. Biol.*, vol. 12, no. 2, pp. 858–858, Feb. 2016.

[201] M. L. Tress, F. Abascal, and A. Valencia, “Alternative Splicing May Not Be the Key to Proteome Complexity,” *Trends Biochem. Sci.*, vol. 42, no. 2, pp. 98–110, 2017.

[202] B. J. Blencowe, “The Relationship between Alternative Splicing and Proteomic Complexity.,” *Trends Biochem. Sci.*, vol. 4, no. 0, p. e07794, May 2017.

[203] M. L. Tress, F. Abascal, and A. Valencia, “Most Alternative Isoforms Are Not Functionally Important,” *Trends in Biochemical Sciences*. 2017.

[204] Y. Liu, A. Beyer, R. Aebersold, J. S. Parker, D. N. Hayes, C. M. Perou, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, N. CPTAC, and et al., “On the Dependency of Cellular Protein Levels on mRNA Abundance,” *Cell*, vol. 165, no. 3, pp. 535–550, Apr. 2016.

[205] N. Mukherjee, L. Calviello, A. Hirsekorn, S. de Pretis, M. Pelizzola, and U. Ohler, “Integrative classification of human coding and noncoding genes through RNA metabolism profiles,” *Nat. Struct. Mol. Biol.*, vol. 24, no. 1, pp. 86–96, Nov. 2016.

[206] X. Yang, J. Coulombe-Huntington, S. Kang, G. M. Sheynkman, T. Hao, A. Richardson, S. Sun, F. Yang, Y. A. Shen, R. R. Murray, K. Spirohn, B. E. Begg, M. Duran-Frigola, A. MacWilliams, S. J. Pevzner, Q. Zhong, S. A. Trigg, S. Tam, L. Ghamsari, N. Sahni, S. Yi, M. D. Rodriguez, D. Balcha, G. Tan, M. Costanzo, B. Andrews, C. Boone, X. J. Zhou, K. Salehi-Ashtiani, B. Charloteaux, A. A. Chen, M. A. Calderwood, P. Aloy, F. P. Roth, D. E. Hill, L. M. Iakoucheva, Y. Xia, and M. Vidal, “Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing,” *Cell*, vol. 164, no. 4, pp. 805–817, 2016.

[207] M. T. Couvillion, I. C. Soto, G. Shipkovenska, and L. S. Churchman, “Synchronized mitochondrial and cytosolic translation programs,” *Nature*, vol. 533, no. 1, pp. 1–17,

2016.

[208] P. Chotewutmontri, A. Barkan, S. Salzberg, T. Fennell, J. Ruan, and N. Homer, “Dynamics of Chloroplast Translation during Chloroplast Differentiation in Maize,” *PLOS Genet.*, vol. 12, no. 7, p. e1006106, Jul. 2016.

[209] K. Rahim, “Applications of Multitaper Spectral Analysis to Nonstationary Data,” 2014.

[210] C. Chen, Z. Li, H. Huang, B. E. Suzek, and C. H. Wu, “A fast Peptide Match service for UniProt Knowledgebase.,” *Bioinformatics*, vol. 29, no. 21, pp. 2808–9, Nov. 2013.

[211] D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, and D. S. Rokhsar, “Phytozome: A comparative platform for green plant genomics,” *Nucleic Acids Res.*, vol. 40, no. D1, 2012.

[212] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, “BLAST+: architecture and applications,” *BMC Bioinformatics*, vol. 10, p. 421, 2009.

[213] R. C. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.

[214] A. Larsson, “AliView: A fast and lightweight alignment viewer and editor for large datasets,” *Bioinformatics*, vol. 30, no. 22, pp. 3276–3278, 2014.

Appendix A: The Slepian Sequences and the multitaper F-test

As seen in Section 2.3.5, in the multitaper method we average the effect of multiple windows on our original signal. The validity of this approach thus depends on the properties of the different window functions a and their Fourier transform A. As mentioned in Section 2.3.5, we want to ensure that the tapers do not cause excessive spectral leakage. The ideal window function should concentrate the energy within a specific frequency width W, minimizing the leakage over the rest of the frequency spectrum. For a signal of length N, this concept can be formulated as follows (Equation 23):

$$\lambda(N, W) = \frac{\int_{-W}^{W} |A(f)|^2 df}{\int_{-1/2}^{1/2} |A(f)|^2 df} \quad (23)$$

We want to maximize λ, which means maximizing the energy of A over the bandwidth W (numerator) with respect to the entire spectrum (denominator, ½ is the Nyquist frequency, assuming 1 as the sampling rate). Finding the window a which maximizes λ leads to the definition of the following differential equation (Equation 24):

$$\frac{\partial \lambda}{\partial a} = 0 \qquad (24)$$

This can in turn be rewritten[124] as (Equation 25):

$$D \cdot a = \lambda a \quad (25)$$

With D being a NxN matrix with components (Equation 26):

$$D_{xy} = \frac{\sin(2\pi W(x-y)}{\pi(x-y)} \qquad (26)$$

As D is a symmetric matrix, the solution to Equation 25 can be found using standard linear algebra methods, creating a set of N eigenvalues λ and N orthogonal eigenvectors v (using the same principle behind PCA on the covariance matrix). The eigenvectors are used as window functions in the multitaper method, and they are known as *Slepian sequences* (or *discrete* *prolate spheroidal sequences, DPSS*), studied by David Slepian and colleagues[125]. The N eigenvalues are used in the multitaper method as weights in the averaging procedure (Equation 27):

$$\hat{B}(f_0) = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{n=0}^{N-1} v_{kn} x_n \, e^{-i2\pi f_0 \, n/N}}{\lambda_k} \quad (27)$$

However, the eigenvalues λ_k are usually used to derive a different set of weights, using an iterative process initialized by the first two λ, known as adaptive weighting[124].

The use of the Slepian sequences lies at the core of the multitaper method, and determined its efficiency in both reducing noise in the PSD estimate and minimizing spectral leakage. Another useful feature of Slepian functions is their orthogonal nature: being uncorrelated, they can be used to provide independent realizations of the same original sample. This allows us to derive statistical confidence in the detection of different frequency components against the null hypotheses of random (or at least *locally* random) noise.

Using a slightly different notation, we can explain the signal x_n as a mixture of two components: one coming from the frequency f_0 with its associated coefficient B and a term η_n representing energy from other frequencies and noise (Equation 28).

$$x_n = Be^{i2\pi f_0 n} + \eta_n \quad (28)$$

We can estimate a mean μ for the coefficient $B(f_0)$ using a least-squares fit in the frequency domain (Equation 29), using the DFT A of the tapers functions a, where A_{k0} represents the amplitude at the zero-frequency of A_k (and can also be calculated using the total signal in a_k [209]):

$$\hat{\mu}(f_0) = \frac{\sum_{k=1}^{K} A_{k0} B_k(f_0)}{\sum_{k=1}^{K} A^2{}_{k0}} \quad (29)$$

The estimate $\hat{\mu}(f_0)$ quantifies the variation captured by our estimates $B(f_0)$ along the different tapers A_k, thus representing how robust a frequency coefficient is in the different (and independent) tapered signals.

This formula is analogous to the calculation of the slope of a simple regression $y \sim bx + \varepsilon$, where $b = Cov(x, y)/Var(x)$.

We can now define the variance of the A component as the sum of two terms (Equations 30 and 31):

1) The amount of variation captured by our estimate:

$$\theta(f_0) = \ |\hat{\mu}(f_0)|^2 \sum_{k=1}^{K} A^2{}_{k0} \quad (30)$$

2) The “unexplained” variance:

$$\psi(f_0) = \ \sum_{k=1}^{K} |B_k(f_0) - \hat{\mu}(f_0)A_{k0}|^2 \ (31)$$

Comparing the two estimated variance components enables to extract an F-test statistic [124], [127], [129], (Equation 32):

$$F(f_0) = (K - 1)\frac{\theta(f_0)}{\psi(f_0)} \quad (32)$$

F-values for each frequency bin can be converted into p-values using a variance-ratio test with 2 and 2K-2 degrees of freedom.

Appendix B: Supplementary Materials

B.1: The Ribo-seq protocol in HEK293

Ribosome profiling. We followed the original protocol[81] with minor modifications. For cell lysis, the cell medium was aspirated and cells were washed with ice-cold PBS containing 100 µg/ml cycloheximide. No cycloheximide was added to the culture medium before the wash. After thorough removal of the PBS, the plates were immersed in liquid nitrogen and placed on dried ice. For cell lysis, we dripped 400 µl of mammalian polysome buffer (20 mM Tris-HCl, pH 7.4, 150 mM NaCl, 5 mM MgCl2, with 1 mM DTT and 100 µg/ml cycloheximide added freshly) supplemented with 1% (vol/vol) Triton X-100 and 25 U/ml Turbo DNase (Life Technologies, AM2238) onto the plates and then placed the plates on wet ice. We scraped the cells off to the lower portion of the dish so that they thawed in lysis buffer. After dispersal of the cells by pipetting, the lysate was triturated ten times through a 26-gauge needle, cleared by centrifugation at 20,000 g for 5 min, flash-frozen in liquid nitrogen and stored at −80 °C until further use. For isolation of ribosome-protected RNA fragments, 120 µl of the lysate were digested with 3 µl of RNase I (Life Technologies, AM2294) for 45 min at room temperature with rotation. Digestion was stopped by the addition of 4 µl of Superase-In (Life Technologies, AM2694). Meanwhile, MicroSpin S-400 HR columns (GE Healthcare, 27-5140-01) were equilibrated with 3 ml of mammalian polysome buffer by gravity flow and emptied by centrifugation at 600g for 4 min. We then immediately loaded 100 µl of the digested lysate on the column and eluted the column by centrifugation at 600g for 2 min. We extracted RNA from the flow-through (approximately 125 µl) using Trizol LS (Life Technologies, 10296-010). We then removed ribosomal RNA fragments using the RiboZero Kit (Illumina, MRZH11124) and separated them on a 17% denaturing urea-PAGE gel (National Diagnostics, EC-829). The size range from 27nt to 30nt, defined by loading with 20 pmol each of Marker-27 nt and Marker-30 nt, was cut out, and the RNA fragments were subjected to library generation using 3'adaptor NN-RA3, 5'adaptor OR5-NN, RT primer RTP and PCR primers RP1 (forward primer) and RPI6-7 (reverse primer, containing barcodes). Libraries were sequenced on a HiSeq 2000 device (Illumina). After initial quality control, we obtained ~29 Million raw reads by pooling the RPI6 and RPI7 samples. The following primers were used:

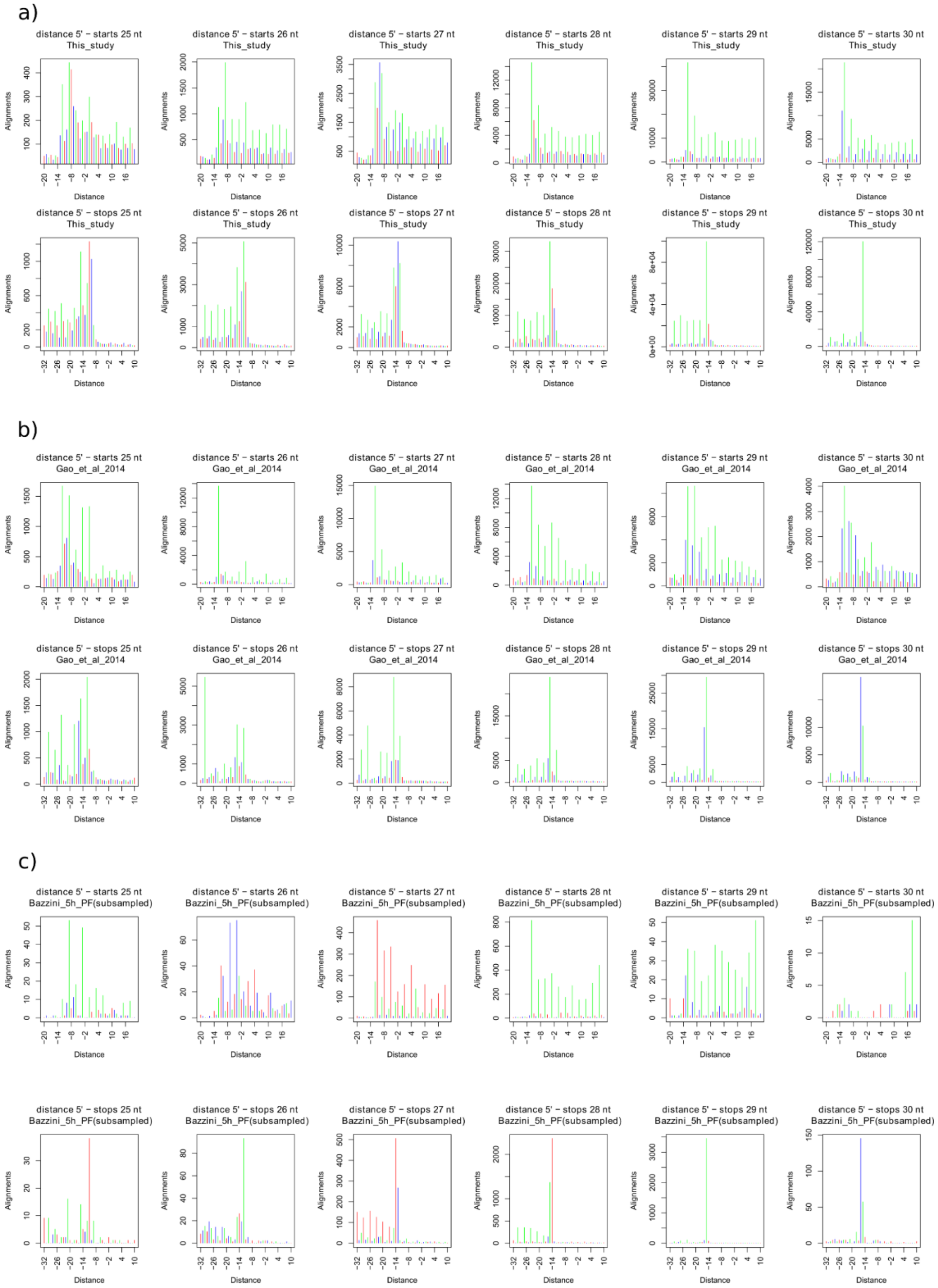
Marker-27 nt, 5'-AUGUACACGGAGUCGAGCUCAACCCGC-P;
Marker-30 nt, 5'-AUGUACACGGAGUCGAGCUCAACCCGCAAC-P;

NN-RA3, 5’-P NNTGGAATTCTCGGGTGCCAAGG-InvdT;
OR5-NN, 5’-GUUCAGAGUUCUACAGUCCGACGAUCNN;
RTP, 5’-GCCTTGGCACCCGAGAATTCCA;
RP1, 5’-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA;
RPI6,5’-
CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA;
RPI7,5’-
CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA.

B.2: Ribo-seq and RNA-seq data processing.

Ribo-seq reads were stripped of their adapters using cutadapt. Randomized UMI sequences were removed and reads were collapsed using custom perl scripts. Reads aligning to rRNA sequences were removed with Bowtie. Unaligned reads were then mapped with STAR using the hg19 genome and the GENCODE 19 annotation in GTF format. For zebrafish, transcript structures annotated in Ensembl (version 76) were used. For both RNA-seq and Ribo-seq, a maximum of four mismatches was allowed, and multimapping of to up to eight different positions was permitted. Alignments flagged as secondary alignments were filtered out, ensuring one genomic position per aligned read. RSEM was run using default parameters. The hg38 version of the human genome, supplied with GENCODE 25 annotation, was used for analyses in Section 3.4.

B.3: Supplementary Figure 1: Metagene analysis for different Ribo-seq datasets



Supplementary Figure 1. Metagene analysis for different Ribo-seq datasets. Aggregate plots for different read lengths (from 25 to 30 nt) are shown, showing distance between 5'ends and annotated start and stop codons. Distinct profiles, in terms of both precision and coverage, emerge in the different datasets. a) HEK293, this study; b) HEK293, Gao et al; c) Zebrafish 5h post-fertilization, Bazzini et al, 2014

B.4: multitaper analysis

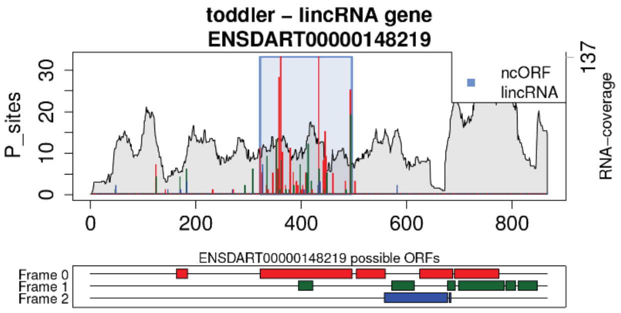
The original multitaper algorithm from Thomson is implemented in R in the package “multitaper”[129]. For each track, we added a stretch of zeros to the input sample to reach a minimum length of 50 nt. We ran the multitaper with 24 tapers, with the time-window parameter set to 12. Moreover, sequences shorter than 500nt were zero-padded to 1,024 data points before the discrete Fourier transform was computed, to obtain adequate frequency resolution in the spectrum. We extracted F-values from the frequency bin closest to 3nt periodicity. Using the *pf* function, we calculated p-values from the F-statistic by using 2 and 2k-2 degrees of freedom, where k is the number of tapers (24 in this study). ORFs and exons with fewer than six P-sites or shorter than 6nt were ignored.

For the simulation tests, we sampled 1,000 CCDS exons from different read lengths and coverage as a positive set. For each exon, we randomly shuffled the P-site positions 1,000 times to obtain a negative set.

B.5: QTI-seq analysis

For every reported QTI-seq peak[84], we selected the closest ORF called by RiboTaper on the basis of the reported distance relative to the annotated start codon. Only AUG start codons were used.

B.6: Supplementary Figure 2: The *toddler* ncORF



Supplementary Figure 2: The *toddler* ncORF. Shown are P-sites positions, RNA-seq coverage and ORF position. Data from Bazzini et al, 2014

B.7: Supplementary Table 1: RiboTaper-detected ORFs in Danio rerio.

Dataset	uORFs	ORFs_ccds	dORFs	lincRNA	Pseudogenes	Processed Transcript
5h_pf_1	64	12624	9	17	2	59

Supplementary Table 1: RiboTaper-detected ORFs in *Danio rerio*. Shown are the genes harboring translated ORFs, divided by ORF category/biotype.

B.8: Evolutionary conservation analysis

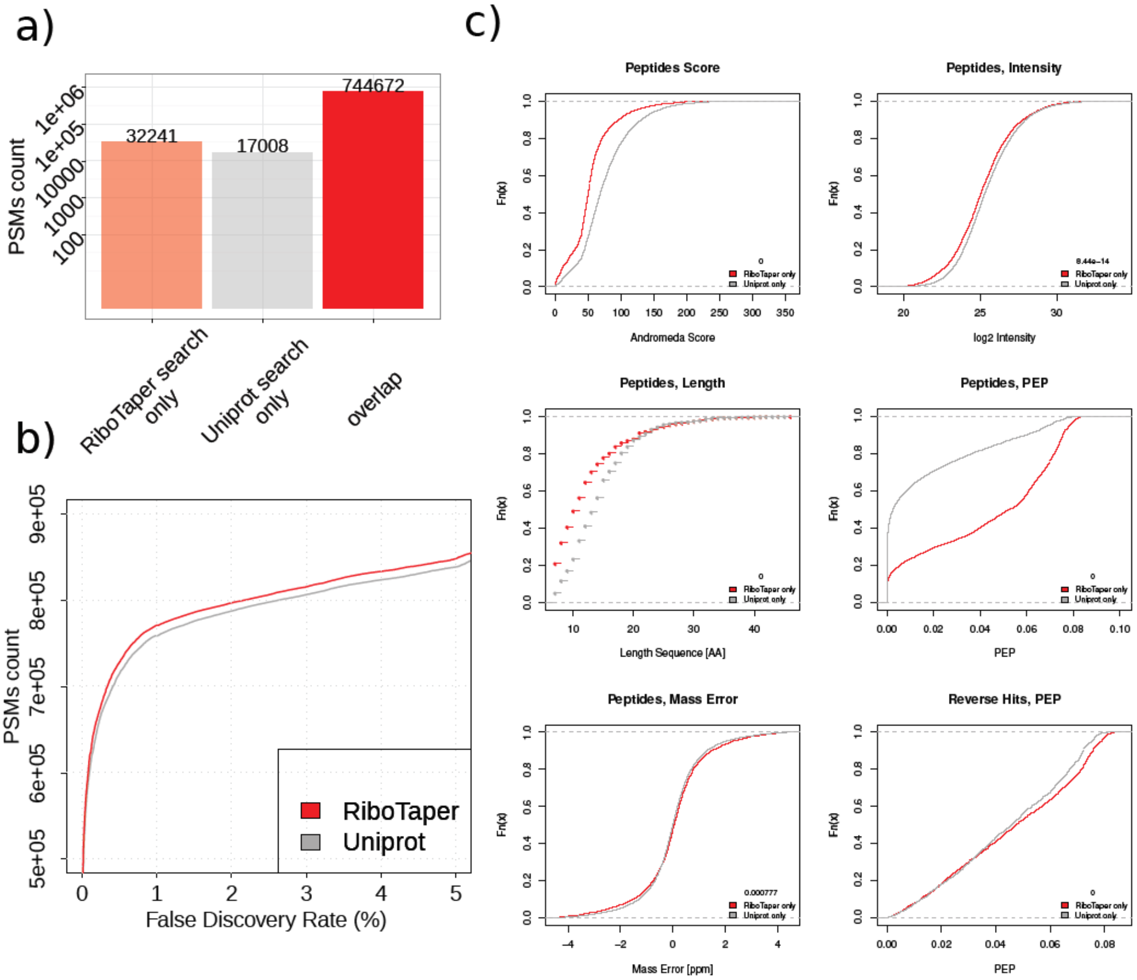
PhastCons scores were extracted as the average over the entire ORF or in 25nt windows around the start and stop codons. ORFs were then scored with PhyloCSF in the “mle” (default) mode, using the “29mammals” parameter set on the 46-vertebrate alignment to the human genome (hg19) after alignment filtering steps as described by Bazzini *et al*[130] . We additionally used the hexamer score from the CPAT tool to assess the coding potential of different ORFs, using the available trained model for the human genome. For each category, the scores were compared against a control set of ORFs matching the length and conservation of the category of interest. For CCDS ORFs and non-CCDS coding ORFs, we selected ORFs shorter than 300nt as meaningful matching controls. SNPs were downloaded as *.gvf* files from Ensembl (v75, 1000 Genomes phase 1). We removed SNPs in reverse orientation, SNPs falling into genomic repeats (using the RepeatMasker track from the UCSC genome browser, March 22, 2015) and rare SNPs with a derived allele frequency of <1%. We then recorded for each ORF and its conceptual translation the number of synonymous and nonsynonymous SNPs compared to the human reference genome, as well as the number of synonymous and non-synonymous sites derived from the degeneracy of the genetic code. For every set of ORFs, we aggregated these numbers and calculated the dN/dS ratio, where dN is the number of non-synonymous SNPs per non-synonymous site and dS is the number of synonymous SNPs per synonymous site. For the CPAT and PhyloCSF scores, p-values were determined via Wilcoxon-Mann-Whitney tests. For dN/dS ratios, p-values were determined via χ2 test, using as expected frequencies the values from SNPs falling in synonymous and non-synonymous sites in the ORF control set, as described previously[157].

B.9: Mass spectrometry preparation and data analysis.

The proteomic data for HEK293 cells was published recently[182] (PRIDE accession PXD002389). Briefly, cells were grown in DMEM (Life Technologies). Lysis was performed in 50 mM ammonium bicarbonate buffer (pH 8.0) containing 2% SDS and 0.1 M DTT. Sulfhydryl groups were alkylated by iodoacetamide added to a final concentration of 0.25 M and incubated for 20 min. Proteins were precipitated; resuspended in 6 M urea, 2 M thiourea and 10 mM HEPES; and digested into peptides by LysC (3 h) and trypsin (overnight, diluted four times with 50 mM ammonium bicarbonate buffer). Peptides were desalted using StageTip purification and subsequently analyzed by online LC-MS/MS on a Q-Exactive instrument (Thermo Fisher) using nano-electrospray ionization. The resolution was set to 70,000 and 17,500 for full and fragment scans, respectively. We identified peptides from MS/MS spectra by searching against the recent UniProt human database (2014-10) or the newly generated HEK293-specific database using ribosome profiling with MaxQuant[165] version 1.5.2.8. For all searches, carbamidomethylation (C) was set as a fixed modification, and oxidation (M), acetylation (protein N-term) and deamidation (NQ) were set as variable modifications. A maximum of two missed cleavages was allowed. The peptide FDR was set to 0.01, the minimal peptide length was set to 7 amino acids, and the main search peptide tolerance was set to 4.5 ppm.

We built custom peptide databases by using all the identified ORFs before filtering for multimapping reads. The FDR was calculated based on the ratio of hits in the positive and decoy databases. Counts and feature distribution (PEP, score, sequence length) from evidence files were compared on the basis of an FDR < 1%, excluding reverse hits and contaminants as well as unique sequence information. iBAQ values were calculated using MaxQuant. Non-UniProt peptide sequences were defined using PeptideMatch[210].

B.10: Supplementary Figure 3: Additional statistics about Ribotaper- and Uniprot-only identified peptides.



Supplementary Figure 3. Additional statistics about Ribotaper- and Uniprot-only identified peptides. a) Overlap of Peptide Spectrum Matches (PSMs) identified in the two strategies. b) False Discovery Rate (FDR) vs. PSMs counts for the two search strategies. c) Comparisons of RiboTaper-only identified peptides vs. Uniprot-only identified peptides (PEP=Posterior Error Probability).

B.11: Ribo-seq data processing in *Arabidopsis Thaliana*

After adapter removal, Ribo-seq reads were searched for expected contaminant RNA sequences in *Arabidopsis*, including rRNA, tRNA, and snoRNA sequences, using bowtie2 (parameter: –L 20). The unaligned reads were then mapped to the *Arabidopsis* genome using TAIR10.29 reference using STAR, version 2.5.1b, allowing up to three mismatches and a maximum of 20 multimapping positions. The best position for multimapping reads was chosen by STAR with these options "–outSAMmultNmax 1 – outMultimapperOrder Random" which randomly selects one alignment over all of the possible best scoring alignments. Raw sequencing data from Juntawong et al. (whole seedlings of 7-d-old nonstress plants; the GEO accession GSE50597, GSM1224475 and GSM1224476), Liu et al. (aerial part of 4-d-old etiolated seedlings exposed to light for 4 h; the GEO accession GSE43703), and Merchante et al. (72-h-old etiolated seedlings with normal air; the SRA accession SRP056795) were downloaded and processed using the same procedures described above. The sequencing summary of each dataset is provided in Supplementary Table 2. Replicates of the same tissue in each dataset were pooled for analysis to achieve high coverage, unless otherwise specified. Several steps of analysis, including calculating correlation among the replicates, assigning reads to genomic features (5'-UTR, CDS, 3'UTR, introns, and intergenic regions), meta-gene analysis over start/stop codons defined by TAIR10 protein-coding genes, and statistical presentation of the data were performed and plotted in R, version 3.2.3, using various R packages (Supplementary Material). Because not all of the datasets to which we compared our results used a strand-specific library construction method and some of the datasets were generated with different mRNA enrichment methods for RNA-seq, TPM values for RNA-seq data on protein-coding genes only were determined by RSEM[98], version 1.2.11, using a non-strand-specific parameter. For comparison with the same sequencing depth among the datasets, 25 million reads were randomly sampled from each Ribo-seq dataset using inhouse scripts.

B.12: Supplementary Table 2: Mapping statistics for the different libraries analyzed in *Arabidopsis thaliana*.

dataset	library	total_reads	rRNA %	snoRNA & tRNA %	remaining reads	mapped reads	Uniquely mapping reads %	Multi-mapping reads %
Hsu	ribo_R1	149,405,767	49.7%	2.1%	72,030,234	52,354,522	91%	9%
Hsu	ribo_R2	143,897,003	38.5%	2.8%	84,385,394	48,625,232	91%	9%
Hsu	ribo_R3	154,597,025	50.6%	5.2%	68,401,924	54,810,821	88%	12%
Hsu	ribo_S1	136,655,109	22.5%	2.3%	102,835,491	32,608,135	85%	15%
Hsu	ribo_S2	169,463,792	36.6%	3.1%	102,050,409	78,838,023	91%	9%
Hsu	ribo_S3	121,836,865	41.3%	2.8%	68,127,544	51,613,460	91%	9%
Hsu	RNA_R1	40,251,993	2.0%	36.8%	24,643,106	22,652,994	74%	26%
Hsu	RNA_R2	49,692,907	1.9%	36.5%	30,610,968	28,350,800	75%	25%
Hsu	RNA_R3	52,556,216	2.0%	36.9%	32,082,943	29,788,222	74%	26%
Hsu	RNA_S1	35,587,149	0.8%	37.7%	21,885,720	20,133,715	71%	29%
Hsu	RNA_S2	56,545,985	0.8%	35.1%	36,272,859	33,576,115	72%	28%
Hsu	RNA_S3	59,009,561	0.8%	36.2%	37,159,151	34,615,016	71%	29%
Juntawong	ribo_noStress1	93452469	63.7%	2.9%	31,235,427	28,962,468	71%	29%
Juntawong	ribo_noStress2	89907795	49.4%	1.6%	44,075,153	37,771,475	64%	36%
Juntawong	RNA_noStress1	25,384,079	46.8%	0.1%	13,487,403	12,889,887	92%	8%

Juntawong	RNA_noStress2	55,925,575	56.3%	0.1%	24,397,043	22,027,558	93%	7%
Liu	ribo_Light1	144,479,049	92.6%	0.9%	9,356,737	8,095,138	67%	33%
Liu	ribo_Light2	149,173,203	64.8%	1.2%	50,692,754	49,027,433	90%	10%
Liu	RNA_Light1	55,762,418	11.1%	0.1%	49,504,687	47,597,072	95%	5%
Liu	RNA_Light2	67,275,771	22.3%	0.0%	52,228,401	51,087,039	95%	5%
Merchante	ribo_Col_Air1	126,326,120	46.0%	0.3%	67,809,572	64,160,828	86%	14%
Merchante	ribo_Col_Air2	103,257,454	66.2%	0.7%	34,215,679	31,064,050	64%	36%
Merchante	RNA_Col_Air1	98,461,958	73.6%	0.5%	25,411,162	23,763,204	87%	13%
Merchante	RNA_Col_Air2	143,960,373	81.6%	0.4%	25,882,228	23,798,627	82%	18%

Supplementary Table 2: Mapping statistics for the different libraries analyzed in Arabidopsis thaliana.

B.13: Supplementary Table 3: Read lengths and cutoffs used to infer P-sites position in *Arabidopsis thaliana*.

	Footprint length	Offset to P-site position
This study	22,23,24,25,26,27,28	6,7,8,9,10,11,12
Juntawong et al. 2014*	25,26,27,28,29,30,31,32,33,34	0,0,0,0,0,0,0,0,0,0
Liu et al. 2013	28,29,32,33,34	12,13,13,14,14
Merchante et al. 2015	23,24,25,28,30,31,32,33	6,7,8,12,13,14,14,14

Supplementary Table 3: Read lengths and cutoffs used to infer P-sites position in *Arabidopsis thaliana*.

B.14: Protein Alignments

Whole-genome assemblies of 15 selected species were downloaded from Phytozome[211], version 11.0. To search for homologs in these genomes, we used tBLASTn[212] with the E-value set to 0.1. Hits were retained if they met the following criteria: (i) the hit-query BLAST alignment must cover at least 30% of the query protein sequence; and (ii) a complete ORF, from start to stop codon, must be present and its length must not be longer than the query by 120 aa. For each unannotated ORF and its homologs, we reconstructed multiple sequence alignments using MUSCLE[213] and visualized them in AliView[214] to manually validate the alignments. Pairwise sequence identities were calculated from the alignments by a custom Python script. All of the alignments and the Python script have been deposited in the Dryad Digital Repository (dx. doi.org/10.5061/dryad.m8jr7).

B.15: Polysome profiling, nuclear-cytoplasmic comparison and 5’end sequencing.

Polysome profiling:

DEXSeq was run to detect differential exon usage between each of the polysome fraction and the cytoplasmic abundance. Differential exons (FDR<0.01) were intersected with ORF coordinates and only exons uniquely mapping to one ORF group were retained. Only genes with multiple translated isoforms were used.

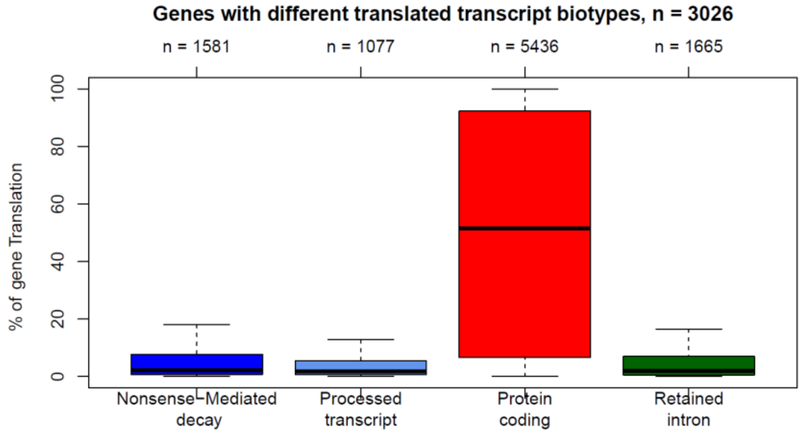
Nuclear-cytoplasmic comparison:

DEXSeq was run to detect differential exon usage between the nuclear and the cytoplasmic fraction. Differential exons (FDR<0.01) were intersected with transcript structures and only exons uniquely mapping to one of the transcript groups were selected.

5’end of endonucleolytic cuts:

Bigwig files for the different libraries were obtained from the GEO accession GSE57433. Coordinates were lifted to hg38 and overlapped with SaTAnn-identified stop codon positions, for both controls and NMD candidates. Stop codon regions of NMD candidates overlapping CDS regions were removed.

B.16: Supplementary Figure 4: Translation quantification on different transcript biotypes.



Supplementary Figure 4. Translation quantification on different transcript biotypes. Percentage of gene translation is shown for different transcript biotypes. Genes were selected when containing multiple translated transcript biotypes (numbers shown on top). Outliers are not shown.

Appendix C: List of main software used in this study.

samtools 1.3.1

bedtools 2.17

cutadapt 1.8

bowtie 1.1.2

bowtie2 2.2.6

STAR 2.5.1b

RSEM 1.2.11

MaxQuant 1.5.2.8

PeptideMatch 1.0

R packages:

multitaper_1.0-12

seqinr_3.3-0

XNomial_1.0.4

genomation_1.4.2

ggplot2_2.1.0

gplots_3.0.1

corrplot_0.77

doMC_1.3.4

foreach_1.4.3

GenomicAlignments_1.8.4

Rsamtools_1.24.0

SummarizedExperiment_1.2.3

BSgenome_1.40.1

rtracklayer_1.32.1

Biostrings_2.40.2

GenomicFeatures_1.24.4

AnnotationDbi_1.34.4

GenomicRanges_1.24.2

GenomicAlignments_1.8.4

Gviz_1.18.1

List of Publications

The following publications were authored during the PhD:

Extensive identification and analysis of conserved small ORFs in animals.
Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, **Calviello L**, Mastrobuoni G, Rajewsky N, Kempa S, Selbach M, Obermayer B.
Genome Biol. 2015 Sep 14;16:179. doi: 10.1186/s13059-015-0742-x.

Detecting actively translated open reading frames in ribosome profiling data.
Calviello L, Mukherjee N*, Wyler E*, Zauber H, Hirsekorn A, Selbach M, Landthaler M, Obermayer B, Ohler U.
Nat Methods. 2016 Feb;13(2):165-70. doi: 10.1038/nmeth.3688. Epub 2015 Dec 14.

Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis.
Hsu PY, **Calviello L**, Wu HL, Li FW, Rothfels CJ, Ohler U, Benfey PN.
Proc Natl Acad Sci U S A. 2016 Oct 21. pii: 201614788. [Epub ahead of print]

Integrative classification of human coding and noncoding genes through RNA metabolism profiles.
Mukherjee N, **Calviello L**, Hirsekorn A, de Pretis S, Pelizzola M, Ohler U.
Nat Struct Mol Biol. 2017 Jan;24(1):86-96. doi: 10.1038/nsmb.3325. Epub 2016 Nov 21.

RNA localization is a key determinant of neurite-enriched proteome
Zappulo1 A*, van den Bruck D*, Ciolli Mattioli C*, Franke V*, Imami K, McShane E, Moreno-Estelles M, **Calviello L**, Filipchyk A, Peguero-Sanchez E, Müller T, Woehler A, Birchmeier C, Merino E, Rajewsky N, Ohler U, Mazzoni EO, Selbach M, Akalin A, and Chekulaeva M
Nat Commun, in press

Beyond Read Counts: Ribo-seq data analysis to understand the functions of the transcriptome.
Calviello L, Ohler U
Trends in Genetics, under review

Annotate and quantify the translated transcriptome with SaTAnn
Calviello L, Zauber H, Hirsekorn A, Selbach M, Ohler U
in preparation

* equally contributing authors

Acknowledgments

Many are the people that should go here, but mostly I want to thank:

Uwe, for taking me on this (sometimes very challenging) adventure, and for his great support;
The reviewers, for reading and commenting over this thesis.
Neel, for constant supervision, discussions, and for making this PhD a fun experience!
Harm, for scientific discussions and jokes that I will surely miss;
Rebecca, for being always nice and helpful regardless of what I ask;
Other members of the lab (especially Katia, Antje, Scott and Mahmoud), for discussions and nice atmosphere;
Polly, for constant enthusiasm and collaboration;
Harm again, for reading and correcting my Google-translated abstract;
Dominique, for reading this dissertation and providing extremely useful comments;
Michaela Liemen and the PhD Office, for helping my poor understanding of the German bureaucracy;
Stack Overflow, for making my bash scripts semi-reliable;
The Bioconductor Project, for providing an excellent platform that revolutionized the way I do this job;
Stoner Meadow of Doom, The Psychedelic Muse, Open Culture, and Wikipedia, for providing excellent music and knowledge, for free and for everybody;
The (mostly Italian) gang of building 89, for providing many beri gut moments;
The BIMSB people (especially Jordi), for being a nice ensemble of humans;
Marta, Jackie, Rekado and Ibanez, for beautiful music and even more beautiful friendship;
My family and friends in Italy, for their support over this long-distance relationship;
You, Aslım, for being a constant loving inspiration for Everything I do. Let’s crack this code of life together!