



UNIVERSITY OF TRENTO - Italy

International PhD Program in Biomolecular Sciences
Centre for Integrative Biology
29th Cycle

**“UNDERSTANDING THE ORGANIZATION AND
FUNCTIONAL CONTROL OF POLYSOMES BY
INTEGRATIVE APPROACHES”**

Tutor

Gabriella VIERO

Institute of Biophysics - CNR, Trento (IT)

Advisor

Toma TEBALDI

Centre for Integrative Biology, University of Trento (IT)

Co-advisor

Guido SANGUINETTI

School of Informatics, University of Edinburgh (UK)

Ph.D. Thesis of

Fabio LAURIA

Institute of Biophysics - CNR, Trento (IT)

Academic Year 2015/2016

The work described in this thesis was carried out at the Laboratory of Translational Architectomics, Institute of Biophysics (National Research Council) of Trento (IT) and at the School of Informatics, University of Edinburgh (UK) between January 2014 and November 2017. I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged. This work has not been submitted previously for any other degree at the University of Trento or Edinburgh or any other university.

November 2017

Fabio Lauria

Contents

Abstract	7
1 Introduction.....	11
1.1 From DNA to proteins	11
1.2 Translation.....	11
1.3 Translation regulation	16
1.3.1 Codon usage bias	18
1.3.2 Ramp	19
1.4 Experimental techniques for study translation at the genome-wide level	20
1.4.1 Polysome profiling	22
1.4.2 Ribosome profiling.....	22
1.4.3 Atomic force microscopy	25
1.5 Mathematical models	26
1.5.1 Deterministic models.....	27
1.5.2 Stochastic models	29
1.5.3 Other models of translation	32
1.6 Translation in motor neuron diseases: the case of Spinal Muscular Atrophy	32
2 Mathematical models of translation.....	35
2.1 riboAbacus.....	35
2.2 riboSim	61
2.2.1 Materials and methods.....	62
2.2.2 Results.....	67
2.2.3 Conclusions	73
3 Development of tools for analysing ribosome profiling data	75
3.1 riboWaltz	75
3.2 riboScan.....	107
3.2.1 Materials and methods.....	109
3.2.2 Conclusions	112

4 Biological case study	113
4.1 Ribosome profiling datasets.....	114
4.2 Ribosome profiling of early-symptomatic SMA mouse brains	116
4.2.1 Ribosome profiling of actively translating ribosomes	116
4.2.2 Ribosome drop-off in SMA	121
4.2.3 Mislocalization of ribosomes along the 3' UTR in SMA?	124
4.3 Ribosome profiling of SMN-specialized ribosomes reveals a role for SMN in translation of the first codons.....	129
4.4 Loss of SMN-specialized ribosomes impacts on active translation in SMA	135
4.5 Conclusions.....	138
5 Discussion	141
Appendix	153
Bibliography	155
Acknowledgements	173

Abstract

Background and rationale

Translation is a fundamental biological process occurring in cells, carried out by ribosomes simultaneously bound to an mRNA molecule (polyribosomes). It has been exhaustively demonstrated that dysregulation of translation is implicated in a wide collection of pathologies including tumours and neurological disorders. Latest findings reveal the existence of translational regulatory mechanisms acting in *cis* or *trans* with respect to the mRNAs and governing the movement and the position of ribosomes along transcripts or directly impacting on the ribosome catalogue of its constituent proteins. For this reason, translational controls also account for widespread uncoupling between transcript and protein abundances in cells.

To explain the poor correlation between transcripts and protein levels, many computational models of translation have been developed. Usually, these approaches aim at predicting protein abundances in cells starting from the mRNA abundance. Despite the efforts of these modelling studies, a consensus model remains elusive, drawing to contradictory conclusions concerning the role of mRNA regulatory elements such as the usage of codons (codon usage bias) and slowdown mechanism at the beginning of the coding sequence (ramp). More recently, following the rapid and widespread diffusion of ribosome footprinting assays (RiboSeq), which enables the dissection of translation at single nucleotide resolution, a number of computational pipelines dedicated to the analysis of RiboSeq data have been proposed. These tools are typically designed for extracting gene expression alterations at the translational level, while the positional information describing fluxes and positions of ribosomes along the transcript is still underutilized.

Therefore, the polysome organization, in term of number and position of ribosomes along the transcript and the translational controls directed in shaping cellular phenotypes is still open to breakthrough discoveries.

Broad objectives

The aim of my thesis is the development of mathematical and computational tools integrated with experimental data for a comprehensive understanding of translation regulation and polysome organization rules governing the number of ribosomes per polysome and the ribosome position along transcripts.

Project design and methods

With this purpose, I developed riboWaves, an integrated bioinformatics suite divided in two branches. riboWaves includes in the first branch two modeling modules: riboAbacus, predicting the number of ribosomes per transcript, and riboSim, predicting ribosome localization along mRNAs. In the second branch, riboWaves provides two pipelines, riboWaltz and riboScan, for detailed analyses of ribosome profiling data aimed at providing meaningful and yet unexplored ribosome positional information. The models and the pipelines are implemented in C and R, respectively. riboAbacus and riboWaltz are available on GitHub.

Results

To predict the number of ribosomes per transcript and the position of ribosomes on mRNAs, I applied riboAbacus and riboSim, respectively, to transcriptomes of different organisms (yeast, mouse, human) for understanding the role of translational regulatory elements in tuning polysome in different organisms. First, I trained and validated performances of riboAbacus taking advantage of Atomic Force Microscopy images of polysomes, while performances of riboSim were assessed employing ribosome profiling data. Predictions provided by riboAbacus and riboSim were evaluated in parallel. I showed that the average number of ribosomes translating a molecule of mRNA can be well explained by the deterministic model, riboAbacus, that includes as features the mRNA levels, the mRNA sequences, the codon usage bias and a slowdown mechanism at the beginning of the CDS (ramp hypothesis). The predictions of ribosome localization by riboSim that used as features the mRNA sequence, the codon usage and the ramp, were run for yeast, mouse and human. I observed a good similarity between the predicted and experimental positions of ribosomes along transcripts in yeast, while poor similarity was obtained between predicted and experimental ribosome positions in the two mammals, suggesting the presence of more elaborate controls that tune ribosomes movement in higher eukaryotes than in simple species.

After having developed two tools for the analyses of RiboSeq data and extraction of positional information on ribosome localization along transcripts, I applied both riboWaltz and riboScan in a case study. The aim was to dissect possible defects in ribosome localization in tissues of a mouse model of Spinal Muscular Atrophy (SMA). SMA is a neurodegenerative disorder caused by low levels of the Survival of Motor Neuron protein (SMN) in which translational impairments are recently emerging as possible cause of the disease. I analysed ribosome profiling data obtained from three different types of RiboSeq variants in healthy and SMA-affected mouse brains at the

early-symptomatic stage of the disease. I observed i) a significant drop-off of translating ribosomes along the coding sequence in the SMA condition (using riboWaltz); ii) in SMA-affected mice, the possible accumulation of ribosomes along the 3' UTR in neuro-related mRNAs (using riboScan); iii) the involvement of SMN-specialized ribosomes in playing a very intimate role with the elongation stage of translation of the first codons of transcripts (riboWaltz), iv) the loss of ribosomes at the 3rd codon in SMA in transcripts bound by SMN-specialized ribosomes and v) a remarkable connection between SMN and the down-regulation of genes in SMA-affected mice. Overall, these findings confirmed previous observation about possible SMN-related dysregulations of local protein synthesis in neurons. More importantly, they unravel a completely new role of SMN in tuning translation at multiple levels (initiation, elongation and the recycling of terminating ribosomes), opening new hypotheses and scenarios for explaining the most devastating genetic disease, leading cause worldwide of infant mortality.

Conclusions

The present work provides a new comprehensive and integrated scenario for better understanding translation and demonstrates that this approach is a very powerful strategy to pave the way for new understanding of fine alteration in polysome organization and functional control in both physiological and pathological conditions.

1 Introduction

In this chapter I present the biological aspects related to this work. I start introducing the central dogma of biology, focusing on the eukaryotic translation and on the main elements involved in the process and discussing the integration of the translational machinery with a variety of regulatory factors. I then give an overview on the experimental assays employed to study translation. Finally, I illustrate the relationship between translation and motor neuron diseases discussing the case of Spinal Muscular Atrophy.

1.1 From DNA to proteins

The central dogma of molecular biology describes the processes that guarantee the maintenance of the genetic information throughout cell division and cell life and its flow from DNA to RNA and from RNA to proteins¹. It consists of three fundamental stages: replication, transcription and translation. Replication is in charge of duplicating the DNA, the primary source of the genetic information stored in the nucleus of eukaryotic cells and the starting point for the processes resulting in the production of functional proteins. DNA is used as template during transcription, which produces different types of RNA by copying portions of the genomic sequences, called genes. Some RNAs (small nuclear RNAs, small nucleolar RNAs, micro RNAs and long non coding RNAs) resulting from this process are engaged in post-transcriptional or translational regulation of gene expression²⁻⁴ while other (transfer RNAs, ribosomal RNAs and messenger RNAs) are intimately involved in translation, that is the last step of the central dogma. In the next chapter I outline how translation works in eukaryotic organisms.

1.2 Translation

Translation is the process by which polypeptide chains (i.e. proteins and peptides) are produced using a molecule of mature messenger RNA (mRNA) as a template. A typical mRNA is composed of two untranslated regions (UTRs) placed at the extremity of the mRNA (called 5' and 3' end) and of a central region, the coding sequence (CDS), that contains the information for synthesising the new protein (Figure 1.1A). Groups of three consecutive nucleotides, called codons or triplets, encode specific amino acids,

the monomers constituting the proteins. A mature mRNA also presents a variety of post-transcriptional modifications, such as the m7G cap at the 5' end and the poly-A tail at the 3' end. The poly(A) tail is bound by multiple poly-A binding proteins (PABP) that interact with the 5' cap through protein-protein interactions⁵⁻⁸. The resulting closed loop plays an important role in defining the mRNA fate, stabilizing the mRNA and increasing its translation efficiency⁹⁻¹¹.

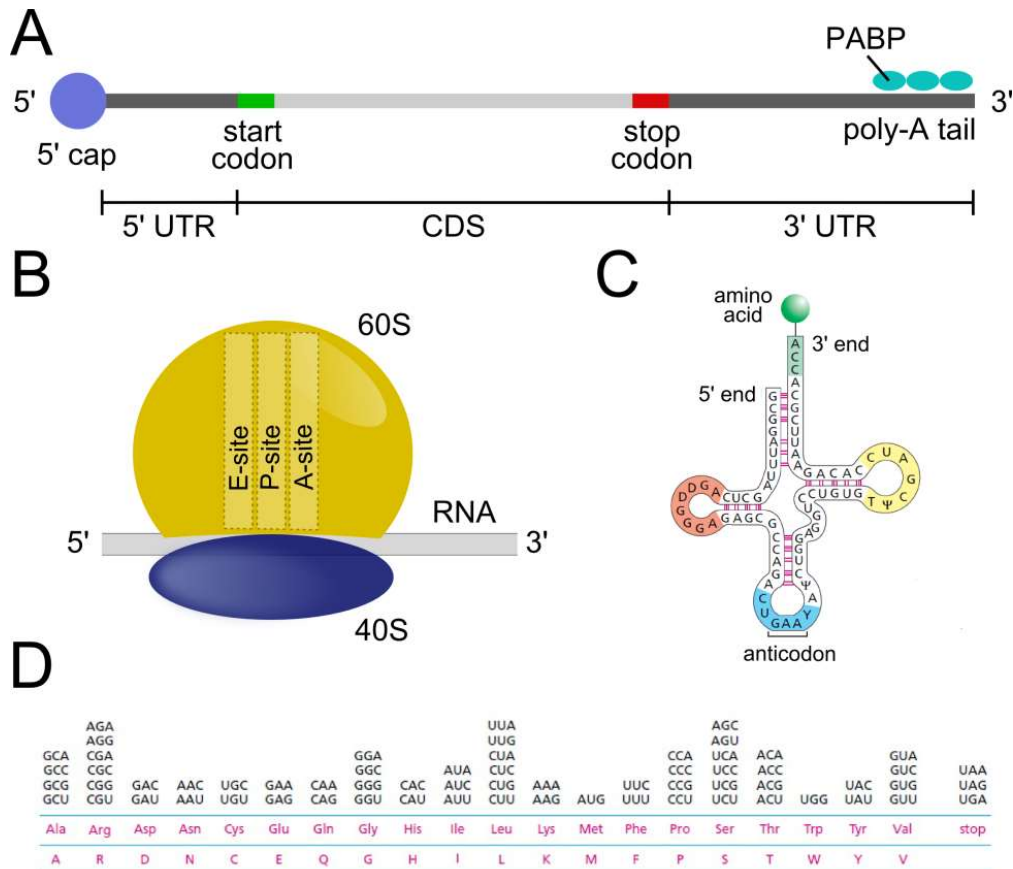


Figure 1.1. **Elements of translation.** Schematic representation of (A) a filament of mature mRNA, (B) a ribosome and (C) a tRNA. (D) Degeneracy of the genetic code in human. In the two lines at the bottom are the proteins (with two different notations) and above the corresponding triplets.

The molecular machineries that carry out the polymerization of peptides are the ribosomes, composed in eukaryotes by four ribosomal RNA (rRNA) and around eighty proteins¹². Eukaryotic ribosomes are formed by the small (40S) and the large (60S) subunit and contain three active sites called E, P and A site (Figure 1.1B). Each site can accommodate, in a different states depending on the stage of translation, a single transfer RNA (tRNA), the molecule that allows the match of the codon with the correct amino acid (Figure 1.1C): for this purpose, the tRNAs include an anticodon, i.e. a

sequence of 3 nucleotides that can base-pair with a triplet, and the associated amino acid.

The relationship between triplets and amino acids is contained in the genetic code¹³ that is highly but not fully conserved among different species^{14,15}. Arranging 4 nucleotides in triplets there are 64 possible combinations, each encoding for one amino acid. However, only 20 different amino acids are used to synthesised protein, meaning that in some case the same amino acid is encoded by multiple codons, namely synonymous codons. In fact, different tRNAs characterized by similar anticodons bring the same amino acid, leading to the so called "codon degeneracy" (Figure 1.1D). Usually, many ribosomes are translating in parallel the same mRNA forming the so-called polyribosome or polysome¹⁶⁻¹⁸.

In eukaryotes translation occurs in the cytoplasm and consists of four main phases: initiation, elongation, termination and recycling. The initiation phase¹⁹ (Figure 1.2) starts with the recruitment on the small subunit of the ribosome of multiple initiation factors (eIF1, eIF1A, eIF3, eIF2–GTP–Met-tRNA^{Met} and probably eIF5, in eukaryotes), what give rise to the pre-initiation complex 43S²⁰. Next step is the interaction of the pre-initiation complex with a mature mRNAs decorated by additional proteinaceous factors (e.g. eIF4F and PABPs), resulting into the initiation complex 48S. During this phase, the 40S scans the mRNA until it reaches the translation initiation site (TIS or start codon). TIS is located at the beginning of the coding sequence (i.e. its 5' extremity). Once the small subunit reaches the TIS, both binding of the ribosomal large subunit (60S) and formation of the ribosome (80S) occur, ending the initiation phase.

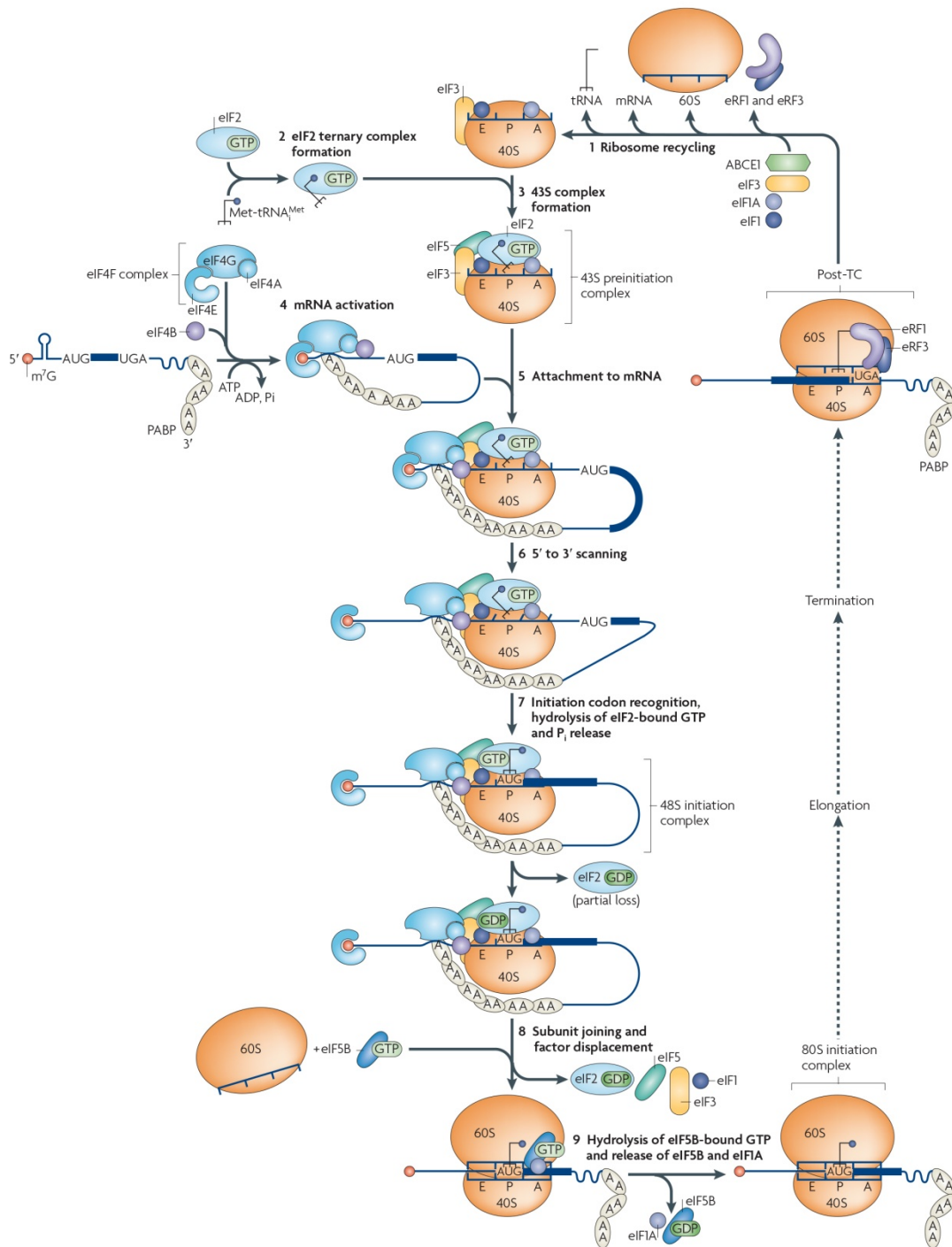


Figure 1.2. **Translation initiation.** The 40S subunit, eIF2–GTP–Met-tRNA ternary complex, eIF1, eIF1A, eIF3, and eIF5 assemble in the 43S pre-initiation complex. The 43S complex binds near to the 5' cap and scans the 5' UTR in a 5' to 3' direction to find the initiation codon. Once found, the 48S initiation complex is formed by switching the scanning complex to a 'closed' conformation. The 60S subunits joins the 48S complexes, the release of eIF5B and eIF1A occurs and 80S ribosomes are assembled. Adapted from Jackson et al²¹.

During the elongation phase²² (Figure 1.3) the ribosome reads the CDS moving towards its 3' end by three nucleotides at a time, adding at each step the correct amino acids to the nascent peptide chain. More in detail, the elongation factor eEF1A-GTP binds the aa-tRNA complex forming the so-called ternary complex. Upon ATP hydrolysis of eEF3-ATP the ternary complex enters the A-site of the ribosome and the hydrolysis of eEF1A-GTP occurs. This reaction leads to conformational changes of the ribosome and the formation of the peptide bond between the new amino acid and the nascent peptide. In particular, during the ribosome translocation step the tRNA placed in the P-site and the A-site move close, leading to rotated state of the ribosomes containing an hybrid state A/P-tRNA²³. eEF1A and eEF3 are then released, while the GTPase eEF2, upon GTP hydrolysis, binds the A-site promoting the advancement of the ribosome towards the 3' end of the mRNA. Finally, eEF2 and the tRNA are released and another cycle can start.

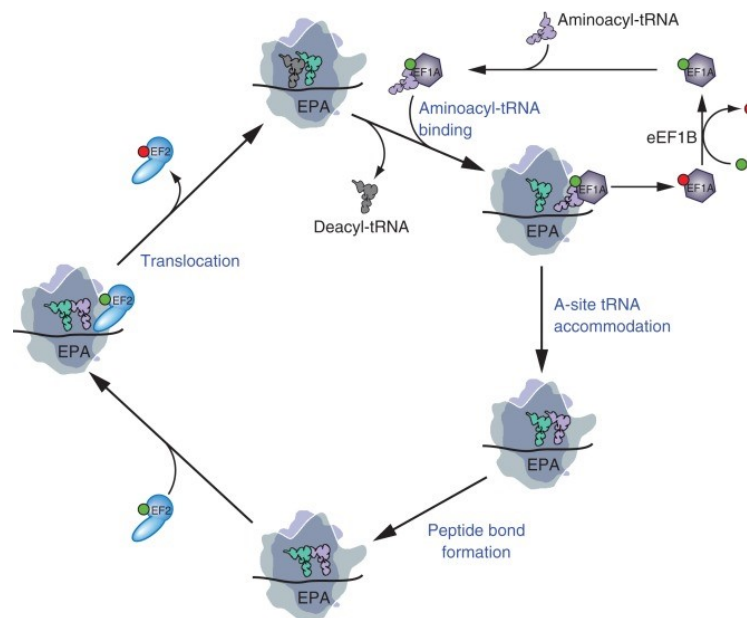


Figure 1.3. **Translation elongation.** eEF1A-GTP-aminoacyl-tRNA ternary complex binds to the ribosomal A site. Following release of eEF1A-GDP, the peptide bond between the amino acid in the A-site and the nascent peptide in the P-site is formed. Binding of eEF2-GTP promotes translocation of the tRNAs into the P and E-sites, and is followed by release of eEF2-GDP. Adapted from Dever and Green²⁴.

When a ribosome reaches the last codon of the CDS (stop codon), the termination phase²⁵ (Figure 1.4) takes place and the termination factor eRF3, together with eRF1 and ATPase ABCE1, binds the ribosome in the A-site, causing the release of the new-formed protein. eRF3 then promotes dissociation of ribosome subunits from the transcript. Note that the latter step may be altered by the binding of IF3 to the pre-

termination complex, which makes ribosomes skip the termination phase causing the readthrough of the stop codon²⁶.

Finally, if the termination phase properly occurs the ribosomes leaving the stop codon may be recycled, i.e. they have a higher probability to start another cycle of translation on the same mRNA rather than turn up in the group of free subunits^{24,27}.

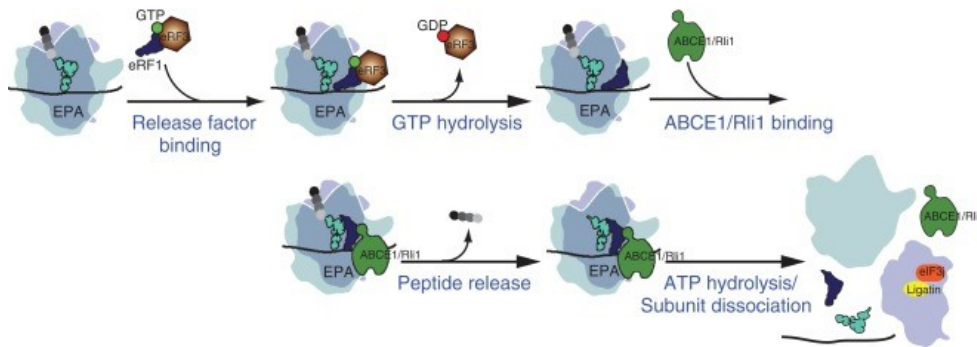


Figure 1.4. Translation termination and recycling. Upon the recognition of a stop codon, the eRF1-eRF3-GTP ternary complex binds to the A site of the ribosome. Following GTP hydrolysis, eRF3 is released. ABCE1/Rli1 binds and facilitates the accommodation of eRF1. ATP hydrolysis finally releases the subunits. Adapted from Dever and Green²⁴.

While the translation phases and the initiation, elongation and termination factors involved in the process have been described for eukaryotes, the abovementioned steps are general and highly conserved in all organisms, from prokaryotes to eukaryotes. In addition, many translation factors have been demonstrated to control translation in a wide range of species, from bacteria to human. For example, hypusination of eIF5A lead to control of both initiation and elongation phases²⁸⁻³⁰, ribosome translocation is tuned through the phosphorylation of the elongation factor eEF2a³¹⁻³³ and the activity of eIF4E and eIF6 may promote the translation of specific mRNAs³⁴. Nevertheless, in the last decades a growing body of evidence showed that organisms control translation using additional and sometimes very sophisticated mechanisms and plenty of molecules to regulate almost each step of protein synthesis, as discussed in the next section.

1.3 Translation regulation

Translation is the most energy consuming process in cells^{35,36} and a primary mechanism for regulating protein expression in a variety of fundamental physiological processes³⁷⁻⁴⁰. Dysregulation of translational control is implicated in a wide collection of pathologies associated to cell proliferation⁴¹⁻⁴⁶. Furthermore, recent findings also

reveals that loss of translational controls connects with the pathogenesis of several neurological diseases, including Alzheimer's disease, fragile X syndrome and spinal muscular atrophy⁴⁷⁻⁵².

It is not surprising that protein levels are largely controlled at the translational rather than transcriptional level⁵³. In fact, latest evidences show a widespread uncoupling between transcript and protein abundances in cells^{54,55}. This uncoupling can be only partially explained by transcription alone (around 40%)⁵⁶. These observations point at post-transcriptional and translational controls as fundamental players in shaping proteomes. Latest findings highlight the complexity of translation and the existence of a wide collection of translational regulatory mechanisms acting both in *cis* (mRNA sequences and secondary structures)^{57,58} and *trans* (ncRNA and RNA binding proteins)^{59,60}. Interestingly these controls can govern translation or even the movement and the position of ribosomes along the mRNAs^{61,62}. This information may lead to new insights into how translational machinery can be controlled through a number of regulatory elements within ribosomes and polysomes.

While ribosomes have been widely studied in various organisms⁶³⁻⁶⁶ providing intriguing evolutionary insights⁶⁷, their organization and coordinated functioning in polysomes is not yet studied in detail. The emerging hypothesis that in cells do indeed exist specialized ribosomes⁶⁸⁻⁷¹ that can be post-translationally modified⁵⁷ open new scenarios about *in situ* translational controls of architectural features of polysomes as unexplored players in tuning translation. This hypothesis is supported by analogous conclusions drawn for transcription, where the structure of the chromatin, a complex DNA/protein/RNA ultra-structural domain within the genome, has been demonstrated to be of crucial importance in tuning gene expression⁷²⁻⁷⁵. This assumption clearly points to a better understanding of the organizing rules of ribosomes along the transcripts, governing parameters such as the number of ribosomes and their positional organization. Acquiring such information would be an important advance to disclose possible causes of the poor correlation between cellular transcriptome and proteome.

Albeit polysomes have been initially described merely as multiple ribosomes moving on a single messenger RNA, recent findings present them as complex integrated platforms where many *cis* and *trans* regulatory elements converge, acting at each phase of translation⁵⁷⁻⁶⁰.

The initiation phase strongly depends on the 5' cap and its interaction with the initiation factors²⁰, the presence of secondary structures of the 5' UTR^{76,77} and the recognition of the correct translation initiation site⁷⁸.

The elongation phase is tuned through many regulatory elements acting both in *cis*^{57,58,79–82} and *trans*^{37,59,60,62,83}. *Cis* regulatory elements mainly consist in the nucleotide composition and GC content⁸⁰ of the coding sequence that may cause mRNA secondary structures^{81,82} and different usage of the codons codifying for the same amino acid (codon usage bias, see next section)⁷⁹. *Trans* regulatory elements include RNA binding proteins that bind the mRNAs^{59,60}, non-coding RNAs that bind ribosomes⁶² and specialised ribosomes^{37,83} that have been demonstrated to determine ribosomal pauses and slowdowns, eventually increasing the complexity of the translational regulatory mechanism.

The canonical termination phase may also be altered leading to either alternative and premature translation termination^{84,85} and ribosome drop-off^{86,87}, or the stop codon readthrough^{88,89}.

Finally, investigations of polysomes by cryo-electron tomography and atomic force microscopy^{90–96} demonstrate the presence of tight ribosome-ribosome interactions and of ribosome cliques (namely clusters) separated by naked portion of mRNAs⁹⁵, revealing highly-organised three dimensional polysomal structures.

1.3.1 Codon usage bias

Among the many translational controls discussed above those acting in *cis* are the most studied. The codon usage bias is a prominent example, since it has been object of several investigations due to its strong association with the nucleotide composition of the mRNAs and consequently its potential role in controlling protein production. Nevertheless, even though its impact on ribosome pauses^{97,98} and drop offs^{99,100} have been largely discussed, its contribution in controlling and translation remains unclear.

The term “codon usage bias” refers to the different frequency (“usage”) of codons that codify for the same amino acid (synonymous codons) in a transcriptome of an organism. Its possible relationship with gene expression and specifically with translation has been proposed in the '80s^{79,101,102}, but only in the last years its connections with the movement of ribosomes along the transcripts has been examined^{103–106}. Recently the “codon optimality”, a scale describing the ratio between the supply of aa-tRNA in the cytoplasmic pool and the frequency of codons along the

mRNAs, has been defined¹⁰⁷. Codon optimality allows to discriminate between optimal codon (decoded with high speed) and non-optimal codons (slowly translated)^{105,108}.

Many hypotheses have been proposed concerning the role played by the codon usage and codon optimality in modulating translation. It has been demonstrated that specific triplets stabilize the mRNAs facilitating ribosome translocation¹⁰⁷. In addition, different usage of codons also implies a different usage of synonymous and non-synonymous aa-tRNAs. The well-known variability of tRNAs concentration in cells correlating with the frequency of the corresponding codon^{109–112} creates a direct link between the nucleotide composition of the mRNAs and the speed of the ribosomes in the elongation phase. In fact, it has been demonstrated that rare tRNAs induce slowdown of ribosomes on non-optimal codon, while frequent tRNAs reduce the time ribosomes spend on optimal codons¹¹³. The presence of rare codons along mRNAs has been shown to increase the efficiency of protein folding^{105,114} and a reduced ribosomal traffic jam along the filaments¹¹⁵, leading to a better control of translation.

Experimental and computational approaches have been employed to investigate the role of codon usage bias in translation. For example, the correlation between the codon usage bias of specifically engineered sequences and the protein abundance obtained after their translation has been computed¹¹⁶, demonstrating that particular sequence of codons may lead to secondary structures along the coding sequence controlling the movement of ribosomes¹¹⁶. Moreover, many bioinformatics analyses examined the relationship between the frequency of synonymous codons along the mRNAs and the correct folding of the nascent peptides during translation¹¹⁷, as well as the connection between the codon usage bias and the experimentally observed enrichments of ribosomes at the beginning of the coding sequences¹¹³. Nevertheless, the results of these studies suffer from different criteria and computational methods used to define the “optimality” of codons, which is not universally established^{118,119}. Therefore the precise role of codon usage in tuning translation, particularly in controlling the number and the localization of ribosomes along the mRNAs is still under debate.

1.3.2 Ramp

One of the most controversial issues related to the codon usage bias is the potential connection with a slow-down mechanism at the very beginning of the coding sequence, known as ramp hypothesis.

The ramp is a region of the coding sequence close to the start codon with an high ribosome density associated to a reduced elongation speed with respect to the

remaining CDS^{120–125}. The ramp has been identified by computational investigation of ribosome profiling data¹²⁶ in many organisms (bacteria^{127,128}, yeast^{115,122,129}, mouse^{81,121} and human^{82,113}) as a sequence ranging from 5 to 50 nucleotides in length, depending on the species^{82,127}. This phenomenon has been extensively studied and a wide range of possibilities about its origin and its effects on translation have emerged over the years. The ramp may be caused by specific nucleotide sequences at the beginning of the coding region and by accumulation of non-optimal codons^{113,122,127} as discussed in the previous section. The ribosome slowdown has been also associated to 2D structures of mRNAs^{116,127}, rapid initiation rates¹²⁹ or to a concurrence of multiple causes^{81,115,128}. This mechanism of ribosome stalling at the beginning of the coding sequence is supposed to reduce the ribosomal traffic jam^{127,128}, minimizing errors in protein synthesis and enhancing protein production¹¹⁵ improving translation. However, as for the codon usage bias, the real contribution of the ramp to the number and the localization of ribosomes along mRNAs remain unclear.

1.4 Experimental techniques for study translation at the genome-wide level

Classical gene expression studies have been based for several years on the assumption that transcriptional levels and corresponding protein level are linearly correlated. In the last years much effort has been directed to investigate the relationship between total mRNA and protein abundances in cells and tissues^{53,56,130–132}, finding that the abovementioned assumption is an oversimplified view and that translation is a major player in shaping the cellular proteome. To further investigate this hypothesis high-throughput techniques such as microarray and Next Generation Sequencing (NGS) have been recently employed for developing genome-wide methods to investigate translation with increasing resolution (Figure 1.5)^{133,134}.

In this section I first present polysome profiling followed by microarray or RNA-Seq analysis^{135–137} and ribosome profiling¹³⁸ as basic technique for portraying translational controls and translation events. Indeed, these are the most commonly used techniques to study translation at the genome-wide level in addition to ribosome immunoprecipitation that exploits the expression in tissues of a ribosomal protein genetically fused to a tag, which allows the identification of ribosome-associated transcripts *in vivo* (e.g. Translating Ribosome Affinity Purification or TRAP¹³⁹). Moreover, polysome and ribosome profiling have been exploited not only for investigating mRNA levels in cell cultures and tissues in different conditions, but also to

calculate the number of ribosomes bound to the same transcript (ribosomes per polysome)¹³⁷ and extract ribosome positional information^{140–142}.

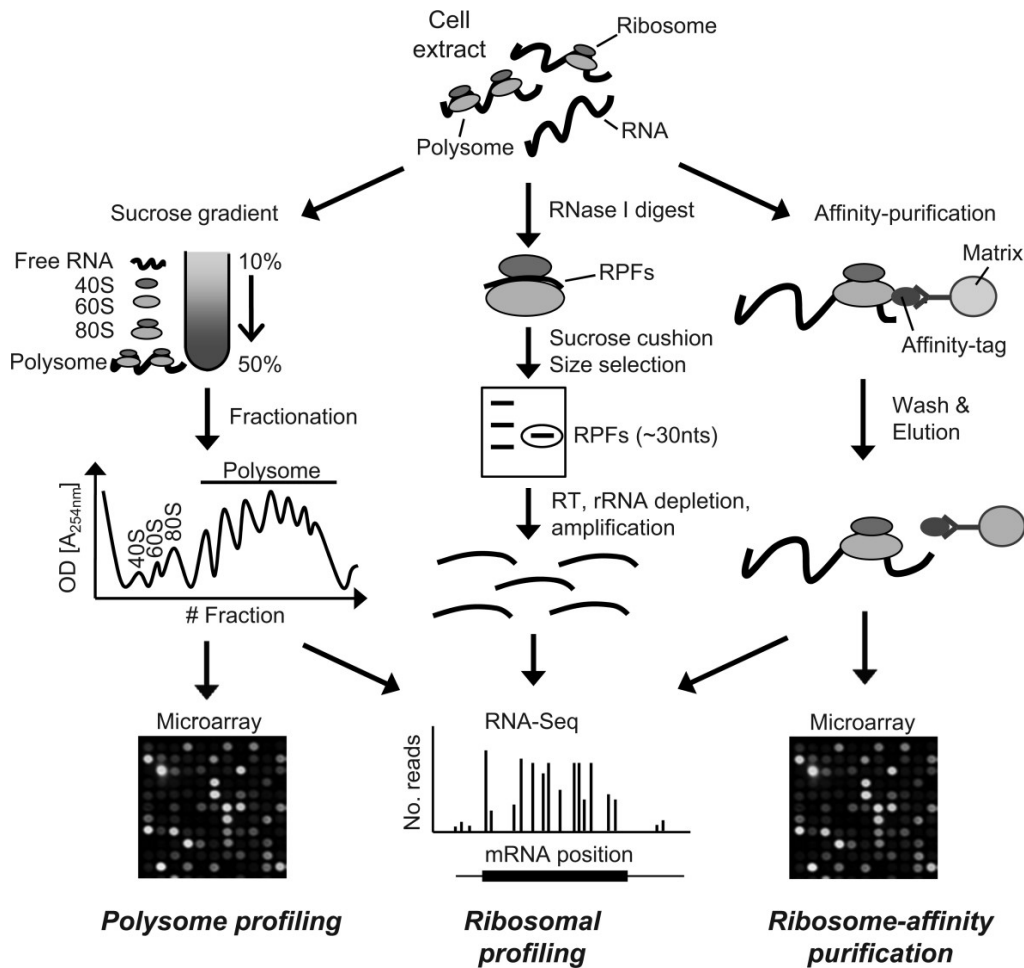


Figure 1.5. **Genome-wide methods to investigate translation.** Polysomal profiling (left): extracts are separated by ultracentrifugation through a linear sucrose density gradient. The gradient is then fractionated allowing the separation of ‘free’ RNA, the small (40S) and large (60S) ribosomal subunits, monosomes (80S) and polysomes. RNA is isolated from individual gradient fractions and pooled for subsequent microarray or RNA-Seq analysis. Ribosomal profiling (centre): extract is treated to digest unprotected and non-ribosome bound regions in the mRNAs. The ribosomes are further enriched through a sucrose cushion, and ribosome protected fragments (RPFs) of RNA are size-fractionated by gel electrophoresis. RPFs of approximately 30 nucleotides are recovered and ligated to sequencing adaptors for reverse transcription, amplification and high-throughput RNA-seq. RAP procedure (right): affinity-tagged ribosomes are captured from extracts with specific antibodies or ligands coupled to a matrix. After several stringent washes, the ribosomes and associated RNAs are released from the matrix and captured RNAs are analysed with DNA microarrays or by RNA-seq. It is possible to combine either polysomal profiling (left), or affinity purification (right), with ribosomal profiling (centre), using either of the aforementioned methods as a vehicle for enriching the sample with ribosomes before isolating ribosome protected fragments. Adapted from King and Gerber¹³⁴.

Second, I briefly examine the use of imaging approaches for obtaining ultrastructural information of ribosome and polysomes at nearly sub-nanometric resolution. In particular, I discuss the use of images of polysomes acquired with an Atomic Force Microscope to compute the number of ribosomes per polysome.

1.4.1 Polysome profiling

Polysome profiling is a classical technique that allows to separate proteins and RNAs (both coding and non-coding) associated to different numbers of ribosomes per polysome, i.e. per transcript. After the removal of mitochondria and nuclei, the cellular lysates are ultracentrifuged on a linear sucrose gradient. Separation of free RNAs (i.e. not-associated to ribosomes), the small (40S) and the large (60S) ribosomal subunits, monosomes (80S) and polysomes can be easily obtained using a fraction collector and employed for indirectly deducing the number of ribosomes per polysome¹³⁷. Typically, polysome profiling is coupled to high-throughput techniques such as RNA-Seq or microarray for the quantification of the transcripts and the determination of the RNA levels at a genome-wide level^{38,143}. In fact, this technique has been used to analyse the variation of mRNAs uploading on polysomes¹⁴⁴, to identify mRNAs controlled at the post-transcriptional levels by RBPs or nc-RNAs¹⁴⁵⁻¹⁴⁸, and study the uncoupling between transcription and translation⁵⁴.

In addition, by measuring the RNA abundance in each sucrose fraction and knowing the sedimentation coefficient of ribosomes, it is also possible to indirectly derive the distribution of the number of ribosomes per transcript¹³⁷.

The estimation of the number of ribosome per transcript provided by this approach is unfortunately indirect and imprecise. For example, polysome fractions collected for microarray analysis may not be clearly resolved and transcripts in specific fractions containing multiple polysomal peaks cannot always be directly assigned to an unambiguous number of ribosomes per polysome. Furthermore, polysomes in higher eukaryotes are known to be associated to a large number of proteins such as enzymes, RNA binding proteins and ncRNAs¹⁴⁹ that can affect the separation and quantification of large polysomes.

1.4.2 Ribosome profiling

Ribosome profiling (RiboSeq) is an experimental technique designed to investigate translation at single nucleotide resolution and genome-wide scale^{123,150}. It is based on the identification of short RNA fragments protected by ribosomes (RPF) from nuclease digestion followed by NGS^{97,151}. Typically, the isolation of RPF is performed starting

from a whole cytoplasmic lysate, that includes both polysomes and 80S monosomes (Figure 1.6), that have been demonstrated to be not-translating^{18,152–155}, even if this is a matter of debate since decades^{156–158}.

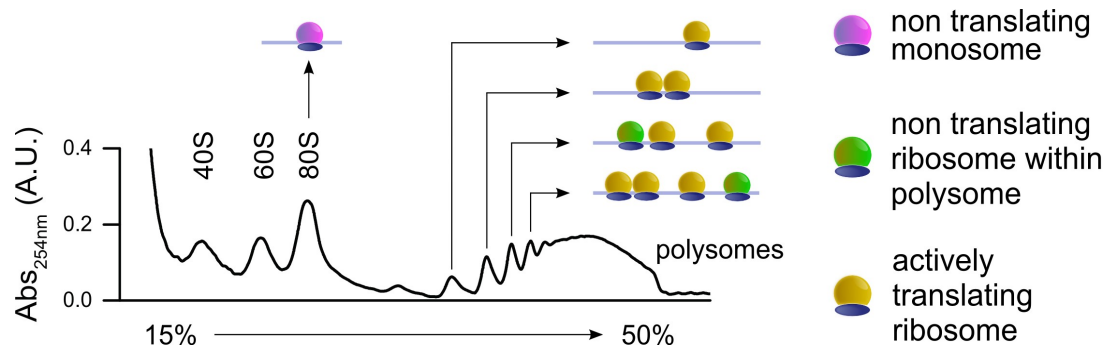


Figure 1.6. **Different classes of ribosomes.** Representative absorbance profile displaying the position of three classes of ribosomes: non-translating ribosomes, non-translating ribosomes associated to polysomes and actively translating ribosomes.

To avoid the isolation of RPF associated to monosomes, ribosome profiling can be applied to already purified polysomes¹⁵⁹. The pre-purification of polysomes removes any possible contamination associated not only to the monosome but also to mRNA fragments associated to the preinitiation complex (48S). Nevertheless, it does not discriminate between ribosomes that are actively translating and ribosomes that might be stalled on transcripts. In fact, it is known that especially in neuronal tissues, physiological paused polysomes do exist^{160,161}. Therefore to overcome possible misleading information about the translational state of transcripts from ribosomes profiling analysis, a third version of ribosome profiling developed by Immagina BioTechnology exploits a new technology called RiboLaceTM to isolate fragments of mRNAs exclusively protected by active ribosomes.

Briefly, Active-RiboSeq uses RiboLaceTM, a new method based on a modified puromycin coupled to magnetic beads. By binding close to the A-site of the ribosome in the not-rotated state, when the acceptor site accommodates the aminoacyl-tRNA engaged by eEF1 α , the puromycin analog can be successfully used to capture active ribosomes.

The pre-print version of the paper, now available in bioRxiv¹⁶², shows that this new tool can be employed to (i) enrich samples with proteins that are either constituent component of the ribosome (RPL26, RPS6) and associated to functional polysomes (eIF4B, PABP, H3K9, eEF1 α), only under conditions of active translation; (ii) capture transcripts undergoing translation in eukaryotic *in vitro* and *in vivo* systems; (iii) describe the variations at the transcript level more precisely than total or polysomal

RNA. For further details about the RiboLace™ technology and the Active-RiboSeq protocol please refer to Clamer et al.¹⁶².

Ribosome profiling coupled with RNA-Seq analysis has been mainly used for computing transcript-specific translation efficiencies (TE)¹²⁶ and performing TE-based differentially analyses¹⁶³. This information is basically identical to what can be obtained by classical polysomal profiling coupled to NGS or microarray^{137,164}. Importantly, the possibility to obtain the precise localization of ribosomes along the mRNAs is unique to Ribo-Seq and is still largely unexploited in the vast majority of published ribosome profiling analysis. Only the last few years have witnessed a rapid adoption of this technique for extracting positional information describing fluxes of ribosomes along the RNA at sub-codon resolution¹⁴⁰⁻¹⁴². This type of analysis is typically based on the so-called ribosome occupancy profiles, i.e. transcript-specific curves showing for each nucleotide along the mRNA sequence the height of the reads signal expressed as the probability to find a ribosome. Analysing ribosome occupancy profiles it is possible for example to reveal novel translated regions¹⁶⁵⁻¹⁶⁷ and ribosome read-through on 3' UTRs^{141,168}. Furthermore, RiboSeq allows to derive translation initiation and elongation rates¹⁶⁹ and estimate codon usage bias identifying translation pauses¹⁷⁰ or ribosomes in specific conformations during the elongation step of translation¹⁷¹.

Analysis of ribosome profiling data

Following the rapid diffusion of ribosome profiling assays, many computational tools and pipelines dedicated to the analysis of RiboSeq data have been developed in the last years. As already mentioned in the previous paragraph, most computational tools are aimed at just computing transcript-specific translation efficiencies for differential analyses in multiple organisms, treatments or conditions (see Babel¹⁶³, Xtail¹⁷², RiboDiff¹⁷³, RUST¹⁷⁴). Typically, these methods do not take into consideration any positional information provided by RiboSeq, since they are based only on the expression levels, i.e. the abundance of mapped RPFs (reads) and on the level of mRNAs obtained by transcriptome analysis that is typically run in parallel to each ribosome profiling.

Occupancy profiles are also the starting point for investigating the presence of alternative translational starting sites i.e. of novel open reading frames (ORF) in known protein coding transcripts or ncRNAs (see AltORFev¹⁶⁶, PROTEOFORMER¹⁷⁵, SPECTre¹⁷⁶, RiboTaper¹⁷⁷ and others^{167,178,179}). They are typically based on statistical methods¹⁷⁷ such as Hidden Markov Model and Bayesian approaches^{167,179} and rely on the identification of the ribosome P-site within the reads^{177,179}. The P-sites position is

employed for the extraction of positional information describing fluxes of ribosomes along the RNA at sub-codon resolution^{140,141} and conformational changes in ribosomes during the elongation step of translation¹⁷¹. Moreover, the identification of ribosome P-site is used for verifying the trinucleotide periodicity of translating ribosomes along coding regions^{123,180}, obtain reliable translation initiation¹⁶⁹ and elongation rates^{169,181-183} and accurately estimate codon usage bias¹⁸⁴. Nevertheless, only very recently and after almost 8 years from the introduction of ribosome profiling two pipelines, specifically dedicated to the identification of the P-site, have been released (Plastid¹⁸⁵ and RiboProfiling¹⁸⁶).

Despite the vast availability of tools for handling ribosome profiling data, carry out a comprehensive analysis of ribosome profiling data is still a complex and time-consuming task. In fact, most of the above-mentioned pipelines perform only one of the many possible RiboSeq data analyses (differential expression analysis, detection of novel ORF, identification of P-site etc.). To overcome this problem, user-friendly genome browsers and on-line environments designed for the storage, the visualization and widespread analyses of ribosome profiling data arose¹⁸⁷⁻¹⁸⁹.

Overall, these applications point to ribosome profiling as a mayor assay for the study of translation from many points of view: the investigation of the role played by controls of translation (e.g. the nucleotide composition of the CDS); the comparison of translational abundances of specific transcript in different conditions; the characterization of polysomes in terms of ribosome number and localization. Nevertheless, both alignment and preprocessing of RiboSeq data may be the cradle of many biases^{190,191} determined, for example, by PCR duplicates¹⁹² ambiguous reads mapped to mRNA isoforms, missing normalizations¹⁹³⁻¹⁹⁵. These biases may lead to particularly noisy occupancy profiles, making the identification of regions associated to ribosome pauses and slowdowns a difficult task. Few works propose original procedures to get rid of RiboSeq bias and improve data analysis^{61,170,196} but a conclusive approach for the extraction of meaningful positional information is still missing.

1.4.3 Atomic force microscopy

A more precise way for obtaining information concerning the number of ribosomes bound to an mRNA is the use of imaging techniques. Given the dimension of ribosomes (around 25 nm in diameter^{64,154}) and of polysomes (ranging between 50-200 nm⁹⁵), imaging approaches that reaches sub-nanometric resolution can be employed. Among these, Electron Microscopy (EM), cryo-EM and Atomic Force Microscopy (AFM) have

been proven to be appropriate for obtaining structural and ultra-structural information on ribosomes in polysomes^{90–96}. Electron microscopes allow to obtain high-resolution information about ribosome–ribosome interactions and the 3D organization of polysomes^{90–94}. Nevertheless, these methods cannot be employed to identify naked filaments of mRNAs, precluding the possibility to precisely count the number of ribosomes per transcript. Moreover, compared to cryo-EM, AFM doesn't need any ex-post image reconstruction procedures thus allowing the acquisition of thousands images of single polysome and no post-processing or reconstruction analysis. From these images the number of ribosomes per polysome can easily obtained with high accuracy^{95,96}.

With respect to estimating the number of ribosomes per polysome using the above-mentioned and indirect fraction by fraction polysome profiling, AFM has some advantages: it can acquire images at single ribosome-resolution, returning a highly resolved distribution of directly counted ribosomes per transcript; it avoids possible biases due to sedimentation characteristics of polysomes (e.g. composition, shape, diffusion coefficients) or dissimilarities in sedimentation in different organism. Therefore, AFM can be of great help for precisely counting the number of ribosomes in thousands of transcripts purified from cells or tissues⁹⁶. The main drawback of the use of AFM images is that it cannot distinguish a transcripts from another, meaning that is possible to obtain the distribution of the number of ribosomes per polysome for a whole transcriptome but it is not possible to obtain transcript-specific information unless using *in vitro* translation systems of single transcripts at a time.

1.5 Mathematical models

The first interactions between mathematics and biological phenomena date back to the XII century, when some probabilistic hypotheses concerning population growth and mortality rates were formulated by Leonardo Fibonacci¹⁹⁷. More recently, similar mathematical approaches were used to study the widespread effects of what is now known as vaccination against specific diseases. For example, in 1760 the Swiss mathematician Daniel Bernoulli investigated the benefits of smallpox exposure against the disease by employing probabilistic procedures¹⁹⁸. From then on, connections between mathematics and biology have been reinforcing, also supported by the advent of the first computers. Nowadays a well-established relationship between mathematical models and the dynamics of biological systems does indeed exist and several applications in population ecology¹⁹⁹, epidemiology²⁰⁰, biophysics (e.g.

crystallography²⁰¹ and electrophysiology²⁰²) and cellular biology²⁰³ and more recently molecular biology²⁰⁴ have been experienced.

Generally speaking, a model is an object that represents and simulates a natural process. A mathematical model is a mathematical theory that well explains natural mechanisms or processes through mathematical entities such as functions, equations, variables, probabilities etc. In this chapter I present the most widespread mathematical models and discuss their applications in the study of translation.

1.5.1 Deterministic models

A deterministic model describes the dynamics of a system through states connected by events, which correspond to the either reversible or irreversible transitions between the states. Each state is associated to specific quantities of all elements involved in the system, represented by variables, and each transition corresponds to an ordinary differential equation (ODE), i.e. a relation between a function f and its derivatives of the form:

$$\frac{d f(x)}{d x} = g(f(x))$$

It is called ordinary differential equation if it contains only one independent variable. Practically, the function f describes a physical quantity and its derivatives represent the rate of its change either in time or space.

Solving these models consists in finding their steady state, starting from a given initial state. This means that the steady state reaches a point such that the variables defining the process do not change neither in time nor in space. In this condition the function(s) describing the system must satisfy the following equality:

$$\frac{d f(x)}{d x} = 0$$

The major advantage in using deterministic models is that, knowing the dynamics of the state transitions in the system, a given initial state always leads to the same solution. In fact, since the system is fully described by selected assumptions and parameters are known with certainty, the predictions are never influenced by uncertain events. Therefore, it is possible to determine the exact state of the system at any time. Nevertheless, deterministic models rely on known parameters and their use is penalised due to the great number of parameters that cannot be experimentally described and that can be represented only as random variables. For example, if the

transition between two states cannot be fully described by one or multiple variables, this event must be either simplified or excluded from the model, making the predictions inaccurate.

Deterministic models of translation

As previously discussed, multiple controls involved in translation may help understanding the reason behind the well-known general uncoupling between transcript and protein abundances in cells^{54,55}. In fact, studying variables in many deterministic models of translation developed since the 1960s²⁰⁵ can increase the prediction of global and transcript-specific protein production rates (hence protein abundances), thus increasing the low correlation observed between transcriptome and proteome in cells.

Typically the deterministic approaches rely on a simplistic representation of translation, modelled as a system composed by a transcript and an infinite pool of aa-tRNAs and ribosomes coupled with ribosome kinetics during the previously described phases of translation: initiation, elongation and termination^{206,207}. In order to find the limiting-step of translation, these models have been also employed to specifically investigate initiation, elongation and termination rates²⁰⁸. To do this, the models complexity has been increased over the years, for example by introducing aminoacyl-tRNA or ribosome competition (i.e. nonspecific binding of aminoacyl-tRNAs in the ribosome A-site²⁰⁹ and a limited supply of free ribosomes in cells²¹⁰, respectively). By introducing these additional assumptions, it was demonstrated that limitation in the number of free ribosomes in cells may prevent ribosomes from stalling along mRNAs during the elongation phase²¹⁰. Additionally, the aminoacyl-tRNA competition was found to be irrelevant for the final protein production rate, even though decrease in ribosome translocation rates may occur under these conditions²⁰⁹. Similar results were obtained by Zhang and Ignatova²¹¹ showing that different levels of aminoacyl-tRNAs and their competition negatively affect ribosome movement along the transcript. Moreover, they showed that the presence of stretches of non-optimal codons within the coding sequences globally influences protein production rates. Nevertheless, opposite findings by Zouridis and Hatzimanikatis²¹² revealed that the codon usage bias tunes ribosome elongation rates and increases the protein production rate to maximum levels. Finally, the crucial role in controlling ribosome translocation by the nucleotide composition of the mRNA has been confirmed by other studies²¹³, where the involvement of the codon usage bias in slowing down the translation initiation phase have been also discussed²¹⁴.

Though many efforts to dissect the mechanism of translation by using deterministic approaches has been put in last years, the existing models point to contradictory conclusions concerning the role of mRNA determinants such as the codon usage and the ramp hypothesis i.e. the slowdown mechanism at the beginning of the coding sequence. Moreover, the connection between these mRNA determinants and polysome features such as the number of ribosome per polysome and the ribosome localization along the mRNA has never been explored in detail. Analogously, even though deterministic models of translation have been extensively exploited for predicting protein production rates and protein abundances, the precise contribution of polysome to the final protein production has never been investigated.

1.5.2 Stochastic models

A stochastic model, as the name suggests, assumes that the evolution of a system relies on single or multiple uncertain events. This means that, simulating a phenomenon described as a set of events connecting different states, the choice of the next reaction to occur is based on a probability distribution. In many cases, the kinetics of the systems moving from one state to another also depends on random variables. Thus, the stochastic nature of this approach gives rise to a variety of paths and, even if a specific state of the system is known, it is impossible to forecast the following ones, and the final state cannot be uniquely determined.

Basically, a stochastic model follows three main steps: i) definition of the initial state; ii) determination of the next reaction and the time it will take; iii) update the system after each reaction occurs. The second and the third steps are reiterated until the system reaches the chosen steady state (if any).

Despite the increased computational complexity of stochastic models compared to the deterministic approaches, the former allow to know exactly which is the state of the system at any time point of the simulation. This means that it is possible to follow the evolution and the variations of all elements involved in the process. Moreover, parameters describing multiple aspects of the modelled system can be easily added and their contribution assessed on the basis of the trend of the simulations.

Stochastic models of translation

Following the same motivations having driven the development of the first deterministic models of translation, many stochastic models have been employed to forecast protein production rates starting from experimentally assessed mRNA levels^{129,181,215,216}. Lots of efforts involving stochastic simulations of translation have

been directed to clarify the contribution of the codon usage bias in controlling ribosome dynamics^{129,215,217–220}.

Similarly to deterministic approaches, stochastic models translation usually describe the initiation, elongation and termination phase by simulating the binding, the movement and the release of ribosomes along a sequence of mRNA^{206,207}. However, these models are based on the totally asymmetric simple exclusion process (TASEP)²²¹, based on the Gillespie algorithm²²² and canonical frequentist probabilities²²³. The TASEP model assumes that a ribosome can move forward one codon at a time in only one direction and only if the next triplet is not occupied by another ribosome. The fundamental steps of these models are the following:

- I. definition of the initial state, typically an empty filament of mRNA and an infinite pool of ribosomes;
- II. generation of a random value to determine the next reaction to occur and the time it will take. There are only three possible types of reactions that can take place:
 - a. the binding of a new ribosome at the start codon;
 - b. the movement of bound ribosomes from one triplet to the next one;
 - c. the detachment of a ribosome from the stop codon;
- III. update the system depending on step II;
- IV. reiterate from step II.

Despite many stochastic simulations of translation based on ribosome kinetics took advantage of the above-mentioned procedure^{215,218,224,225}, more refined models have been developed to investigate the contribution of additional parameters such as the presence of mRNA 2D structures²²⁶, and ribosomes and tRNAs competition^{181,216,217,227} in tuning translation.

For example, Mao and co-workers²²⁶ investigated the effects of 2D structures of the transcripts on translation rates, showing a consistent slowdown of ribosomes in presence of mRNA secondary structures. Regarding ribosomes and aa-tRNAs competition, Chu and collaborators²¹⁷ showed that a limited number of free ribosomes in cells negatively affect the global translation elongation rates more than codon usage bias or possible differences in aminoacyl-tRNA levels can do. In fact, they demonstrated that optimal codons mostly control local and mRNA-specific ribosome translocation rates and that aa-tRNAs competition exert only minor control of translation²¹⁷. The minor role of aa-tRNA abundances in controlling translation was

confirmed by Gritsenko and co-workers¹⁸¹ that on the other hand hypothesized a more complex dynamics of translation regulation by aa-tRNA, based not only on their levels but also on possible aa-tRNA post transcriptional modifications. On the contrary, Gorgoni and collaborators²²⁴ demonstrated that the abundance of aa-tRNA in cells is sufficient to prompt ribosome queues along the transcript.

Recently, the widespread diffusion of ribosome profiling assays gives rise to a collection of stochastic models that take advantage of RiboSeq data, which allows to estimate the position of ribosomes along transcripts expressed as number of reads obtain^{121,129,181,219,228}. These models, by tuning their parameters to obtain the best fit of the experimental data, allow the estimation of translation rates: Ciandrini and collaborators²¹⁹ inferred a set of translation initiation rates for yeast, also revealing that the codon usage bias alone is not sufficient to control ribosome localization along the transcripts; Zupanic and co-workers¹²¹ computed transcript-specific termination rates in mouse, suggesting that translation premature termination is due to either mRNA post transcriptional modifications or stretches of non-optimal codons. Other RiboSeq-based stochastic models led to discordant conclusions about the role of codon usage along in determining ribosome localization and slowdowns along the transcripts. For example, Raveh and collaborators²²⁰, exploiting a model simultaneously simulating translation of multiple mRNAs, described both global and local effects of codon usage bias on ribosome translocation. On the contrary, Shah and co-workers¹²⁹ revealed a connection between high ribosome densities along the coding sequence (especially close to the start codon) and a rapid initiation rate, discarding the hypothesis of slow, non-optimal codons at the beginning of the CDS.

All together, these studies show the great versatility of stochastic simulations of translation, especially using ribosome profiling data. In particular, their ability in dissecting translation by taking into account multiple features (ribosome and aa-tRNA competition, mRNA 2D structures, codon usage bias etc.) emerges. Nevertheless, many contradictory hypotheses about the role exerted by elements affecting translation (e.g. tRNA levels, optimal and non-optimal codons) have been proposed and a consensus is still lacking. Furthermore, the comparisons between predicted and experimental ribosome profiling data may lead to inaccurate results due the existence of global but not local similarities that are typically ignored. Tuning the parameters of the model to reach the best fit of RiboSeq data also presents a second shortcoming: the contribution of the single feature is lost in the simultaneous optimization of all feature values. Lastly, none of these models pays specific attention in extracting polysome organizational rules such as the number or position of ribosomes along transcripts and connections of these features with the considered elements.

1.5.3 Other models of translation

Beyond the deterministic and stochastic procedures described in the previous sections, models based on Bayesian probability²²⁹ and other statistical approaches^{230,231} have been developed for studying translation. For example, Gilchrist and collaborators showed a strong connection between codon usage bias and translation elongation rate in yeast²²⁹, and a significant impact of ribosome recycling and potential non sense-errors have been assessed²³⁰.

Furthermore, studies that mix deterministic and stochastic approaches have been employed for predicting protein production rates²³² and demonstrate the role of aminoacyl-tRNAs abundance and diffusion in cells in limiting translation elongation rates²³³. Finally, recent techniques based on Boolean logic^{234,235} attempted to find the best combination of models for representing translation by exploiting the advantages of multiple mathematical and computational approaches.

1.6 Translation in motor neuron diseases: the case of Spinal Muscular Atrophy

Spinal Muscular Atrophy is the leading cause of infant mortality associated to genetic diseases²³⁶. SMA was described for the first time at the end of the 19th century²³⁷ and is classified as a motor neuron disease, i.e. a progressive neurological disorders characterized by degeneration of motor neurons, the cells that control voluntary muscle activity. In particular, SMA affects lower alpha motor neurons, whose axons arise either from the brainstem or the anterior horn of the spinal cord, directly innervating skeletal muscles²³⁸. Lower alpha motor neurons are responsible for the innervation of the extrafusal muscle fibres at the neuromuscular junction. Consequently, their degeneration results in hypotonia and muscle weakness²³⁹.

SMA has an incidence of around 1 in 6000-10,000 live births and the frequency of the carrier is approximately of 1 in 54²³⁶. Patients are classified into five main classes on the basis of parameters as the age of onset, the patient phenotypes and their motor functions: from type 0 (affected by the most severe form²³⁶) to type IV (with milder symptoms, reach adulthood having all major motor functions²⁴⁰). Type I is the most common form, accounting for about 50% of the SMA cases. SMA of type I has an onset before 6 months of age and death occurs within two years, often determined loss of respiratory functions^{241,242}.

SMA is caused by the loss or mutation of human Survival Motor Neuron gene (*smn*)²⁴³. In the human genome, *smn* is present in two different copies: *Smn1* and *Smn2*, derived from an inverted duplication of *Smn1*²⁴⁰. A single nucleotide mutation differentiates the two copies of *smn*, leading to the skipping of the exon 7 during splicing, which in turn gives rise to a truncated protein that is rapidly degraded. A small percentage of *smn1* is properly translated into functional SMN proteins allowing a correct development of the organism at the embryonic stage.

Even though the genetic cause of SMA has been well-established, the molecular mechanisms that links SMN depletion to the pathogenesis of SMA is still unclear. Recent findings connect SMN to the translational machinery showing a mislocalization of its components in SMN-depleted cells²⁴⁴ and its association to polysomes *in vivo*²⁴⁵ and *in vitro*²⁴⁶. Translation locally occurring in the axons has been also demonstrated²⁴⁷ pointing to local protein synthesis as a fundamental process that allows highly specialized cells such as neurons to regulate their structure and function²⁴⁸. Moreover, SMN have been shown to be associated to SMA Type I leads to a reduction in the number of ribosomes in polysomes in tissues of central nervous system of a mouse model at early and late symptomatic stage and that this defect correlates with SMA disease progression²⁴⁹. Nevertheless, links between SMN and polysome organisation such as the exact ribosome localization along the mRNAs are still open to breakthrough discoveries.

2 Mathematical models of translation

2.1 riboAbacus

Translation is the most energy consuming process in cells^{35,36} and a primary mechanism for regulating protein expression in a variety of fundamental physiological processes³⁷⁻⁴⁰. Translation occurs in polysomes¹⁶⁻¹⁸, highly-structured complexes where several controls converges: recent findings reveal the existence of a wide collection of translational regulatory mechanisms acting in *cis* (mRNA sequences and secondary structures)^{57,58} and *trans* (ncRNA and RNA binding proteins) of mRNAs. As a consequence, these controls may account for the discussed widespread uncoupling between transcript and protein abundances in cells^{54,55}, that can be only partially explained by transcription alone (around 40%)⁵⁶. In fact, multiple regulatory elements can in some cases govern the movement, position and, as consequence, the number of ribosomes within polysomes^{61,164} and thus the whole translation process. In particular, the number of ribosomes bound to the transcripts coupled with mRNA levels is likely to affect the abundance of proteins in cells.

Unfortunately, up to now no experimental techniques allow to calculate the number of ribosomes per transcript with single-transcript resolution at genome-wide level. Thus, the only way to clarify the contribute of the number of ribosomes per polysome in shaping proteomes is a dedicated mathematical model of translation that i) takes into consideration the mRNA levels and ii) estimates the number of ribosomes per polysome.

Many deterministic models of translation have been already developed^{205,212,214,250-252}, aimed at predicting protein abundances in cells and increasing the low correlation observed between transcriptome and proteome in cells. Nevertheless, none of them pay specific attention in extracting polysome organizational rules (such as number and position of the ribosomes) and the precise contribution of the number of ribosomes per transcript to the final protein production is still unexplored. Moreover, despite the many efforts of these modelling studies, a consensus model remains elusive, drawing to contradictory conclusions concerning the role of mRNA determinants. In particular, the contribution of the codon usage¹⁰³⁻¹⁰⁶ and the ramp hypothesis¹²³ (a region at the

beginning of the coding sequence characterized by high ribosome density associated to ribosome slowdown^{120–122,124,125}) in tuning translation is still unclear.

Here, I propose riboAbacus, a mathematical model to predict the number of ribosomes per transcript exploiting for the first time imaging data of polysomes acquired by atomic force microscopy (AFM) and obtained by Dr. Gabriella Viero (Laboratory of Translational Architectomics, IBF-CNR, Trento) in collaboration with Dr. Lorenzo Lunelli (Bruno Kessler Foundation, Trento). In fact, recent studies showed that AFM can be of the greatest help for precisely and counting the number of ribosomes at the genome-wide scale in polysomes purified from cells or tissues^{91,93,95}. I used as input the kinetic constants of elongation and as mRNA determinants the gene expression level, the codon usage bias and the ramp. The model was trained using the experimental distribution of the number of ribosomes per transcript from images of polysomes of the human Hek-293 (GSM936076) cell line, to optimize two ramp parameters: the ramp length and ribosome slowdown rate. These two parameters has been set to best fit the experimental data. I then performed two rounds of validations using: i) the whole transcriptome of human MCF-7 (GSE48213) cell line and ii) the globin mRNA (globin transcript). In both cases I found a good fit between experimental and predicted data. Finally, the predicted number of ribosomes per transcript was used to calculate protein levels of mRNAs expressed in three datasets (the human medulloblastoma, primary mouse motoneurons and NIH3T3 mouse fibroblasts), significantly increasing the correlation between transcript and protein abundances. This result demonstrates the usefulness of the prediction of the number of ribosomes per transcript to reduce the distance between transcriptome and proteome in any biological sample.

The work presented in this chapter has already been published in: “RiboAbacus: a model trained on polyribosome images predicts ribosome density and translational efficiency from mammalian transcriptomes”, Lauria et al., *Nucleic acids research*, 43(22), e153 and riboAbacus is available at <http://fabiolauria.github.io/RiboAbacus/>. A copy of the paper is reported below. My contribution in the following work consisted in i) developing the mathematical model; ii) performing the training and the validation of riboAbacus including the features and the statistical analyses; iii) computing the translational efficiency and comparing it with proteome data; iv) writing the manuscript, the C script and its documentation.

RiboAbacus: a model trained on polyribosome images predicts ribosome density and translational efficiency from mammalian transcriptomes

Fabio Lauria¹, Toma Tebaldi², Lorenzo Lunelli³, Paolo Struffi², Pamela Gatto², Andrea Pugliese⁴, Maurizio Brigotti⁵, Lorenzo Montanaro⁵, Yari Ciribilli⁶, Alberto Inga⁶, Alessandro Quattrone², Guido Sanguinetti⁷ and Gabriella Viero^{1,*}

¹Institute of Biophysics, CNR Unit at Trento, Via alla Cascata, 56/C-38123 Povo (TN), Italy, ²Laboratory of Translational Genomics, Centre for Integrative Biology, Via delle Regole, 101-38123 Mattarello (TN), Italy, ³Laboratory of Biomolecular Sequence and Structure Analysis for Health, Fondazione Bruno Kessler, Via Sommarive, 18-38123 Povo (TN), Italy, ⁴Mathematics Department, University of Trento, Via Sommarive, 14-38123 Povo (TN), Italy, ⁵Department of Experimental, Diagnostic and Specialty Medicine, University of Bologna, Via S. Giacomo, 14-40126 Bologna, Italy, ⁶Laboratory of Transcriptional Networks, Centre for Integrative Biology, Via delle Regole, 101-38123 Mattarello (TN), Italy and ⁷School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, Midlothian EH8 9AB, UK

Received May 29, 2015; Accepted July 20, 2015

ABSTRACT

Fluctuations in mRNA levels only partially contribute to determine variations in mRNA availability for translation, producing the well-known poor correlation between transcriptome and proteome data. Recent advances in microscopy now enable researchers to obtain high resolution images of ribosomes on transcripts, providing precious snapshots of translation *in vivo*. Here we propose RiboAbacus, a mathematical model that for the first time incorporates imaging data in a predictive model of transcript-specific ribosome densities and translational efficiencies. RiboAbacus uses a mechanistic model of ribosome dynamics, enabling the quantification of the relative importance of different features (such as codon usage and the 5' ramp effect) in determining the accuracy of predictions. The model has been optimized in the human Hek-293 cell line to fit thousands of images of human polysomes obtained by atomic force microscopy, from which we could get a reference distribution of the number of ribosomes per mRNA with unmatched resolution. After validation, we applied RiboAbacus to three case studies of known transcriptome-proteome datasets for estimating the translational efficiencies, resulting in an increased correlation with corresponding proteomes. RiboAbacus is an intuitive tool that allows an im-

mediate estimation of crucial translation properties for entire transcriptomes, based on easily obtainable transcript expression levels.

INTRODUCTION

Translation, the synthesis of proteins by ribosomes using an mRNA template, is a fundamental process in biology. It relies upon complex interactions between molecular actors that modulate this process at a number of translation check-points: initiation (1–3), elongation (4–6), termination and ribosome recycling (7,8). Moreover, mRNA determinants such as codon usage bias (9), GC content (10), 5' mRNA structures (11,12), cis regulatory elements (13), protein–protein interaction (14,15), ribosome pausing (16–18), alternative termination (19) and drop off (20,21) influence translational efficiencies or translation rates *in vivo*. In cells, several ribosomes translate the same mRNA forming the so-called polyribosome or polysome (22–24). At steady state, the total number of ribosomes per transcript are the result of an equilibrium among initiation, elongation and termination events. The precise contribution of the number of ribosomes per transcript to the final protein production remains elusive and unexplored because of the challenge posed by obtaining experimental genome-wide distributions of ribosome number per transcript.

Translation has been the subject of intense modelling efforts in the last five decades, using various mathematical and computational approaches (18,25–32). These models aimed at predicting protein production rates and understanding the role of mRNA features or contributions

*To whom correspondence should be addressed. Tel: +39 0461 314033; Fax: +39 0461 314875; Email: gabriella.viero@cnr.it

of translation stages. Several models purely deal with biophysical theoretical descriptions of ribosome fluxes along mRNAs (29,31,33), while recent experimental methods to study translation using ribosome footprinting (17,34) or polysome profiling (32) motivated new mathematical modelling approaches based on genome-wide maps of ribosome occupancy and/or ribosome density along transcripts (18,31,35). Despite the many insights afforded by these modelling studies, a consensus model remains elusive, as different modelling approaches/assumptions often lead to contradictory conclusions concerning the role of mRNA determinants (in particular the contribution of codon usage), the interplay between initiation and elongation, translational rates and efficiencies. Employing ribosome profiling data to develop mathematical models is undoubtedly promising, but several problems have been encountered. For example, biases determined by alignment of ambiguous RNA reads to mRNA isoforms, artefacts caused by missing normalization (36), fragment bias that depends on the length of the sequenced fragments (37–39) can introduce errors that may affect the robustness of translation efficiencies (TEs) calculated using these data. Ribosome profiling has been extensively used for obtaining estimates of ribosome occupancy per transcript. These estimates are essential for parameterizing mechanistic models of translation, however their reliability is questionable, as they are computed by collapsing ribosome positional information from thousands of copies of the very same transcript. Another technique for obtaining ribosome occupancy, ribosome density and the number of ribosomes per transcript could be the employment of polysomal profiling followed by microarray or RNA-seq (40–43). Unfortunately, this approach provides an indirect estimation of the number of ribosomes per transcript. A more precise way for obtaining this information is the employment of imaging techniques, followed by ribosome counting (44). In principle this approach allows to determine the exact number of ribosomes with a single transcript resolution, if a polysome can be univocally identified.

Recently, much effort has been directed at elucidating by imaging the three-dimensional (3D) structure of polysomes in bacteria (45) and eukaryotes using Cryo-ET and atomic force microscopy (AFM) (44,46–48). The emerging model describes polysomes as groups of tightly interacting ribosomes. In addition, independent groups of ribosomes, or ribo-cliques, spaced by naked mRNA can be observed along the same transcript, as demonstrated by AFM (44). Despite the unique advantages of Cryo-ET for obtaining high-resolution information about ribosome–ribosome interactions (48), it cannot be employed to identify coding mRNA filaments uncovered by ribosomes, precluding the possibility to precisely count the number of ribosomes per transcript. Therefore, AFM is of major help for precisely and univocally counting the number of ribosomes in thousands of transcripts purified from cells or tissues.

Ribosome profiling studies introduced the concept of ‘5' ramp’, identified as a region of about 50 codons (34). This region immediately follows the start codon, where ribosomes display on average an increased density, probably moving with a reduced elongation speed (36) with respect to the remaining coding sequence (CDS). Although definitive molecular evidences and mechanistic explanation are

still missing, a body of clues indicates the existence of the ramp effect (49), that has been identified in bacteria (35,50), yeast (18,34,35) and mammals (31,36,51,52). While existing mathematical models of translation have often included a heuristical ramp effect, to our knowledge the ramp parameters have never been systematically explored or optimized.

Here, for the first time, we exploit the rich data provided by AFM images to calibrate a mechanistic model of translation. We develop RiboAbacus, a new mathematical model of translation calibrated using thousands of single-polysome AFM images. The output of RiboAbacus is the prediction of transcript-specific ribosome numbers and ribosome occupancy from transcriptome data. The model takes into account the main steps of the elongation phase to predict in a transcript specific fashion the number of ribosomes per transcript and derive the corresponding translational efficiency (TE). The proposed method has also been compared with polysome profiling in yeast, showing an increased resolution in determining the number of ribosomes per transcript, and a general agreement for single transcript predictions. We took advantage of the experimental distribution of the number of ribosome per transcript in one human cell line (HeK-293) to tune RiboAbacus parameters (ramp length and slowdown) during the training of the model. A second genome-wide dataset (human MCF-7) and one enriched in a single transcript (rabbit globin from *in vitro* translation system) were used for validation. Finally, the predicted number of ribosomes per transcript was employed to calculate the TE of mRNAs expressed in three additional biological systems: the human medulloblastoma cell line DAOY (53), primary mouse motoneurons from stem cells (54) and NIH3T3 mouse fibroblasts (55), significantly increasing the experimental correlation between transcript and protein abundances. This application illustrates the effectiveness of model-based predictions in estimating proteome abundances from transcriptome data. In synthesis, RiboAbacus is an intuitive tool that allows an almost immediate estimation of crucial translation properties for entire transcriptomes, based on easily obtainable transcript expression levels.

MATERIALS AND METHODS

Chemicals

All solution used for polysome purifications has been prepared in RNase-free water containing 100 µg/ml cycloheximide in order to prevent ribosome subunit disassembly. All reagents, unless otherwise cited, were of molecular biological grade and purchased from Sigma.

Cell culture and human polysomal purification

The baker's yeast *Saccharomyces cerevisiae* wild-type strain BY4741 (MATa, his3D1, leu2D0, met15D0, ura3D0) was obtained from the EUROSCARF repository (EUROpean *Saccharomyces Cerevisiae* ARchive for Functional analysis, Institute for Molecular Biosciences, Johann Wolfgang Goethe-University Frankfurt, Germany, www.euroscarf.de). A single yeast colony was grown overnight to stationary phase in 5 ml of YPDA growth medium (1% Yeast Extract, 2% Peptone, 2% Dextrose and 200 mg/l Adenine) at

30°C. The day after the culture was diluted 1/10 in 20 ml of fresh YPDA and allowed to reach the mid-log growth phase. Translation was blocked by adding 0.01 mg/ml cycloheximide. Yeast cells were then collected by centrifugation and lysed with little modifications to Arava's protocol (40). Briefly, yeast cells were transferred to 2 ml round bottom tubes with 1 ml of freshly prepared lysis buffer (20 mM Tris-HCl, pH 8.0, 140 mM KCl, 1.5 mM MgCl₂, 0.5 mM dithiothreitol (DTT), 0.01 mg/ml of cycloheximide, 1% Sodium DeoxyCholate, 1% Triton X-100, 20U RNase inhibitor) and washed twice. Cells were then lysed using 0.7 ml of lysis buffer with 0.6 vol of pre-chilled acid-washed glass beads (0.45–0.55 mm, Sigma-Aldrich). Complete lysis was performed through six cycles of vortexing (30 s) followed by incubation in ice (1 min). Lysates were harvested by collecting supernatants from two subsequent rounds of cold centrifugation with increasing speed (2600 and 7200 g, respectively). Lysates were then diluted to 0.8 ml with lysis buffer and stored at –80°C. Polysomes were purified as described below for human cellular lysates.

Hek-293 and MCF-7 cells were seeded at a density of 2.5×10^4 cells/cm² and maintained for 3 days in growth medium (Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, 100 units/ml penicillin and 100 mg/ml streptomycin at 37°C, 5% CO₂). At 80% confluence, cells were incubated for 3 min with cycloheximide (100 µg/ml) at 37°C to interfere with the translocation step during protein synthesis, blocking translational elongation and trapping ribosomes on the mRNA. Cells were washed with phosphate buffered saline (PBS + cycloheximide 100 µg/ml) and scraped directly on the plate with 300 µl lysis buffer (10 mM NaCl, 10 mM MgCl₂, 10 mM Tris-HCl, pH 7.5, 1% Triton X-100, 1% sodium deoxycholate, 0.2 U/µl RNase inhibitor (Fermentas), cycloheximide 10 µg/ml and 1 mM DTT). After a nuclei and cellular debris removal by centrifugation (5 min at 12 000 g at 4°C), the supernatant was directly transferred onto a 15–50% linear sucrose gradient containing 30 mM Tris-HCl, pH 7.5, 100 mM NaCl, 10 mM MgCl₂ and centrifuged in a Sorvall ultracentrifuge on a swinging rotor for 100 min at 180 000 g at 4°C. The fractions corresponding to the 80S peak and to the polysomes were collected monitoring the absorbance at 254 nm. Each fraction was aliquoted, flash frozen in liquid N₂ and stored at –80°C before AFM imaging.

Preparation of polysomes from rabbit reticulocytes (RRL)

Briefly, 1 ml of untreated rabbit reticulocytes (RRL) prepared according to Jackson and Hunt (56) was complemented with 20 µM hemin (Fluka), 50 µg/ml creatine phosphokinase, 10 mg/ml creatine phosphate (Fluka), 50 µg/ml of bovine liver tRNAs and 5 mM of D-glucose. Endogenous RNAs were translated in 80 µl reactions containing 40 µl of the complemented, untreated RRL in the presence of 75 mM KCl, 0.5 mM MgCl₂, amino acids (20 µM each), 5 mM DTT and 0.1 U/µl RiboLock RNase (Fermentas) for 10 min at 30°C. Reactions were stopped by cooling the tube on ice for 1 min and adding 320 µl of ice-cold, low salt buffer (15 mM NaCl, 1 mM MgCl₂, 10 mM Tris pH 7.4, 1

mM DTT, 0.12 mg/ml cycloheximide). Polysome purification following the above-mentioned protocol.

qPCR from RRL polysomal fractions

Nine fractions were collected monitoring the absorbance at 254 nm. From 0.5 ml of each fraction, total RNA was isolated after proteinase K treatment, phenol-chloroform extraction and isopropanol precipitation and resuspended in 20 µl of RNase free water. For each fraction, 4 µl of total RNA was reverse-transcribed using the iScript™ cDNA Synthesis Kit (Biorad) in a final volume of 20 µl. One microlitre of cDNA and 400 nM of each primer were used in combination with the KAPA SYBR Green kit (KAPA Biosystems) in a final volume of 10 µl. Forty amplification cycles (95°C for 15 s, 55°C for 20 s, 72°C for 25 s) were run in a CFX-96 C1000 thermal cycler (Biorad) using primers specific to rabbit beta-globin (forward: 5'-TTTGCTAAGCTGAGTGAAGTGC; reverse: 5'-CCAGCCACCACCTTCTGATA), rabbit 15-lipoxygenase (forward: 5'-TTCTGTCCCCCTGACGATCT; reverse: 5'-GATCTCTCGGCACCAGCTCT) and rabbit 18S rRNA (forward: 5'-ACGGCCGGTACAGTGAAACT; reverse: 5'-GACCGGGTTGGTTTTGATCTG). qPCR amplification efficiency was calculated for each gene using a relative standard curve derived from a cDNA of total RNA isolated from RRL. The Ct values were determined by the CFX Manager 2.1 (Biorad) applying multi-variable, non-linear regression model to individual well fluorescence traces. The amount of each target gene was quantified relative to the fraction n° 14 and normalized to the level 18S gene, according to Pfaffl equation (57). qPCR reactions were carried out in triplicates.

Atomic force microscopy imaging

For AFM imaging a 20 µl of Hek-293 or RRL polysomal fraction were adsorbed for 3 min on freshly cleaved mica pretreated with Ni²⁺ for 3 min. The samples were then covered with 100 mM Hepes, pH 7.4, 10 mM NaCl, 10 mM MgCl₂, 100 µg/ml cycloheximide and 3% (w/v) sucrose. After 1 h of incubation at 4°C, the sample was extensively and gently washed with DEPC-water containing 100 µg/ml cycloheximide and dried at 20°C for at least 1 h.

Imaging was performed using a Cypher AFM (Asylum Research, Santa Barbara, CA, USA) in AC mode, using Asylum routines for the IGOR software environment (WaveMetrics, Portland, OR, USA). Scans have been acquired using OMCL-AC240TS tips (Olympus) with nominal spring constant of 2 N/m. The scanning parameters were as follows: typical driving frequency 70 kHz in air, scanning rate 1–2 Hz. AFM images were levelled line by line and rendered using the Gwyddion (gwyddion.net) software package. Images were analysed in ImageJ (58) to count ribosomes in polysomes, manually picking the ribosomal particles and assigning them to their respective polysomes using a custom ImageJ macro. Thousand polysomes were analysed (# objects = 3300 for yeast; # objects = 2251 for Hek-293; # objects = 696 for MCF-7, # objects = 901 for polysomes from RRL) picking more than 20 000 ribosomes.

Model

In order to provide the modelling of translation, the elongation phase was divided in nine different steps, each of them linked to a flux (measured in ribosomes/sec) representing the transition of ribosomes from one stage to the next, similarly to what was done in (29).

The model assigns to each codon of an mRNA nine ordinary differential equations describing the rate of change in the number of ribosomes at position n (referred to the position of the P site of ribosomes), S_n , at the different stages (Equations (1)-(11)); note that all the equations are transcript specific). We then set all fluxes equal to 0 to compute the steady-state values of the variables, which is obtained by solving the resulting algebraic system. This procedure, with the addition of a set of initial conditions i.e. of the hypothesis that at time 0 there are no ribosomes along the transcript, allows to compute the steady-state number of ribosomes bound to an mRNA in this condition. See also 'Results' section for the assumptions of the model and Supplementary File 1 for further information on fluxes, notations and all parameters involved in the model.

$$\frac{dS_n^{(2)}}{dt} = V_I + V_n^{(-2)} - V_n^{(2)} \quad n = 1 \quad (1)$$

$$\frac{dS_n^{(2)}}{dt} = V_{n-1}^{(1)} + V_n^{(-2)} - V_n^{(2)} \quad \forall n = 2 \dots N-1 \quad (2)$$

$$\frac{dS_n^{(3)}}{dt} = V_n^{(2)} + V_n^{(-3)} - V_n^{(-2)} - V_n^{(3)} \quad \forall n = 1 \dots N-1 \quad (3)$$

$$\frac{dS_n^{(4)}}{dt} = V_n^{(3)} - V_n^{(-3)} - V_n^{(4)} \quad \forall n = 1 \dots N-1 \quad (4)$$

$$\frac{dS_n^{(5)}}{dt} = V_n^{(4)} - V_n^{(5)} \quad \forall n = 1 \dots N-1 \quad (5)$$

$$\frac{dS_n^{(6)}}{dt} = V_n^{(5)} - V_n^{(6)} \quad \forall n = 1 \dots N-1 \quad (6)$$

$$\frac{dS_n^{(7)}}{dt} = V_n^{(6)} - V_n^{(7)} \quad \forall n = 1 \dots N-1 \quad (7)$$

$$\frac{dS_n^{(8)}}{dt} = V_n^{(7)} + V_n^{(-8)} - V_n^{(8)} \quad \forall n = 1 \dots N-1 \quad (8)$$

$$\frac{dS_n^{(9)}}{dt} = V_n^{(8)} - V_n^{(-8)} - V_n^{(9)} \quad \forall n = 1 \dots N-1 \quad (9)$$

$$\frac{dS_{n+1}^{(1)}}{dt} = V_n^{(9)} - V_n^{(1)} \quad \forall n = 1 \dots N-2 \quad (10)$$

$$\frac{dS^T}{dt} = V_n^{(9)} - V_T \quad n = N-1 \quad (11)$$

To be able to solve the system described above, two other equations are necessary: (i) the formula representing the

number of codons at position n for transcript r that are not covered by the tail of the preceding ribosomes ($C_{n,r}$)

$$C_{n,r} := \begin{cases} M_r - \sum_{\sigma=1}^9 S_{n+4}^{(\sigma)} & n \in [1, N-5] \\ M_r & n \in [N-4, N-1] \end{cases} \quad (12)$$

where M_r is the total number of transcript of species r , and (ii) the probability for a ribosome at position n to move to the next codon (U_n)

$$U_n := \begin{cases} \frac{C_{n+6} - \sum_{j=n+1}^{n+L} \sum_{\sigma=1}^9 S_j^{(\sigma)}}{M_r - \sum_{j=n+1}^{n+L} \sum_{\sigma=1}^9 S_j^{(\sigma)}} & n \in [1, N-(L+1)] \\ 1 & n \in [N-L, N-1] \end{cases} \quad (13)$$

Equation (12) arises from the assumptions made on the ribosome footprint (see 'Results' section and Figure 2A) and the position of the ribosome site with respect to the covered portion of mRNA. Indeed, to obtain the number of free codons at position n , we have to subtract to the maximum amount of codons in that position of the transcript (coincident with the total number of transcripts of species r i.e. M_r) the number of codons occupied by the tail of a ribosome at position $n+4$. Since ribosomes leave mRNAs when they reach position N , the number of codons not covered by any ribosome tail is equal to M_r for the last 4 codons. As introduced before, Equation (13) is related to the probability of ribosomes at position n to move forward (U_n): since RiboAbacus considers M_r copies of the mRNA species r , the number of ribosomes bound at a specific codon (S_n) ranges from 0 (if all mRNA copies have that position empty) to M_r (if all mRNA copies have that position occupied by ribosomes). (S_n) changes codon to codon and the probability to move forward (U_n) depends on the number of free codons close to the tail of ribosomes. In this way the presence of ribosomes along the transcript influences the translocation probability of ribosomes positioned on upstream codons. Basically, (13) coincides with the probability of having a free codon at position $n+5$ for only the transcripts not presenting ribosomes between the triplets $n+1$ and $n+9$, avoiding in this way overestimations of U_n . Note that even if Equation (12) is needed exclusively to compute V_I (translation initiation) and calculate the probability U_n , both the equations are crucial to allow the maintenance of correct distances between the head of each ribosome and the tail of the next one. More precisely, they are necessary to properly compute the only flux related to ribosomes translocation, i.e. $V_n^{(9)}$, whereas all the other steps of the process are not affected by them.

Being a probability, U_n has to satisfy the following condition:

$$0 \leq U_n \leq 1 \quad \forall n = 1 \dots N-1 \quad (14)$$

To avoid any physical overlap between two consecutive ribosomes, the total number of ribosomes bound to transcripts

r has to satisfy the following condition:

$$0 \leq \sum_{\sigma=2}^9 S_1^{(\sigma)} + \sum_{n=2}^{N-1} \sum_{\sigma=1}^9 S_n^{(\sigma)} \leq \frac{M_r \cdot N}{L} \quad (15)$$

Solving directly such a complex system is potentially problematic due to the heavy computational load. This can be alleviated by observing that the equations for the last codon are considerably simpler, since $V_n^{(8)}$ and $V_n^{(9)}$ are related to the presence of ribosomes on subsequent codons and hence are trivially zero for the final codon. This allows to devise an efficient backward solution, by fixing a range of values for the exit flux and then computing for each of them the number of ribosomes bound to the mRNA. At this point, we choose the maximum exit flux such that conditions (14)-(15) are both satisfied.

Since the precise nature of the ramp is still controversial (see Figure 2C), we model the ramp effect by enforcing a lower speed of ribosomes along the first n codons of the transcripts, where n represents the ramp length. Thus, for this portion of the mRNA, we simply multiply the fluxes $V_n^{(1)} \dots V_n^{(9)}$ by a constant (ranging from 0 to 1) corresponding to the ribosome slow down rate we want to test and then we proceed as described before.

Assignment of images to transcripts and calculation of translation efficiency

The output provided by RiboAbacus contains three values for each transcript: (i) the number of ribosomes per transcript, (ii) ribosome occupancy and (iii) TE. The transcriptome-wide distribution of the number of ribosomes per transcript was compared with the experimental distribution obtained from AFM images, both for the training and the validation of the model. As the identity of the individual transcripts imaged by AFM is not known, we couldn't connect directly specific mRNAs to AFM images. We reasonably assumed AFM images to be representative of the distributions of polysomes and transcripts in cells, meaning that the probability of finding the polysome of a certain transcript in an AFM image is proportional to its abundance, easily measurable by experimental approaches. For this reason, in the distributions of the number of ribosomes per transcript obtained with RiboAbacus, we included transcriptome-wide measurements of mRNA levels, given by FPKM (fragments per kilobases per million mapped reads) measurements retrieved from RNA-seq experiments available in literature. RNA-seq provides an empirical distribution of abundance of individual transcripts in a population of cells; the predicted distribution of ribosome counts per transcript was obtained by weighing the predicted number of ribosome on a specific transcript (obtained from RiboAbacus) by its relative frequency (measured by RNA-seq). This marginal distribution can then be directly compared with the distribution of number of ribosomes per transcript measured by AFM. The distance between the experimental and the predicted distribution was calculated constructing two vectors containing at the n -position the experimental and the predicted number of transcripts with exactly n ribosomes attached to them respectively, exploiting then the Euclidean metric to obtain

the distance of interest. More precisely, we used the following formula:

$$\sqrt{\sum_{i=0}^{50} (x_i - y_i)^2} \quad (16)$$

where x_i and y_i represent the frequency of mRNAs associated with the number of ribosomes per transcript i respectively for the experimental and the predicted distribution. This distance ranges from 0 (if the two distributions are identical) and 1.

In our study the ribosome occupancy for transcript r (RO_r) represents the percentage of nucleotides covered by ribosomes for each mRNA and is computed multiplying the predicted number of ribosomes per transcript for the ribosome footprint (L) and normalizing the result for the length of the transcript (N):

$$RO_r = \frac{\#ribosomes \cdot L}{N} \quad (17)$$

Translation efficiency (TE_r) is obtained by multiplying the ribosome occupancy by the transcript expression levels (M_r):

$$TE_r = RO_r \cdot M_r \quad (18)$$

In the paragraph 'RiboAbacus improves predictions of proteome data from transcriptome' the predicted protein abundances were then obtained fitting the length versus ribosome occupancy plot with a negative exponential and using that curve to compute a set of length-specific correction factors, as suggested in (34). We finally used these values in the TE formula obtaining a corrected translation efficiency (cTE).

Statistical analysis

Cross-validation was performed by splitting the Hek-293 transcriptome dataset in two halves. The first was used as training set to optimize the two ramp parameters, the second was used as test set to evaluate the fit of the model. The procedure was repeated 100 times. In parallel, we also approached cross-validation by splitting in two halves the experimental AFM data and we calculated the distance between the two splitted experimental distributions. Also this procedure was repeated 100 times.

To compare the experimental and predicted distributions of the number of ribosomes per transcript, two analytical approaches are used: the first measures the Euclidean distance between the discrete distributions (see the previous subsection), while the second is based on the Kullback-Leibler divergence. In this latter case, using the chi-square minimization method, we first fitted the two distributions with the number of Gaussian curves corresponding to the best fit (usually two or three). Then we computed the Kullback-Leibler divergence between the related curves from experimental and predicted data. After weighting the divergence value to the area under the Gaussian curves that fit the predicted distribution, we finally summed the obtained values.

The Wilcoxon–Mann–Whitney test was used to compare similarity distributions in the paragraph 'Feature analysis'.

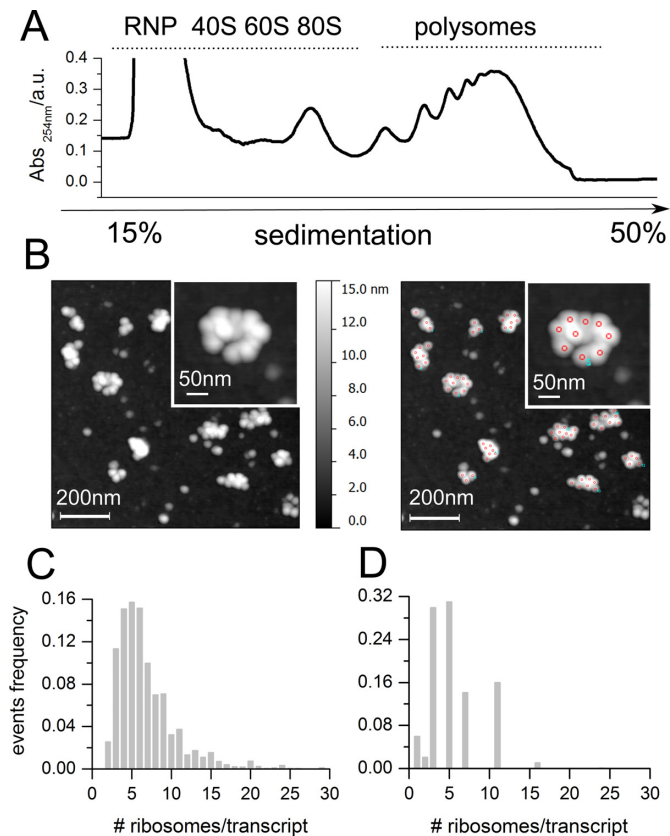


Figure 1. Distribution of the number of ribosomes per transcript by atomic force microscopy (AFM). (A) Representative absorbance profile for sucrose gradient sedimentation of yeast. (B) Example of AFM image of yeast polysomes after absorption on mica (left panel) and example of ribosome detection and counting (right panel, red circles). (C and D) Comparison between the distribution of the number of ribosomes per transcript in yeast obtained using AFM (C) and obtained by Arava and collaborators ((40), D). The number of polysomes considered for counting the number of ribosomes per transcript by AFM is 3300, obtained from 40 independent images.

Williams's test was used to analyse differences between two Pearson coefficients in paragraph 'RiboAbacus improves predictions of proteome data from transcriptome'. The test determines if two dependent correlations are significantly different (59). Williams's test only requires the sample size value and the two correlation values to be compared, and it is the optimal choice for our purposes since it properly works with dependent correlations (60).

RESULTS

Obtaining the distribution of ribosomes per transcripts by atomic force microscopy and comparison with polysome profiling

AFM has been proven to be a powerful approach for studying polysomes and obtaining a great amount of data and information concerning the overall organization of polysomes and the distribution of the number of ribosomes per transcript from thousands of native human polysomes (44). With respect to other methods (40), this technique allows to count the number of ribosomes per transcript at single

ribosome resolution and to obtain genome-wide distributions.

To demonstrate the advantages of AFM method, we compared our approach with polysomal profiling coupled to microarray (40) in yeast. Yeast polysomes were isolated from cellular lysates by sucrose gradient sedimentation (Figure 1A). Then, polysomes were imaged by AFM (Figure 1B, left panel) and the number of ribosomes per polysome (i.e. per transcript) was obtained (Figure 1B, right panel). The transcriptome-wide distribution of the number of ribosomes per transcript was determined and shown in Figure 1C. For the analysis, we took into consideration polysomes with high and medium molecular weights that reflect the steady state distribution of ribosomes per transcript for mRNAs with different lengths (40). Given the fact that it is impossible to obtain pure all-steady state polysomes from a cell lysate, we cannot exclude that in the polysome fraction corresponding to medium molecular weight polysomes, some growing polysomes with long transcripts could be possibly present. Next, we employed the dataset of the number of ribosomes per transcript from (40) and compared this distribution (Figure 1D) with ours (Figure 1C). It is clear that AFM provides a much higher resolution of the distribution of ribosomes per transcript. This is to be expected, as AFM enables a direct measurement at single ribosome resolution of the number of ribosomes per transcript. On the other hand, polysome profiling returns this number indirectly employing the absorbance profiles of a sucrose-gradient separation, followed by logarithmic extrapolation of the number of ribosomes in each fraction, microarray analysis (40) or hybridization/blotting (61) and bootstrapping methods for assigning the number of ribosomes to specific transcripts. Transcripts have to be grouped according to the sucrose fraction that corresponds to a fixed number of ribosomes per transcript. This approach leads to the low resolution of the distribution shown in Figure 1D that would be unsatisfactory as training dataset for RiboAbacus. To date no methods exist to assign with single-ribosome resolution the number of ribosomes per transcript in a genome-wide manner. Thus AFM and polysome profiling appear as complementary techniques because AFM has the advantage of ribosome-resolution and polysome profiling of assigning a specific number of ribosome per transcript. For the purpose of modelling the AFM distribution is the most apt technique, because it is reasonable to assume that a big sample size of single-polysomes AFM images is representative of polysomes and transcripts in cells. This means that the probability of finding the polysome of a certain transcript in an AFM image is proportional to the abundance of its mRNA, i.e its transcript expression level. For this reason, RiboAbacus takes into account the level of the transcript, as measured by RNA-seq.

Assumptions and model development

Before developing the model, we made some preliminary assumptions. RiboAbacus aims at estimating ribosomal densities for an entire transcriptome, without need to know the exact position of ribosomes along the mRNA, an information that could be provided by stochastic (32,62) or probabilistic (27) models. Although some authors considered in

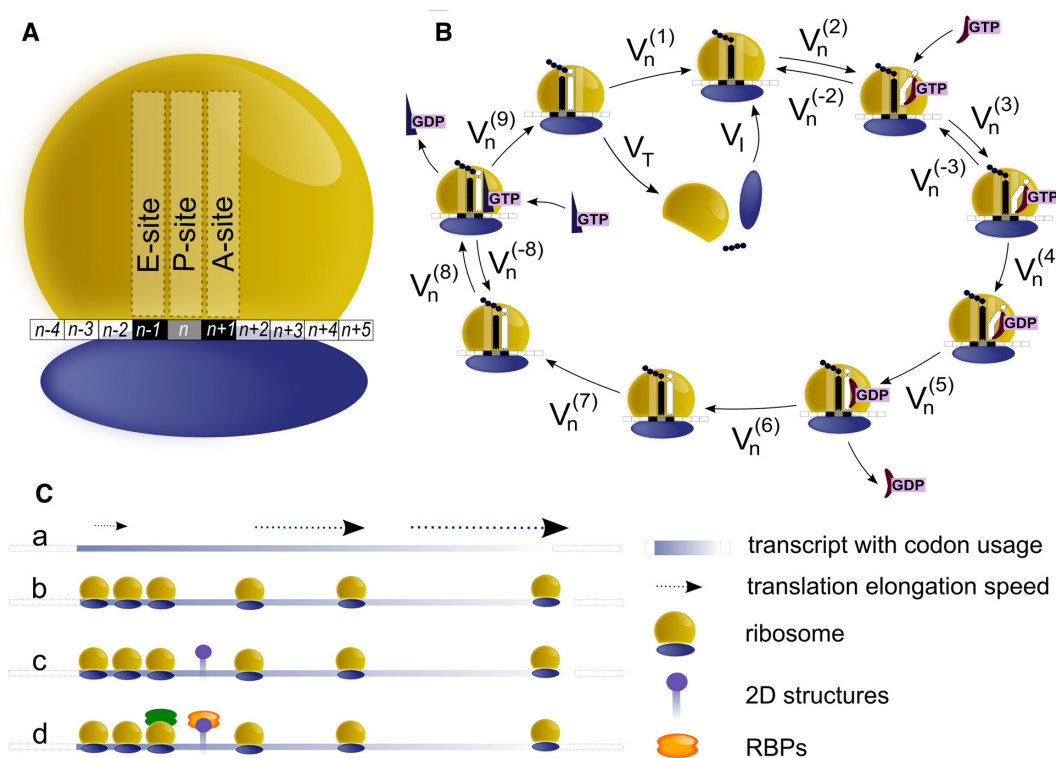


Figure 2. Model description and assumptions. (A) Schematic representation of a ribosome and the portion of the transcript covered. The length of the ribosome footprint is 10 codons. (B) Scheme of the elongation phase, illustrating the chemical reactions considered by the model. Reversible and irreversible reactions during the elongation phase are a simplified version of (64,65) in accordance to (29,66). The kinetic constants for such reactions are taken from (67–70). (C) Schematic representation of the ramp hypotheses: (a) elongation speed is reduced while ribosomes are located on the ramp; (b) the ramp region displays a higher density of ribosomes with respect to the average observed along the remaining transcript; (c) the slowdown effect of the ribosomes along the ramp region could be the consequence of mRNA complex secondary structures; or (d) the presence of RNA binding proteins bound to the region.

their models the availability of ribosomes or their concentrations in the cytoplasm (30,32), our approach does not need to evaluate ribosome competition effects because each transcript is analysed independently. Moreover, we can consider the number of free ribosomes as not limiting. This latter assumption is reasonable given the conditions used for obtaining the training dataset (see next section). In fact under our experimental condition, the number of free ribosomes was calculated to be $2.5 \cdot 10^5$ ribosomes/cell. This value is similar to what observed in rapidly growing cells of *S. cerevisiae* (63) as a not limiting condition.

Moreover, we considered a ribosome coverage of 10 codons (34) (Figure 2A). The choice of this parameter is of utmost importance to avoid collisions between neighbouring ribosomes. To allow the maintenance of correct distances between the head of each ribosome and the tail of the next one, we calculated the probability of any ribosome to move forward and to start a new cycle of translation (i.e the probability for the first 6 codons of the transcript to be uncovered). To do so we defined the codons occupied by E, P and A sites. From now on we will refer to the position of a ribosome as the position of its P site: for example, if a ribosome is at codon position n , this means that its E, P and A sites cover the $n - 1$ th, n th and $n + 1$ th codons, respectively (Figure 2A). The 3 codons upstream the A site and the 4 codons downstream the E site are therefore also covered by the same ribosome given the ribosome coverage of

30 nt. The ribosome coverage length and the position of the ribosome centered in n position (as in Figure 2A) allow to precisely define the overall occupancy of the ribosome and the probability of a ribosome to bind the transcript, start a new cycle of translation and move forward.

The core of the model is based on the elongation phase of translation that was divided into nine steps (Figure 2B) and modelled as nine ordinary differential equations, similarly to what was done in (29). Since the release of the tRNA from the E site is the first reaction that takes place once the ribosome has reached this site and is positioned on a new codon, we considered this reaction as the first step of the elongation phase for each triplet (Figure 2B). In fact, when a ribosome translocates from the codon at position n to the position $n + 1$ it becomes ready to accept a new tRNA but its E site is still occupied by the old tRNA. Therefore the tRNA release is the very first reaction related to the $n + 1$ th codon. The nine steps of the elongation phase can be described as follows: (i) tRNA release from the E site; (ii) binding of the tRNA along with the elongation factor eEF1A and the GTP (the so-called ternary complex) at the A site in a codon-independent process; (iii) binding of the ternary complex at the A site (codon-dependent process); (iv) GTP hydrolysis; (v) eEF1A-GDP position change; (vi) eEF1A-GDP release; (vii) accommodation of the tRNA in the A site and transpeptidation; (viii) eEF2-GTP binding; (ix) ribosome translocation. With respect to the others, ri-

bosomes placed at start and stop codons present some differences, leading to slightly different formulations of the equations for these positions (see ‘Materials and Methods’ section for further details). Ribosomes starting a new cycle of translation do not translocate from previous codons rather entering a new cycle from the tail of the transcript. Ribosomes that reach the end of the transcript and leave the last codon, release the completed polypeptide chain and temporarily detach from the transcript. Since these ribosomes do not translocate to next triplets, this step can be considered the last one of the process and the flux of ribosomes that leave the stop codon was used as starting point to infer the total number of ribosomes per transcript.

As general input parameters, we considered the organism specific codon usage bias values (downloaded from <http://www.kazusa.or.jp/codon>) and the kinetic constants of translation elongation (Supplementary File 1). As transcript specific input we used the following information: (i) the transcript sequence (from ENSEMBL 73) and (ii) the transcript expression level (from RNA-Seq data).

As mentioned in the ‘Introduction’ section, independently of the possible mechanisms giving rise to the ramp (Figure 2C), RiboAbacus includes the ramp effect with two tunable parameters: the ramp length and the ribosome slowdown rate. These parameters were optimized to minimize the distance between the distribution of ribosomes per transcript predicted by the model and the distribution experimentally obtained by AFM (see the following section).

See the ‘Use of RiboAbacus’ section for more details on how to use the software.

Training the model with Hek-293 transcriptome

Given the assumptions of the previous section, we optimized the unknown parameters linked to the ramp (length and slowdown rate), using the experimental distribution of the number of ribosomes per transcript obtained from AFM images of polysomes purified from human Hek-293 cells (Figure 3A and B). The experimental distribution of Hek-293 polysomes was hypothesized to be the sum of normal distributions that we fitted with Gaussian curves (Figure 3C). We obtained as best fit of the experimental dataset three curves with values 5.0 ± 1.3 , 9.0 ± 2.4 and 15.0 ± 5.7 ribosomes per transcript ($R^2 = 0.997$). This experimental distribution was taken as reference for training RiboAbacus.

To predict the distribution of ribosomes per transcript of the Hek-293 transcriptome, we used as input the expression levels of Hek-293 mRNAs determined by RNA-seq (GSM936076) and the corresponding transcript sequences. The transcriptome of Hek-293 consists of 14230 transcripts, whose distribution of the CDS lengths is shown in Figure 3D. To optimize the ramp parameters, we adopted a grid approach and selected 91 different combinations of the two parameters. We run the model for each combination, obtained the corresponding distribution of the number of ribosomes per transcript and compared it with the experimental one. Two examples of these comparisons are displayed in Figure 4A, without ramp (left panel) and with slowdown rate 60% and ramp length of 40 codons (right panel). The distance between experimental and predicted

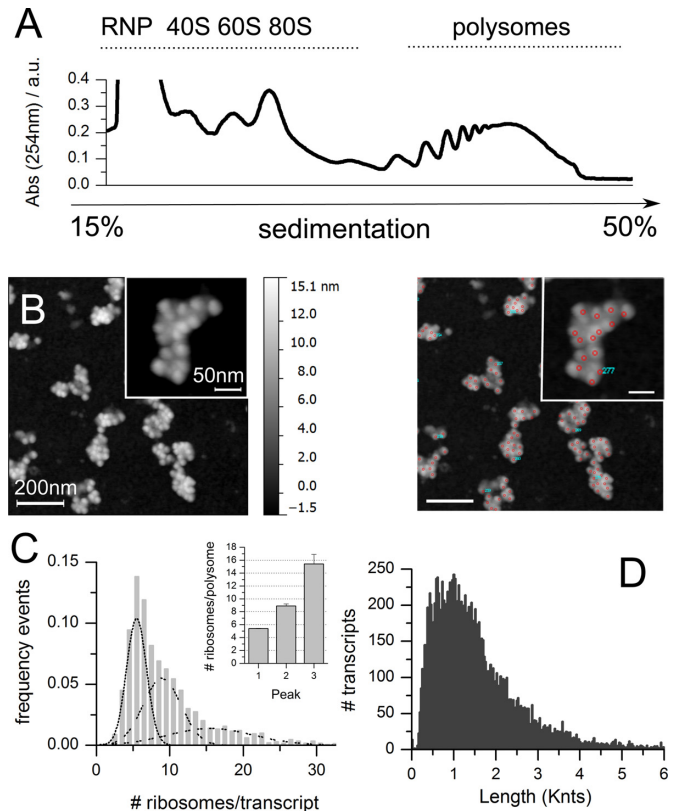


Figure 3. Training dataset: experimental determination of the number of ribosomes per transcript in Hek-293. (A) Example of polysomal profile of Hek-293 lysates. (B) Example of AFM image of Hek-293 polysomes after absorption on mica (left panel) and example of ribosome counting (right panel). (C) Distribution of the number of ribosomes per transcript for Hek-293 transcriptome, as determined from experimental AFM data. The distribution was fitted with three Gaussian curves, with means plotted in the inset ($R^2 = 0.997$). The total number of polysomes considered is 2446, obtained from 20 independent images. (D) Nucleotide length distribution of transcript coding sequences (CDSs) in Hek-293, based on expressed transcripts (GEO ID: GSM936076).

distributions was estimated as described in ‘Materials and Methods’ section. The procedure was repeated 100 times with 50–50 cross-validation (see ‘Materials and Methods’ section and Supplementary Figure S1).

The matrix of average distance values resulting from the combinations of the ramp parameters is displayed in Figure 4B. It is worth noting that without considering the ramp hypothesis in the model (i.e the ramp and the length parameters are equal to 0), the distance value (0.196) between the predicted and experimental distributions is high (Figure 4A, left panel). Similarly, the model with a slowdown rate of 90% displayed the maximum distance values within the matrix with the worst match with ramp length set at 20 codons (distance value 0.334). On the contrary, lower distance values were observed with slowdown rate ranging between 60 and 80% and ramp length between 20 and 80 codons. The best approximation within experimental data (distance = 0.056) was obtained in the case of ramp length equal to 50 codons and slowdown rate of 70% (Figure 4C and D; see also Supplementary File 2 for the complete RiboAbacus results). After fitting this distribution with three

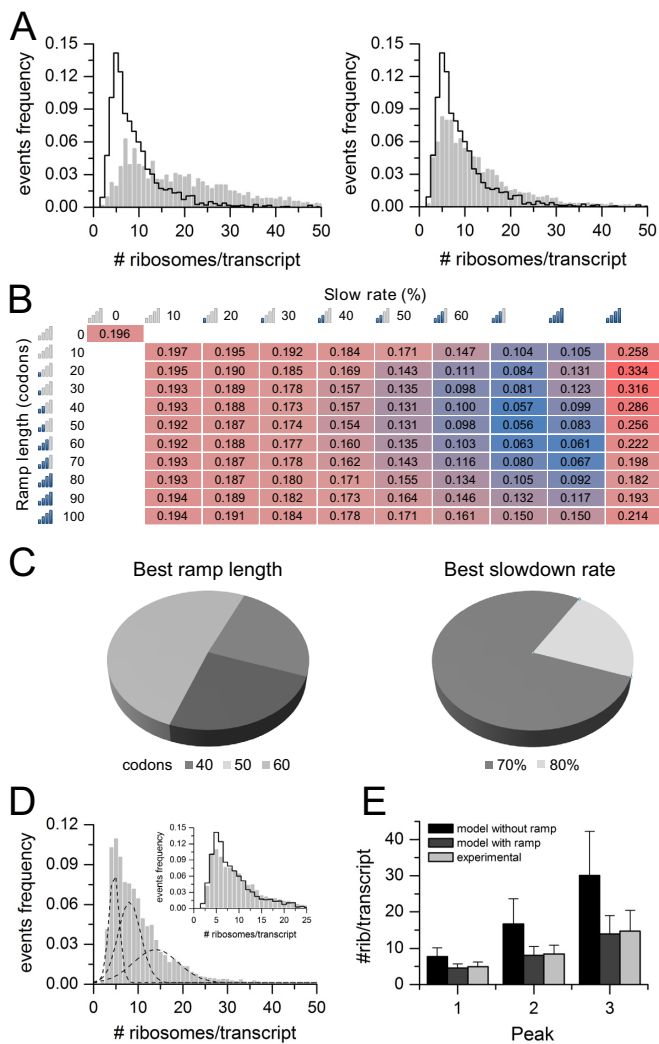


Figure 4. Optimization of model parameters in Hek-293. (A) Comparison between the experimental distribution (black line) and the predicted distribution (grey bars) of the number of ribosomes per transcript, setting the ramp length parameter to 0 (left panel) and to 40 codons with 60% slowdown rate (right panel). (B) Heatmap showing the average distance (100 cross validations) between the experimental and the predicted distribution of the number of ribosomes per transcript, varying the ramp length parameter (from 0 to 100 codons) and the ribosome slowdown rate parameter (from 0 to 90%). Higher distances are highlighted in red gradient, smaller distances in blue gradient. The minimum distance value is obtained with ramp length of 50 codons and ribosome slowdown rate of 70%. (C) Pie charts showing the results of 100 ramp parameters optimizations performed with 50–50 cross validations on the Hek-293 transcriptome. (D) Predicted distribution of the number of ribosomes per transcript, determined by RiboAbacus with optimized ramp parameters (ramp length 50 codons and ribosome slowdown 70%) fitted with three Gaussian curves. The inset shows the comparison with the experimental distribution (black line). (E) Bar plot showing the estimated means of the three Gaussian curves that fit the distribution of the number of ribosomes per transcript, according to experimental data (light grey), predictions from RiboAbacus with ramp length equal to 0 (black), predictions from RiboAbacus with optimized ramp parameters (dark grey).

Gaussian curves similarly to what performed for the experimental data, the predictions nicely matched the experimental values. Computing the distance between the mean of the predicted and experimental Gaussian curves, we obtained differences <1 ribosome per transcript (Figure 4E). On the contrary, comparing the experimental means with those derived from the model with ramp length equal to 0 (black bars in Figure 4E), the differences between the predicted and experimental curves are 3 ribosomes per transcript for the first peak and up to 8–16 ribosomes per transcript for the other two. Interestingly, the absence of the ramp clearly overestimates the number of ribosomes per transcript as displayed in Figure 4E and Supplementary Figure S2. Moreover, the model without the ramp predicts a high coverage ($\sim 70\%$) for both the short and long transcripts, in disagreement with what was observed in yeast in (40), whereas adding the ramp to RiboAbacus the relationship between these two parameters follows an exponential decay trend (Supplementary Figure S3A). It is worth noting that the optimized ramp length value closely matches what experimentally observed by ribosome profiling data (34). Given the results of the training, the two ramp parameters were set to 70% (ramp slowdown) and 50 codons (ramp length) in the following validations.

Feature analysis

To understand the contribution of each transcript feature to RiboAbacus predictions, we started running the model with CDS length as the only feature. Then we progressively added the following features, one at a time: expression level, codon usage bias and optimized ramp parameters. At each step we computed the distance between the resulting predicted distribution of the number of ribosomes per transcript and the experimental distribution. We defined the distribution similarity as 1-distance. In this way the similarity value is 1 if the distributions are identical and close to 0 if huge differences are present. The results of the analysis repeated 100 times with 50–50 cross-validation are displayed in Figure 5, showing a significant improvement in the fit of the model upon addition of each new feature. The inclusion of mRNAs abundances and the codon usage bias leads to an average increment of the similarity of 0.03 and 0.04, respectively. Interestingly, the ramp effect increases the similarity of 0.15, up to 0.95. It is noteworthy that only the combination of all features can properly predict the number of ribosomes. In fact, the removal of one feature leads to lower similarity values (see Supplementary Figure S4). Overall, these results pinpoint the importance of modelling a slowdown mechanisms, such as the ramp and/or initiation rates, and suggest that codon usage bias is not a major determinant of the number of ribosomes per transcript observed in human polysomes, as suggested in (51,52).

Model validation with MCF-7 transcriptome

To validate RiboAbacus, we employed the experimental distribution of the number of ribosomes per transcript obtained from the breast cancer carcinoma cell line MCF-7. The experimental distribution was determined from AFM images by counting the number of ribosomes per polysome

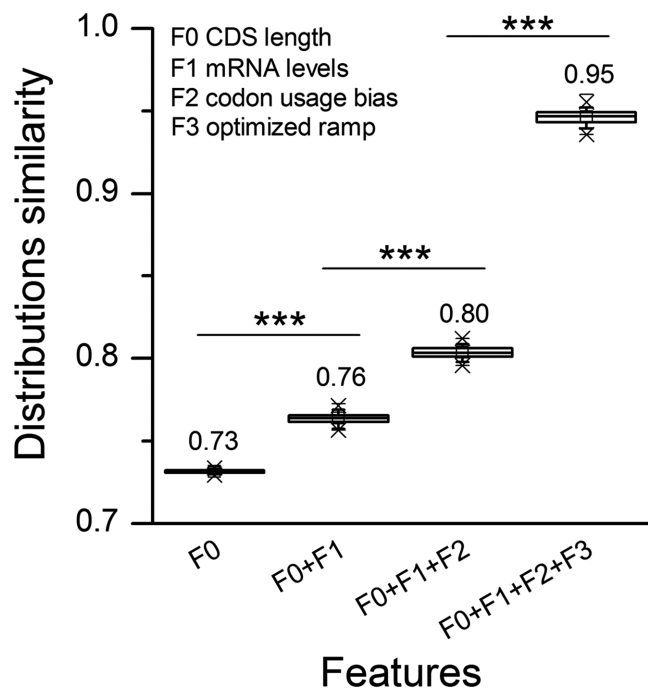


Figure 5. Contribution of transcript features to RiboAbacus predictions. Boxplot showing the similarities (calculated as 1-distance) between the predicted distribution of the number of ribosomes per transcript and the experimental distribution, progressively adding transcript features to the model (F0: CDS length, F1: mRNA level, F2: codon usage bias, F3: optimized ramp parameters). Similarities were calculated in 100 rounds of cross-validation. Statistical significances from Wilcoxon–Mann–Whitney test are shown: (***) P -value < 0.001).

after sucrose gradient sedimentation of cell lysates (Figure 6A) in the same way used for Hek-293.

We then run RiboAbacus, using the ramp parameters previously optimized in Hek-293 and the abundances of MCF-7 transcripts obtained from RNA-seq data (GSE48213, 29 087 transcripts) and the corresponding transcript sequences (Ensembl 73). Similarly to what observed during the training, the predicted distribution of the number of ribosomes per transcript without ramp (grey bars in Figure 6B) poorly matches the experimental distribution (black line Figure 6B). In this case, the distance between the two distributions is 0.234 (Supplementary Figure S5). The introduction of the optimized ramp parameters (length 50 codons and slowdown rate 70%, Figure 6C and Supplementary File 2) leads to a clear improvement of the prediction and a consequent decrease of the distance value to 0.099.

In the case of MCF-7, the experimental distribution was best fitted with two Gaussian curves (Figure 6A), with means of 6.5 ± 2.3 and 12.4 ± 3.9 ribosomes per transcript ($R^2 = 0.996$). Comparing these values with the means of the two Gaussian curves obtained with the optimized parameters, we found a good agreement (Figure 6D). In fact, fitting the data obtained without the ramp, we observed a difference between the experimental and the predicted mean of ~ 3 ribosomes per transcript for the first curve and 8 for the second. Similarly to what observed in Hek-293, the mean values of the optimized model better approximate the experimental means, with differences of around 1 ribosome

per transcript. Noteworthy, the optimal ramp slowdown region is cell line independent and characterized by a length between 30 and 70 codons, with a slowdown rate ranging from 60 to 90% (see Figure 4B and Supplementary Figure S5). Overall, these results confirm the ability of the model to consistently estimate the number of ribosomes per transcript.

Model validation in the rabbit reticulocyte system

The use of AFM allowed us to precisely describe the composition, in term of ribosomes per transcript, of thousands of polysomes, i.e. to precisely count with single transcript resolution how many ribosomes are engaged in translation for transcripts expressed in cells. Even though this prediction originates from the most extensive census of ribosome numbers available in literature, AFM cannot recognize the identity of the corresponding transcripts when using a cell lysate. This means that we cannot associate to one specific transcript a specific number of ribosomes using a cell lysate, nor measure the abundance of each transcript in the images. To overcome this problem, we took advantage of the well-known *in vitro* translation system based on RRL lysates. In this system, unless treated with micrococcal nucleases, two proteins are preferentially produced: globin and lipoxigenase. Indeed, globin represents the great majority of synthesized proteins (71). In addition, given the difference between the length of the two transcripts, it is possible to isolate sucrose fractions that are highly enriched of polysomes formed by the globin transcript. We therefore used this system as additional validation model to count the number of ribosomes per transcript in a population composed of a known transcript. This model has the advantage of offering a single transcript validation of RiboAbacus predictions.

We purified rabbit reticulocyte polysomes by sucrose gradient fractionation (Figure 7A) and purified RNA along the gradient to identify by qPCR the sucrose fraction enriched in globin polysomes (Figure 7B). The polysome fraction with the peak of globin mRNA (arrow in Figure 7A and B) was analysed by AFM imaging (Figure 7C) to determine the experimental distribution of the number of ribosomes per transcript (Figure 7D). The experimental mean number of ribosomes per transcript (4.7 ± 0.89) was compared to the number predicted by RiboAbacus in absence or presence of the ramp parameters optimized in Hek-293. RiboAbacus predicted 9 ribosomes per globin transcript without the ramp assumption, and 4 ribosomes per transcript with the optimized ramp parameters (Figure 7D), a number very close to the mean of the experimental distribution. This transcript-specific validation further demonstrates that RiboAbacus is a powerful model for accurately predicting the number of ribosomes per transcript.

RiboAbacus improves predictions of proteome data from transcriptome data

The great advantages of modelling translation are mainly the possibility to (i) predict protein levels starting from transcript abundances and (ii) obtain information about how mRNA determinants or other parameters can contribute in defining protein production. The quantification of protein

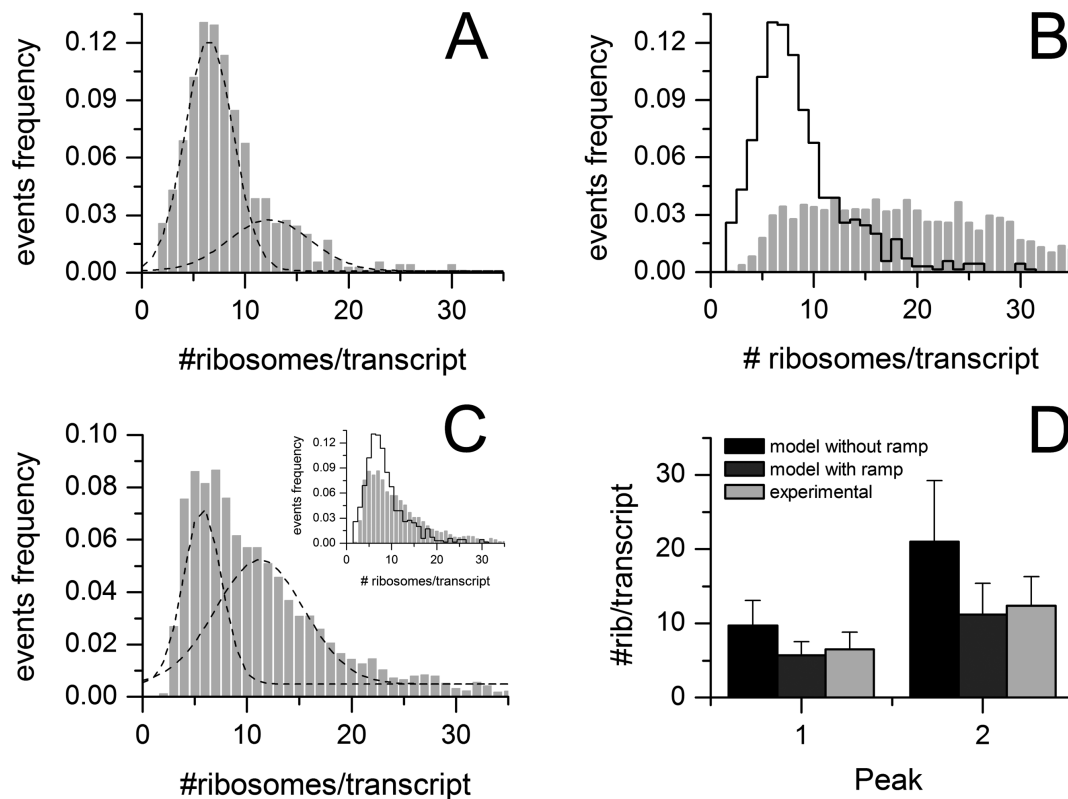


Figure 6. Validation of RiboAbacus in MCF-7. (A) Distribution of the number of ribosomes per transcript for MCF-7 transcriptome, as determined from experimental AFM data. The distribution was best fitted with two Gaussian curves ($R^2 = 0.996$). (B) Comparison between the experimental distribution (black line) and RiboAbacus predicted distribution (grey bars) of the number of ribosomes per transcript, setting the ramp length parameter to 0. (C) Distribution of the number of ribosomes per transcript predicted by RiboAbacus with the previously optimized ramp parameters (length 50 codons and slowdown rate 70%) best fitted with two Gaussian curves. The inset shows the comparison with the experimental distribution (black line). (D) Bar plot showing the estimated means of the two Gaussian curves that fit the distribution of the number of ribosomes per transcript, according to experimental data (light grey), predictions from RiboAbacus with ramp length equal to 0 (black), predictions from RiboAbacus with optimized ramp parameters (dark grey).

levels is sometimes challenging for lowly expressed proteins, difficult samples such as tissues, biopsies, single cells and subcellular compartments such as axons. The experimental detection of transcript levels is far more easy and cost effective, but it has been shown that mRNAs levels poorly correlate with protein levels in several organisms (53,72). We wondered whether this discrepancy between transcriptome and proteome could be reduced by using the number of ribosomes per transcript predicted by RiboAbacus (Supplementary File 2).

To prove this, we selected three studies where protein and transcript abundances have been experimentally determined (53–55) and we checked whether RiboAbacus predictions were able to increase the correlation between experimental transcriptomes and proteomes. For each dataset we computed the predicted number of ribosomes per mRNA and obtained the corresponding ribosome occupancy values using the ramp parameters optimized in Hek-293 (Supplementary File 2). Plotting the ribosome occupancy as a function of the corresponding mRNAs length, we observed that the relationship between these two parameters follows an exponential decay trend (see Supplementary Figure S3A). This means that the shorter the transcript, the higher the ribosome occupancy, supporting previous obser-

vations (34,40). To calculate the TE we introduced a correction parameter that takes into account this effect, as previously suggested in (34). In this way we obtained a cTE. For each dataset we then computed the correlation between the predicted cTEs and the experimental protein levels. To measure whether RiboAbacus significantly improved the correlation between transcriptomes and proteomes, we performed the Williams's tests (Table 1). In parallel this comparison was repeated running RiboAbacus without ramp parameters, to understand the role of slowdown effects also in this context.

The first transcriptome/proteome dataset used was obtained from human medulloblastoma cell line DAOY (53). In this case, the experimentally measured protein quantities better correlate with the predicted cTE ($R = 0.531$) than with transcript levels ($R = 0.425$). The increase of the correlation is statistically significant (P -value $3.5 \cdot 10^{-3}$), suggesting that RiboAbacus better approximates protein production (Figure 8A and B). Moreover, the correlation calculated without the ramp hypothesis ($R = 0.468$, P -value = 0.25) is not significantly higher than the experimental correlation, confirming the important role of slowdown mechanisms in correctly modelling the process of protein production (Figure 8C). Similarly, we applied RiboAbacus to pri-

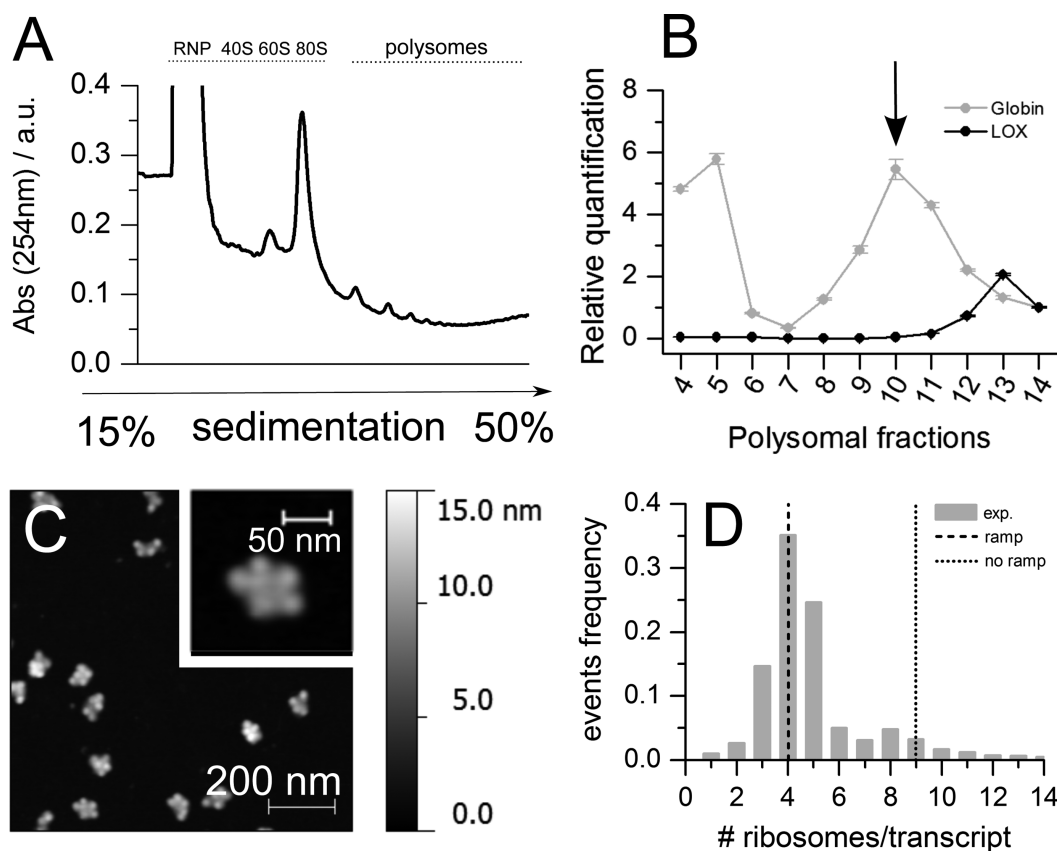


Figure 7. Validation of the model with the globin transcript in rabbit reticulocytes (RRL). (A) Representative absorbance profile for sucrose gradient sedimentation of rabbit reticulocyte lysates after incubation at 37°C for 10 min. (B) PCR quantification of globin and lipoxygenase (LOX) transcripts along the sucrose gradient fractions. The fraction with the highest abundance of the globin transcript is marked with a black arrow. This fraction was chosen for AFM imaging. (C) Example of AFM image of RRL polysomes after absorption on mica. (D) Comparison between the experimentally determined distribution of the number of ribosomes per transcript ($\#$ counted objects = 901; mean $\#$ ribosomes/transcript 4.7 ± 0.8 , in agreement with what observed in (22)), the number predicted with ramp length equal to 0 ($\#$ ribosomes/transcript = 9, dotted line) and with the optimized ramp parameters ($\#$ ribosomes/transcript = 4, dashed line).

Table 1. List of transcriptome/proteome and cTE/proteome correlations with and without the ramp hypothesis for three different transcriptome/proteome datasets

Cell line	Number of transcripts	Transcriptome/proteome correlation	cTE (ramp)/proteome correlation	Correlation increase P -value (ramp)	cTE (no ramp)/proteome correlation	Correlation increase P -value (no ramp)
DAOY	904	0.425	0.531	$3.50 \cdot 10^{-3}$	0.468	0.254
Motoneuron	5600	0.473	0.532	$2.92 \cdot 10^{-5}$	0.480	0.631
NIH3T3	5830	0.615	0.655	$2.95 \cdot 10^{-4}$	0.655	$6.04 \cdot 10^{-4}$

The number of transcripts involved and the P -values from Williams's tests are also reported for each analysis.

mary mouse motoneurons (54). Again, RiboAbacus significantly increased the correlation between transcript and protein levels using the optimized ramp parameters ($R = 0.532$ versus $R = 0.473$, P -value = $2.92 \cdot 10^{-5}$), but not without the ramp hypothesis ($R = 0.480$, P -value = 0.63). Finally, we run RiboAbacus on a third transcriptome-proteome dataset from NIH3T3 mouse fibroblasts (55). Using this dataset the increase in correlation with optimized ramp parameters is smaller than in previous cases, ($R = 0.615$ versus $R = 0.655$, P -value = $2.95 \cdot 10^{-4}$). In contrast to previous examples, the increase in correlation was significant also without the ramp hypothesis ($R = 0.653$, P -value = $6.01 \cdot 10^{-4}$).

This slight increase could be due to the higher initial correlation between the experimental transcriptome and proteome.

Use of RiboAbacus

RiboAbacus is coded in C and available in GitHub at <http://fabiolauria.github.io/RiboAbacus/>. Two input files are needed: a list of transcript CDSs with related expression levels and a list of organism-specific codon usage bias values. The transcript file must contain for each transcript two lines: the first reporting the expression level, along with general information about the transcripts (gene ID, transcript

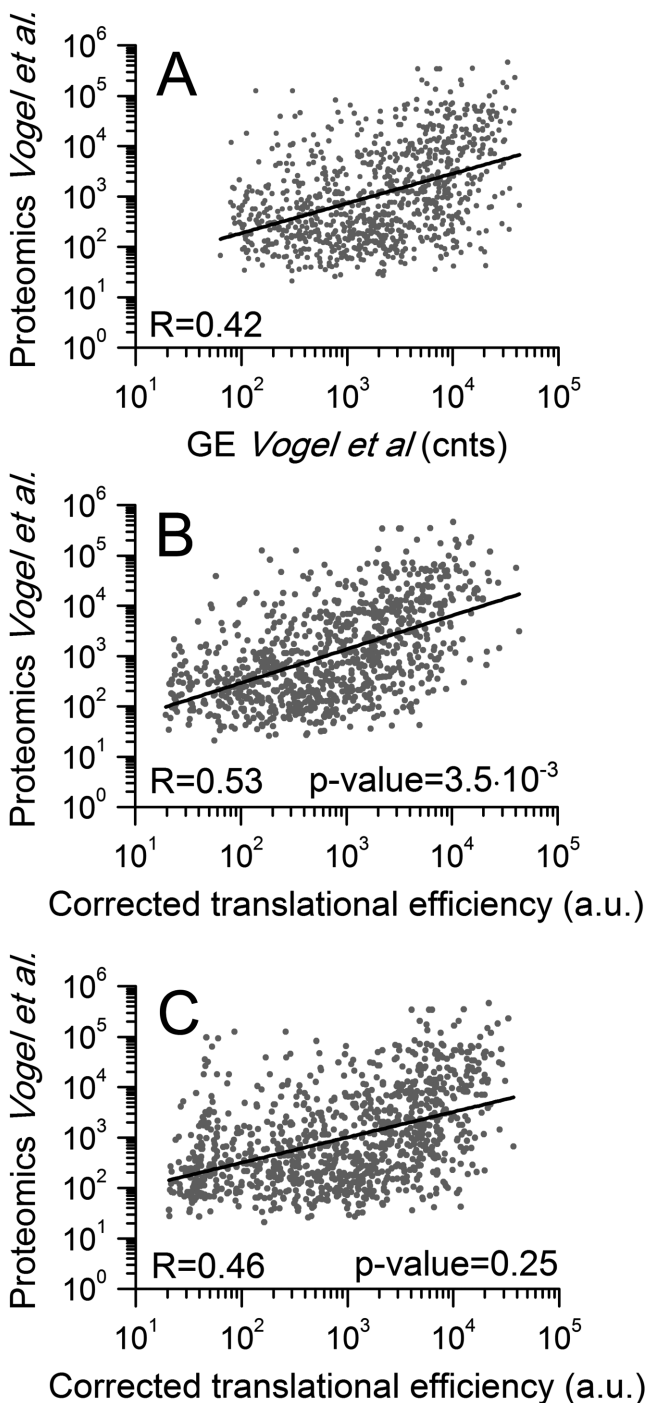


Figure 8. Improved correlation between transcript and protein abundances using translation efficiencies (TEs) calculated by RiboAbacus. (A) Scatterplot of experimental transcript abundances versus protein abundances (53). (B) Scatterplot of cTEs calculated by RiboAbacus with the optimized ramp parameters versus protein abundances. (C) Scatterplot of cTEs calculated by RiboAbacus with ramp equal to 0 versus protein abundances.

ID, protein ID and protein level) and the second reporting the CDS. The codon usage file must contain the list of codons and the corresponding codon usage bias values, arranged in two columns. We provide three options for *Homo*

sapiens (default), *Mus musculus* and *S. cerevisiae*. Note that the set of kinetic constants is fixed (see Supplementary File 1). RiboAbacus outputs two files: the first contains for each transcript the number of ribosomes, the ribosome occupancy and the TE; the second file contains the frequencies of the number of ribosomes per transcript, that can be used to build the transcriptome-wide distribution. For further information please refer to the Readme file in GitHub.

A run of RiboAbacus on an entire transcriptome takes less than a minute on a standard personal computer.

DISCUSSION

We developed RiboAbacus, a model trained on experimental imaging-derived data, able to quickly and accurately predict the steady state number of ribosomes per transcript in entire transcriptomes.

The number of ribosomes bound to a mRNA directly contributes to the final amount of the corresponding protein in cells, since ribosomes are the molecular machines responsible for protein synthesis. Therefore, understanding the contribution of the number of ribosomes bound to a transcript is of major importance for unravelling the impact of translational controls and possibly using transcriptome data to predict TEs. Nevertheless, measuring numbers of ribosomes bound to transcripts is challenging, leading researchers to neglect this important parameter in the development of mathematical models of translation.

The distributions of the number of ribosomes per transcript obtained from AFM images, that underpins RiboAbacus predictions, have been compared with polysome profiling in yeast, using the well known dataset from (40). We were able to show that AFM enables to reach an unparalleled resolution in determining the number of ribosomes per transcript. On the other hand, high-throughput approaches based on hybridization or sequencing allow the identification of transcripts, that is not possible in AFM. Nevertheless, a general agreement for single transcript predictions between arrays and RiboAbacus has been shown (Supplementary Figure S6).

RiboAbacus takes as input a list of transcripts whose sequence and expression levels are known, and the organism codon usage bias and the translational kinetic constants. As experimental reference for tuning the model output, we took advantage of experimental data obtained from AFM images of purified polysomes that uniquely allows the precise count of ribosomes per transcript. Without additional parameters, RiboAbacus predictions overestimate the number of ribosomes per transcript (Figure 4A, left panel). Such overestimation has already been observed in other models, indicating that codon usage alone is not sufficient to account for ribosome dynamics (29,32).

We thus took into consideration the existence of 5' slowdown mechanisms that may give rise to the so-called ramp described in yeast by ribosome protecting assays (34). The possible biological reasons for the existence of the ramp are still under debate and the conclusions discordant (18,31,34,36,49–52). A hypothesis is that regions rich of rare codons could affect the waiting time for the correct tRNA binding to ribosomes (51,52). In addition, the presence of RNA structures, produced by intramolecular base pairing,

could also induce a slowdown movement of mRNA helicases (35) and a consequent stalling of ribosomes. Most probably, these two features contribute simultaneously to final ramp effects (31,50). Regardless of the specific mechanism involved, we decided to model the ramp effect by introducing in RiboAbacus two ramp parameters: ramp length and ramp slowdown rate. We optimized these parameters in Hek-293, computing the best fit with the experimental data. Interestingly, the optimal value of the ramp length (50 codons; Figure 3B) is in agreement with data available in literature (34,35,51,52). Importantly, our results highlight that codon usage bias plays a minor role than the ramp hypothesis in the accuracy of prediction (Figure 5). Therefore, our predictions indicate that the ramp, or any slowdown events taking place at the beginning of the CDS, plays an important role in determining the overall number of ribosomes per transcript.

Another confirmation of the importance of slowdown mechanisms can be observed inspecting the ribosome occupancy or coverage (i.e. the percentage of nucleotides covered by ribosomes). Using the optimized ramp parameters we could observe that the ribosome occupancy per mRNA was inversely proportional to the length of the CDS, similarly to what was experimentally observed in other studies (34,40). It is then possible that additional translation mechanisms, such as different initiation rates or ribosomes drop off (20,21), can play a role to avoid the loading of high number of ribosomes on long transcripts keeping the mRNA coverage at a low level. Indeed, we found that ribosome occupancy is almost constant for transcripts longer than 2000 nt even if the total number of ribosomes per mRNA increases with their length.

Using RiboAbacus, we tried to understand the contribution that the number of predicted ribosomes per transcript may give to explain the total protein level in cells. In fact, mRNA abundances, measured by microarray or by next-generation sequencing (NGS) techniques, are widely used as proxies for protein measurements, but a general poor correlation between the experimental measures of mRNA and protein levels has been reported in many works in mammalian cells. For example, (53) showed that the mRNA abundance in cells may account for approximately one-third of the downstream protein production yield ($R^2 = 0.29$). Computational approaches have been attempted in order to identify and select mRNA features that could bridge the gap between transcriptome and proteome measurements by employing multivariate linear regression models. In *S. cerevisiae*, a set of transcript-specific features (including codon usage, transcript length, ribosome density, evolutionary conservation) was selected to maximally increase the prediction of protein levels from mRNA levels (from 0.69 to 0.76 in (73), from 0.76 to 0.86 in (74)). In mammalian systems, Vogel and co-workers (53) identified 25 mRNA features that increased the coefficient of determination from 0.29 to 0.67 on a subset of 512 transcripts. Thus, we asked what could be the overall contribution of the number of ribosomes uploaded on transcripts in determining the proteome. RiboAbacus is able to estimate the cTE for each transcript given its abundance. We found that the number of ribosomes per transcript significantly increased the experimental correlation in three different datasets: from 0.42

to 0.53 in human medulloblastoma cell line DAOY, from 0.47 to 0.53 in mouse motor neurons and from 0.61 to 0.66 in mouse fibroblasts NIH3T3. Interestingly, without the introduction of the ramp parameters, the increase in correlation is considerably lower and, with the exception of the last dataset, not significant. This result is an additional clue that the slowdown of ribosomes at the 5' of the CDS should play a pivotal role in regulating the final protein abundance. According to the improved correlation provided by RiboAbacus, up to 10% of protein levels can be explained by the number of ribosomes per transcript.

While the improvement in explanatory power afforded by RiboAbacus is both sizeable and statistically significant, there remains a considerable amount of proteomic variability unaccounted for, pointing to the need of additional translational regulation mechanisms. Future work is needed to better understand how additional translational controls could be included in mathematical models to improve the correlation between transcript and protein levels.

RiboAbacus stands as a simple and immediate approach that may be useful to deal with problems that we have with other methods for studying translation. In fact it can predict numbers of bound ribosomes and transcript-specific translation properties solely from global gene expression. As such, RiboAbacus can be applied to any gene-expression dataset, requiring much fewer experimental resources than polysome profiling methods and representing a quick complementary method to more expensive and demanding experimental techniques to study translational control of gene expression. It can also be used to predict protein levels and translational properties in systems (e.g. biopsies, single cells, subcellular compartment etc.) where a proteomic quantification is still challenging.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

AXonomIX research project financed by the Provincia Autonoma di Trento, Italy; European Research Council [MLCS 306999 to G.S.]. Funding for open access charge: International AXonomIX Research Project financed by the Provincia Autonoma di Trento, Italy.

Conflict of interest statement. None declared.

REFERENCES

1. Sonenberg, N. and Hinnebusch, A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**, 731–745.
2. Aitken, C.E. and Lorsch, J.R. (2012) A mechanistic overview of translation initiation in eukaryotes. *Nat. Struct. Mol. Biol.*, **19**, 568–576.
3. Jackson, R.J., Hellen, C.U. and Pestova, T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.*, **324**, 113–127.
4. Ingolia, N.T., Lareau, L.F. and Weissman, J.S. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, **147**, 789–802.
5. Stadler, M. and Fire, A. (2011) Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA*, **17**, 2063–2073.

6. Li, G.-W., Oh, E. and Weissman, J.S. (2012) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, **484**, 538–541.
7. Nürenberg, E. and Tampé, R. (2013) Tying up loose ends: ribosome recycling in eukaryotes and archaea. *Trends Biochem. Sci.*, **38**, 64–74.
8. des Georges, A., Hashem, Y., Unbehauen, A., Grassucci, R.A., Taylor, D., Hellen, C.U., Pestova, T.V. and Frank, J. (2014) Structure of the mammalian ribosomal pre-termination complex associated with eRF1-eRF3-GDPNP. *Nucleic Acids Res.*, **42**, 3409–3418.
9. Novoa, E.M. and Ribas de Pouplana, L. (2012) Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet.*, **28**, 574–581.
10. Lynn, D.J., Singer, G.A. and Hickey, D.A. (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.*, **30**, 4272–4277.
11. Gray, N.K. and Hentze, M.W. (1994) Regulation of protein synthesis by mRNA structure. *Mol. Biol. Rep.*, **19**, 195–200.
12. Kudla, G., Murray, A.W., Tollervey, D. and Plotkin, J.B. (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*, **324**, 255–258.
13. Pircher, A., Bakowska-Zywicka, K., Schneider, L., Zywicki, M. and Polacek, N. (2014) An mRNA-derived noncoding RNA targets and regulates the ribosome. *Mol. Cell*, **54**, 147–155.
14. Shalgi, R., Hurt, J.A., Krykbaeva, I., Taipale, M., Lindquist, S. and Burge, C.B. (2013) Widespread regulation of translation by elongation pausing in heat shock. *Mol. Cell*, **49**, 439–452.
15. Friend, K., Campbell, Z.T., Cooke, A., Kroll-Conner, P., Wickens, M.P. and Kimble, J. (2012) A conserved PUF-Ago-eEF1A complex attenuates translation elongation. *Nat. Struct. Mol. Biol.*, **19**, 176–183.
16. Wolin, S.L. and Walter, P. (1988) Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J.*, **7**, 3559–3569.
17. Ingolia, N.T., Brar, G.A., Rouskin, S., McGeachy, A.M. and Weissman, J.S. (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.*, **7**, 1534–1550.
18. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. and Plotkin, J.B. (2013) Rate-limiting steps in yeast protein translation. *Cell*, **153**, 1589–1601.
19. Noller, H.F. (2012) Evolution of protein synthesis from an RNA World. *Cold Spring Harb. Perspect. Biol.*, **4**, a003681.
20. Herr, A.J., Wills, N.M., Nelson, C.C., Gesteland, R.F. and Atkins, J.F. (2001) Drop-off during ribosome hopping. *J. Mol. Biol.*, **311**, 445–452.
21. Cruz-Vera, L.R., Magos-Castro, M.A., Zamora-Romo, E. and Guarneros, G. (2004) Ribosome stalling and peptidyl-tRNA drop-off during translational delay at AGA codons. *Nucleic Acids Res.*, **32**, 4462–4468.
22. Warner, J.R., Rich, A. and Hall, C.E. (1962) Electron microscope studies of ribosomal clusters synthesizing hemoglobin. *Science*, **138**, 1399–1403.
23. Palade, G.E. (1955) A small particulate component of the cytoplasm. *J. Biophys. Biochem. Cytol.*, **1**, 59–68.
24. Wettstein, F.O., Staehelin, T. and Noll, H. (1963) Ribosomal aggregate engaged in protein synthesis: characterization of the ergosome. *Nature*, **197**, 430–435.
25. Gerst, I. and Levine, S.N. (1965) Kinetics of protein synthesis by polyribosomes. *J. Theor. Biol.*, **9**, 16–36.
26. MacDonald, C.T. and Gibbs, J.H. (1969) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Biopolymers*, **7**, 707–725.
27. Gilchrist, M.A. and Wagner, A. (2006) A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *J. Theor. Biol.*, **239**, 417–434.
28. Mitarai, N., Sneppen, K. and Pedersen, S. (2008) Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. *J. Theor. Biol.*, **382**, 236–245.
29. Zouridis, H. and Hatzimanikatis, V. (2007) A model for protein translation: polysome self-organization leads to maximum protein synthesis rates. *Biophys. J.*, **92**, 717–730.
30. Reuveni, S., Meilijson, I., Kupiec, M., Rupp, E. and Tuller, T. (2011) Genome-scale analysis of translation elongation with a ribosome flow model. *PLoS Comput. Biol.*, **7**, e1002127.
31. Dana, A. and Tuller, T. (2012) Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS Comput. Biol.*, **8**, e1002755.
32. Ciandrini, L., Stansfield, I. and Romano, M.C. (2013) Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS Comput. Biol.*, **9**, e1002866.
33. Von der Haar, T. (2012) Mathematical and Computational Modelling of Ribosomal Movement and Protein Synthesis: an overview. *Comput. Struct. Biotechnol. J.*, **1**, e20120400.
34. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. and Weissman, J.S. (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, **324**, 218–223.
35. Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Rupp, E. and Ziv-Ukelson, M. (2011) Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.*, **12**, R110.
36. Zupanec, A., Meplan, C., Grellescheid, S.N., Mathers, J.C., Kirkwood, T.B., Hesketh, J.E. and Shanley, D.P. (2014) Detecting translational regulation by change point analysis of ribosome profiling data sets. *RNA*, **20**, 1507–1518.
37. Bohnert, R. and Ratsch, G. (2010) rQuant. web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res.*, **38**, W348–W351.
38. Hansen, K.D., Brenner, S.E. and Dudoit, S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.*, **38**, e131.
39. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. and Pachter, L. (2010) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.*, **12**, R22.
40. Arava, Y., Wang, Y., Storey, J.D., Liu, C.L., Brown, P.O. and Herschlag, D. (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 3889–3894.
41. MacKay, V.L., Li, X., Flory, M.R., Turcott, E., Law, G.L., Serikawa, K.A., Xu, X.L., Hookeun, L., Goodlett, D.R., Aebersold, R. et al. (2004) Gene expression analyzed by high-resolution state array analysis and quantitative proteomics response of yeast to mating pheromone. *Mol. Cell. Proteomics*, **3**, 478–489.
42. Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K. Y.S., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W. et al. (2011) FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, **146**, 247–261.
43. Qin, X., Ahn, S., Speed, T.P. and Rubin, G.M. (2007) Global analyses of mRNA translational control during early *Drosophila* embryogenesis. *Genome Biol.*, **8**, R63.
44. Viero, G., Lunelli, L., Passerini, A., Bianchini, P., Gilbert, R.J., Bernabò, P., Tebaldi, T., Diaspro, A., Pederzoli, C. and Quattrone, A. (2015) Three distinct ribosome assemblies modulated by translation are the building blocks of polysomes. *J. Cell. Biol.*, **208**, 581–596.
45. Brandt, F., Etchells, S.A., Ortiz, J.O., Elcock, A.H., Hartl, F.U. and Baumeister, W. (2009) The native 3D organization of bacterial polysomes. *Cell*, **136**, 261–271.
46. Brandt, F., Carlson, L.-A., Hartl, F., Baumeister, W. and Grünewald, K. (2010) The three-dimensional organization of polyribosomes in intact human cells. *Mol. Cell*, **39**, 560–569.
47. Afonina, Z.A., Myasnikov, A.G., Shirokov, V.A., Klaholz, B.P. and Spirin, A.S. (2014) Formation of circular polyribosomes on eukaryotic mRNA without cap-structure and poly (A)-tail: a cryo electron tomography study. *Nucleic Acids Res.*, **42**, 9461–9469.
48. Myasnikov, A.G., Afonina, Z.A., Ménétret, J.-F., Shirokov, V.A., Spirin, A.S. and Klaholz, B.P. (2014) The molecular structure of the left-handed supra-molecular helix of eukaryotic polyribosomes. *Nat. Commun.*, **5**, doi:10.1038/ncomms6294.
49. Tuller, T. and Zur, H. (2015) Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res.*, **43**, 13–28.
50. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. and Blüthgen, N. (2013) Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.*, **9**, doi:10.1038/msb.2013.32.
51. Li, Q. and Qu, H.-Q. (2013) Human coding synonymous single nucleotide polymorphisms at ramp regions of mRNA translation. *PLoS One*, **8**, e59706.
52. Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I. and Pilpel, Y. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.
53. Vogel, C., Abreu Rde, S., Ko, D., Le, S.Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M. and Penalva, L.O. (2010)

- Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol. Syst. Biol.*, **6**, doi:10.1038/msb.2010.59.
54. Hornburg,D., Drepper,C., Butter,F., Meissner,F., Sendtner,M. and Mann,M. (2014) Deep proteomic evaluation of primary and cell line motoneuron disease models delineates major differences in neuronal characteristics. *Mol. Cell Proteomics*, **13**, 3410–3420.
 55. Schwanhäusser,B., Busse,D., Li,N., Dittmar,G., Schuchhardt,J., Wolf,J., Chen,W. and Selbach,M. (2011) Global quantification of mammalian gene expression control. *Nature*, **473**, 337–342.
 56. Jackson,R.J. and Hunt,T. (1983) Preparation and use of nuclease-treated rabbit reticulocyte lysates for the translation of eukaryotic messenger RNA. *Methods Enzymol.*, **96**, 50–74.
 57. Pfaffl,M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.*, **29**, e45.
 58. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.
 59. Williams,E.J. (1959) *Regression Analysis*. Wiley, NY.
 60. Steiger,J.H. (1980) Tests for comparing elements of a correlation matrix. *Psychol. Bull.*, **87**, 245–251.
 61. Arava,Y., Boas,F.E., Brown,P.O. and Herschlag,D. (2005) Dissecting eukaryotic translation and its control by ribosome density mapping. *Nucleic Acids Res.*, **33**, 2421–2432.
 62. Gilchrist,M.A. (2007) Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol. Biol. Evol.*, **24**, 2362–2372.
 63. Warner,J.R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.*, **24**, 437–440.
 64. Frank,J. and Gonzalez,R.L. Jr (2010) Structure and dynamics of a processive Brownian motor: the translating ribosome. *Annu. Rev. Biochem.*, **79**, 381–412.
 65. Wohlgenuth,I., Pohl,C. and Rodnina,M.V. (2010) Optimization of speed and accuracy of decoding in translation. *EMBO J.*, **29**, 3701–3709.
 66. Zouridis,H. and Hatzimanikatis,V. (2008) Effects of codon distributions and tRNA competition on protein translation. *Biophys. J.*, **95**, 1018–1033.
 67. Bilgin,N., Claesens,F., Pahverk,H. and Ehrenberg,M. (1992) Kinetic properties of *Escherichia coli* ribosomes with altered forms of S12. *J. Mol. Biol.*, **224**, 1011–1027.
 68. Rodnina,M.V., Pape,T., Fricke,R., Kuhn,L. and Wintermeyer,W. (1996) Initial binding of the elongation factor Tu·GTP·aminoacyl-tRNA complex preceding codon recognition on the ribosome. *J. Biol. Chem.*, **271**, 646–652.
 69. Pape,T., Wintermeyer,W. and Rodnina,M.V. (1998) Complete kinetic mechanism of elongation factor Tu-dependent binding of aminoacyl-tRNA to the A site of the *E. coli* ribosome. *EMBO J.*, **17**, 7490–7497.
 70. Savelsbergh,A., Katunin,V.I., Mohr,D., Peske,F., Rodnina,M.V. and Wintermeyer,W. (2003) An elongation factor G-induced ribosome rearrangement precedes tRNA-mRNA translocation. *Mol. Cell*, **11**, 1517–1523.
 71. Soto Rifo,R., Ricci,E.P., Decimo,D., Moncorge,O. and Ohlmann,T. (2007) Back to basics: the untreated rabbit reticulocyte lysate as a competitive system to recapitulate cap/poly(A) synergy and the selective advantage of IRES-driven translation. *Nucleic Acids Res.*, **35**, e121.
 72. Ghazalpour,A., Bennet,B., Petyuk,V.A., Orozco,L., Hagopian,R., Mungrue,I.N., Farber,C.R., Sinsheimer,J., Kang,H.M., Furlotte,N. et al. (2011) Comparative analysis of proteome and transcriptome variation in Mouse. *PLoS Genet.*, **7**, e1001393.
 73. Tuller,T., Kupiec,M. and Ruppin,E. (2007) Determinants of protein abundance and translation efficiency in *S. cerevisiae*. *PLoS Comput. Biol.*, **3**, e248.
 74. Gunawardana,Y. and Niranjana,M. (2013) Bridging the gap between transcriptome and proteome measurements identifies post-translationally regulated genes. *Bioinformatics*, **29**, 3060–3066.

RiboAbacus: a model trained on polyribosome images predicts ribosome density and translational efficiency from mammalian transcriptomes

F. Lauria, T. Tebaldi, L. Lunelli, P. Struffi, P. Gatto, A. Pugliese, M. Brigotti, L. Montanaro, Y. Ciribilli, A. Inga, A. Quattrone, G. Sanguinetti, G. Viero

Supplementary figures

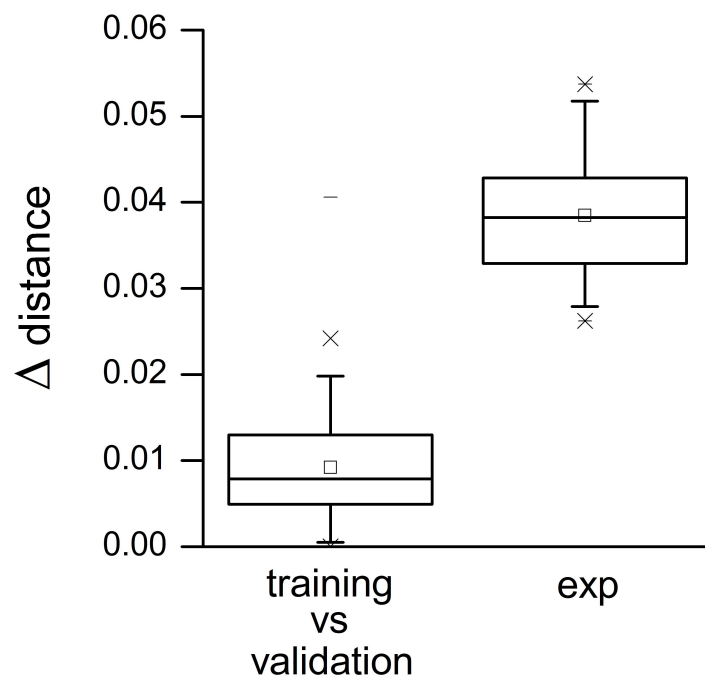


Figure S1: Left: boxplot of Δ differences between a) the distance between the experimental distributions of the number of ribosomes per transcript and the distribution predicted by RiboAbacus with training set, b) the distance between the experimental distributions of the number of ribosomes per transcript and the distribution predicted by RiboAbacus with validation set (100 rounds of 50-50 cross-validation on transcripts, Hek-293 dataset). Right: distances between the distributions of the number of ribosomes per transcript when splitting experimental AFM data (100 rounds of 50-50 cross-validation on AFM data, Hek-293 dataset). Wilcoxon-Mann-Whitney p-value < 0.001 .

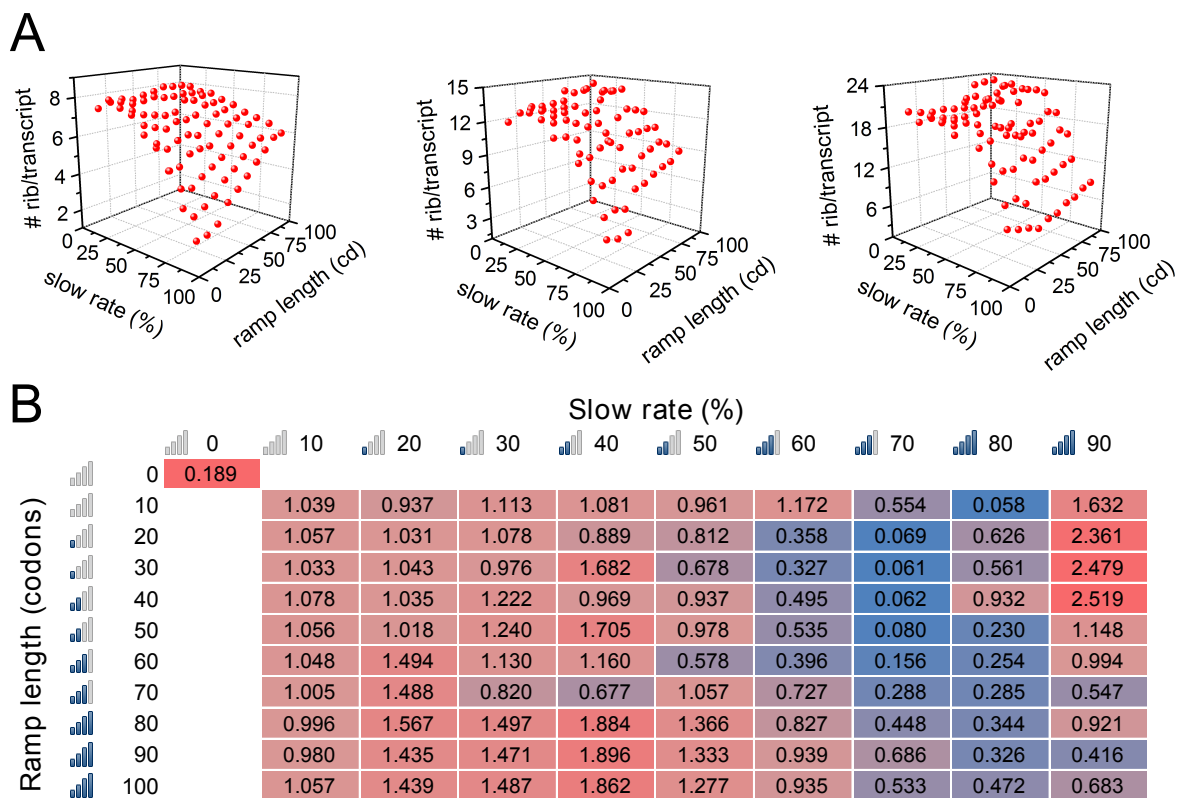


Figure S2: **Model performance in Hek-293 at varying ramp parameters.** (A) 3D scatterplots showing the means (z-axes) respectively of the three Gaussian curves fitting the ribosome per transcript distributions obtained with different combinations of ramp length and ribosome slowdown rate. (B) Heatmap showing the weighted sum of the Kullback-Leibler divergences computed between the Gaussian curves that fit experimental and predicted data at varying ramp parameters.

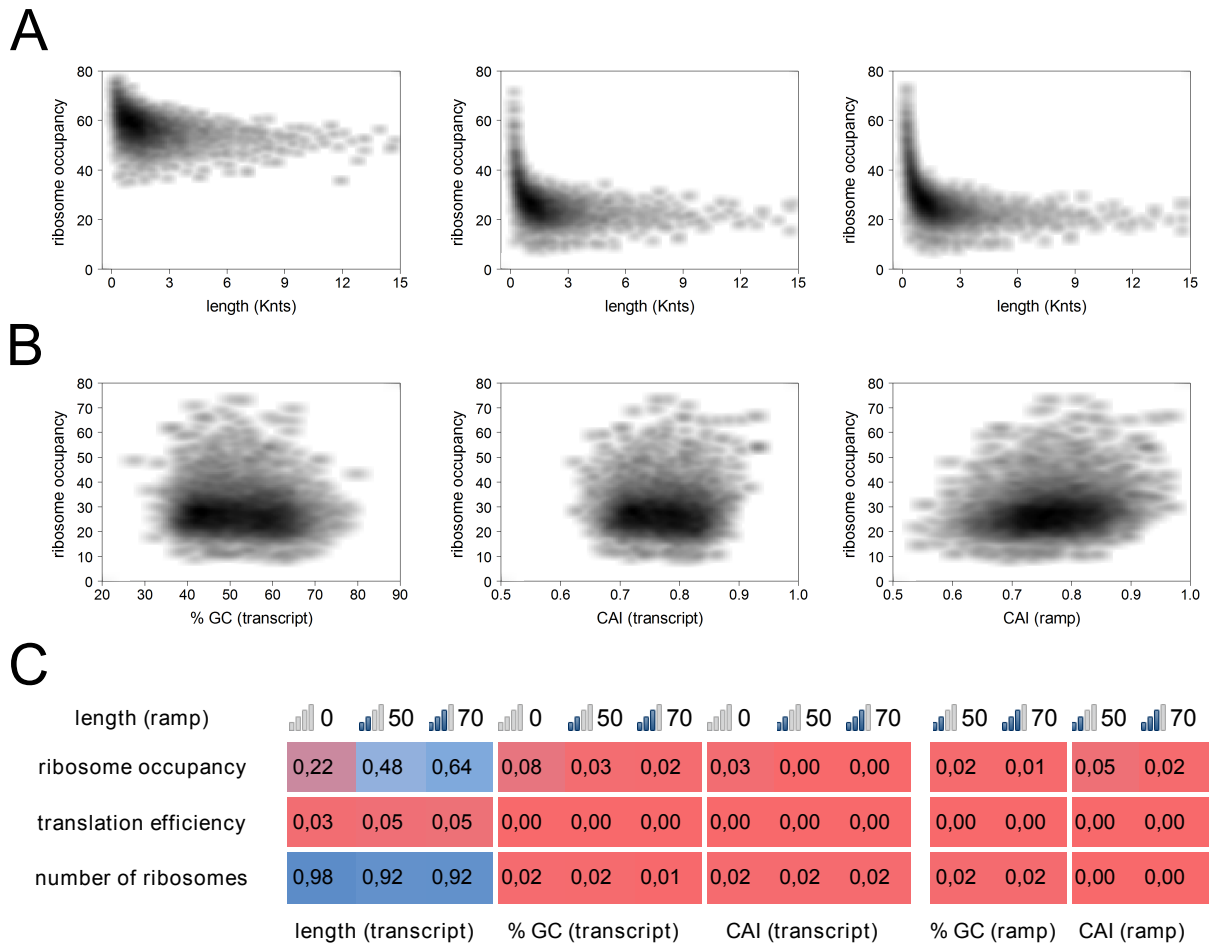


Figure S3: **Correlation between transcript features and RiboAbacus predictions for Hek-293 transcriptome.** (A) Scatterplots of transcript length versus ribosome occupancy. Example for model without ramp hypothesis (left panel), ramp length 50 codons and slowdown rate 70% (middle panel) and ramp length 70 codons and slowdown rate 70% (right panel) are shown. Each dot represents an mRNA. (B) Scatterplots of three transcript features (transcript GC content, Codon Adaptation Index (CAI) of the whole coding sequence and CAI of the ramp region) versus ribosome occupancy. Examples are shown for ramp length 50 codons and slowdown rate 70%. (C) Correlations between transcript features (coding sequence length, coding sequence and ramp GC content, coding sequence and ramp CAI) and RiboAbacus predictions (ribosome occupancy, translation efficiency and number of ribosomes per transcript).

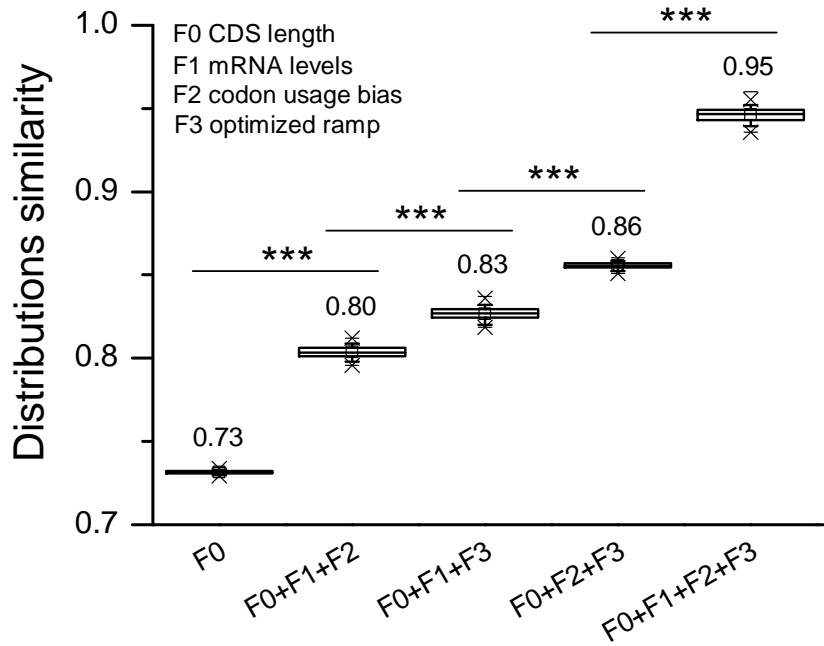


Figure S4: **Contribution of transcript features to RiboAbacus predictions.** Boxplot showing the similarities (calculated as 1-distance) between the experimental distribution of the number of ribosomes per transcript and RiboAbacus predicted distributions, based on different combinations of features (F0: CDS length, F1: mRNA level, F2: codon usage bias, F3: optimized ramp parameters). Similarities were calculated in 100 rounds of cross-validation. Statistical significances from Wilcoxon-Mann-Whitney test are shown: (***, p-value < 0.001).



Figure S5: Heatmap showing, for the MCF-7 dataset, the average distance (100 cross validations) between the experimental and the predicted distribution of the number of ribosomes per transcript, varying the ramp length parameter (from 0 to 100 codons) and the ribosome slowdown rate parameter (from 0 to 90%). Higher distances are highlighted in red gradient, smaller distances in blue gradient. The minimum distance value is obtained with ramp length of 50 codons and ribosome slowdown rate of 70%.

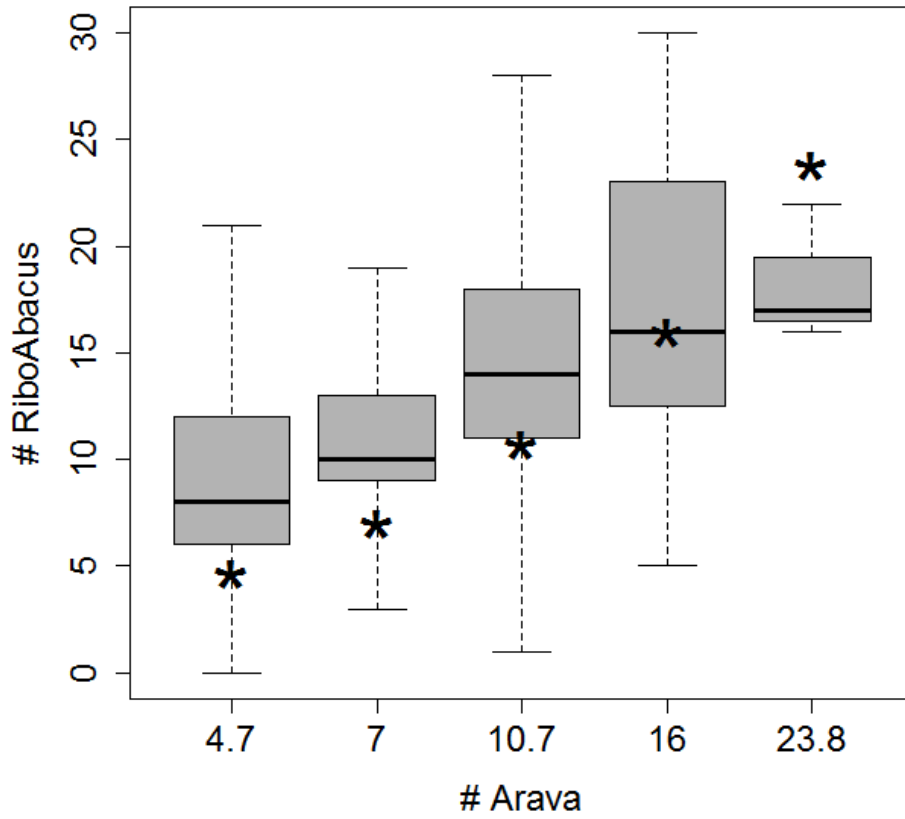


Figure S6: Single transcript comparison between the estimation of the number of ribosomes per transcript in yeast provided with polysome profiling followed by microarray (Arava et al. 2003, reference 40 in the main text) and the estimation provided by RiboAbacus trained on the AFM distribution shown in Figure 1C. Transcripts associated with steady-state polysomes have been classified in 5 populations, according to the discrete (fraction by fraction) estimation of the number of ribosomes per transcript provided by Arava. For each population, the distribution of the number of ribosomes per transcript provided by RiboAbacus is drawn as a box-whisker plot. Asterisks mark the correspondence between measures on the x and the y axes.

2.2 riboSim

Polysomes are characterised by the number of ribosomes per transcript and their localization along the mRNA. In the previous chapter I showed that riboAbacus provides a good estimation of protein abundances starting from mRNA levels by accurately forecasting the number of ribosomes bound to the mRNAs. Because of the deterministic nature of the model, no information about their localization along the transcript can be obtained. An important information about the biology of polysomes is the fact that ribosomes can accumulate along transcripts due to ribosome slowdown^{82,253} and stalling^{97,254}, two scenarios possibly connected to many pathologies such as neurodegenerative diseases^{47,255}, diabetes and multi-systemic failure²⁵⁶. To extract ribosome positional information for investigating translation at single nucleotide resolution and genome-wide scale, ribosome profiling (RiboSeq)^{123,150} has been recently developed. One of the outcomes of RiboSeq is the so-called ribosome occupancy profile, i.e. a transcript-specific curve where the height of the signal is proportional to the probability to find a ribosome within a specific mRNA position along the sequence. Thus, ribosome occupancy profiles display regions of the transcript where ribosomes slowdown and ribosome stalling is more likely occurring. Nevertheless, the contribution translation acting in *cis* (mRNA sequences and secondary structures)^{57,58} and *trans* (ncRNA and RNA binding proteins) in determining ribosome pausing cannot be easily established.

The last years have witnessed the development of many stochastic models of translation aimed at predicting transcript- and codon-specific initiation, elongation and termination rates and consequently protein abundances^{121,181,215,217,218,224,228,232,257}. Many of them are based on the fitting of ribosome profiling data (i.e. the number of mapped fragments per transcripts and ribosome occupancy profiles)^{121,181,219,228}. At present, none of them pay specific attention in investigating ribosome localization and at least two shortcomings can be identified in these approaches. In fact, the comparisons between predicted and RiboSeq ribosome occupancy profiles i) are usually based on the ribosome coverage along the whole transcript, not taking into account the many local fluctuations of the ribosome occupancy profiles and ii) are aimed at simultaneously tuning the parameters of the model to reach the best fit of the experimental data. The first issue may lead to inaccurate results due the existence of global similarities not always representative of the local ones, while the second makes it difficult to understand the precise role of the individual determinants, whose individual contribution to the final simulation cannot be distinguished.

To address these shortcomings I developed RiboSim, a stochastic model of translation that simulates the binding, the movement and the release of ribosomes from transcripts. Different determinants (i.e. CDS length, codon usage bias and ramp) are progressively step by step to evaluate their contribution. This progressive approach allows to individually investigate each feature and to assess its contribution in forecasting ribosome localization. At each step, a codon-by-codon comparison between predicted and experimental ribosome occupancy profiles enables to monitor the effect of the inclusion of every new feature to the model, so that the higher the increase of the profile similarities the higher the benefit of including the determinant in the model.

From the biological point of view, determinants inducing an increase in the correlation between predicted and experimental profiles can be labelled as mayor players in controlling ribosome movement localization along the transcripts and, more in general in tuning translation.

2.2.1 Materials and methods

Assumptions

Due to the high complexity of translation initiation^{19,20} and the lack of conclusive experimental measurement of its kinetics constants, riboSim represents the entire initiation step by simulating the binding of a new ribosome at the start codon. As discussed below, for the same reason the initiation probability is set as the average of the codon usage values of the whole coding sequence.

In order to have a comprehensive description of polysome organization as ribosome positions and ribosome number per transcripts, I evaluated in parallel the positional information provided by riboSim with the number of ribosomes predicted by riboAbacus. Hence, I made the following assumptions to make the two models as coherent as possible.

- First, given the facts that riboSim analyses each transcript independently and that in cells free ribosomes are highly abundant^{258,259}, I assumed that competition of ribosome for translation is not a limiting step for the simulation.
- Second, I assumed that a ribosome reaching the end of the coding sequence dissociates and then starts a new cycle of translation. Hence, the probability of translation termination was always set equal to the initiation one (when different from zero). Note that the deterministic nature of riboAbacus automatically led to this assumption when the system reached the steady

state, while for the stochastic simulation with riboSim, this assumption must be explicitly included.

- Third, the number of codons covered by one ribosome (ribosome footprint) is set to 10 codons¹³⁸, 4 triplets at the left of the P-site and 5 triplets at its right respectively (Figure 2.1A). As for riboAbacus, this parameter is required for maintaining the minimum distance between the P-sites of consecutive ribosomes (Figure 2.1B) and for the choice of the correct initiation and elongation probabilities in the case of stalled ribosomes along the CDS. Ribosome footprint is also required for generating the predicted ribosome density profiles, as explained below.
- Finally, at each step of the progressive approach features already investigated in riboAbacus are included step by step, i.e. the CDS length, the codon usage bias and the ramp hypothesis.

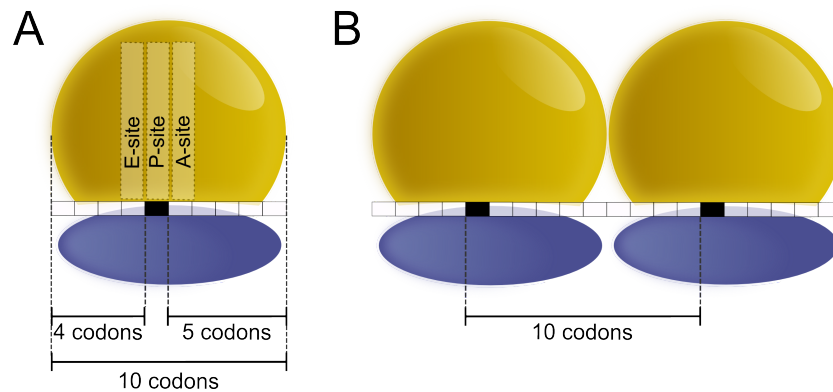


Figure 2.1. **Ribosome footprint.** (A) Schematic representation of a ribosome and the covered portion of the transcript. The length of the ribosome footprint is 10 codons. (B) Schematic representation of adjacent ribosomes: the distance between their P-site is defined by the chosen ribosome footprint.

Note that, differently from riboAbacus, riboSim does not take into account the mRNA level as a feature for improving the goodness of the predicted ribosome occupancy profiles. Indeed, the sequencing of ribosome protected fragments allows to determine, by sequence alignment, the identity of the transcript generating the read. It is then possible to compare each predicted profile directly to the corresponding experimental one and the abundance of single transcripts doesn't provide any additional information and in no way it influences the resulting correlations.

Model

To provide a stochastic simulation of translation, I employed the totally asymmetric simple exclusion process (TASEP)²²¹, based on the Gillespie algorithm²²². riboSim reproduces the elongation phase of the process, simulating the binding of ribosomes at the start codon, their movement along the coding sequence and the dissociation of the two subunits when the stop codon is reached. The TASEP algorithm ensures that ribosomes can move only in one direction (from the 5' to the 3' end of the CDS) and only if the triplet downstream the 3' head of each ribosome is not occupied by another one (Figure 2.2).



Figure 2.2. **Scheme of riboSim.** Ribosomes start a new cycle of translation with rate p_i moving from codon i to the next one with elongation probability p_i . This probability is set to zero if the movement of each ribosome is prevented by the presence of stalled ribosomes on the following codons. Translation termination occurs with rate p_T .

For each transcript the simulation is divided in the following steps:

1. definition of the initial condition of the system: an empty mRNA;
2. generation of a random value to determine the next reaction to occur among the following:
 - I. the binding of a new ribosome at the start codon;
 - II. the movement of already bound ribosomes from one codon to the next one;
 - III. the detachment of a ribosome from the stop codon.

Note that the probability of the first reaction is zero if there is already a ribosome whose footprint is covering the start codon, thus preventing new ribosomes from starting translation. Similarly, ribosomes along the coding sequence cannot move forward if codons that follows are already occupied by the footprint of a ribosome.

3. update of the system i.e. update the number and/or the position of the ribosomes along the mRNA;
4. back to step 2 until the steady state of the system is reached;

Steady states and ribosome occupancy profiles

The Gillespie algorithm is typically applied to the study of systems composed by a finite number of molecules linked by sets of either reversible or irreversible reactions^{260,261}, and allows to easily understand when the steady state of the system (if any) is reached. In fact, when the abundance of a chosen molecule either exceed or fall below a threshold or the probability for some reactions to occur decreases to zero, it can be reasonably assumed that the simulation can be terminated.

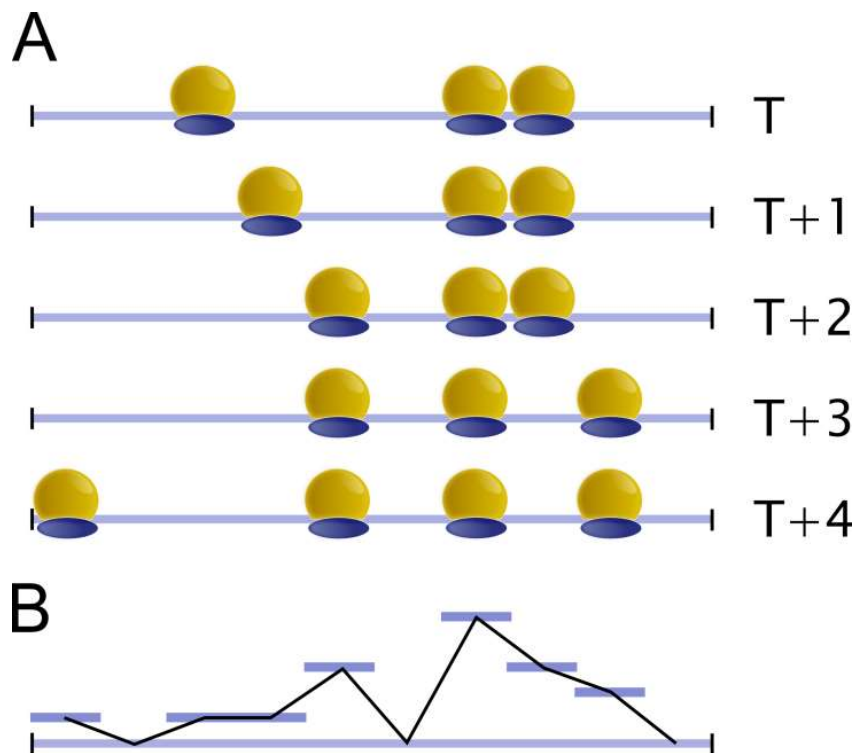


Figure 2.3. **Simplified scheme simulated ribosome movements.** (A) After T iteration of riboSim, the number of ribosomes bound to a transcript is almost constant even if their distribution along the mRNA changes at each time point of the simulation. (B) Trivial example of occupancy profile collecting the 5 snapshots in (A).

riboSim, in line with others TASEP models of translation doesn't deal with a finite number of molecules since it considers a single mRNA and an unlimited pool of ribosomes. Moreover, the whole system keeps changing at each step and a true equilibrium does not exist: a new cycle of the algorithm is always associated to a reaction involving one ribosome. After a certain number of iterations, the number of ribosomes per transcript reaches a plateau with potential variations of ± 1 ribosome

upon binding or release reactions. Thus, I defined as the steady state of the system the moment corresponding to a stable number of total ribosomes bound to the mRNA.

Given an experimental transcriptome, I computed the number of iterations of the algorithm (T) required to reach the steady state for the longest transcript of transcriptome under study. Then I applied the simulation to the whole set of mRNAs T times, to ensure the arrangement of ribosomes along all the transcripts and the definition of a steady state. At this point, for each mRNA I run the algorithm for additional T cycles collecting at each step a snapshot of the simulation to obtain, by merging all the snapshots, the ribosome occupancy profile associated to the transcript, as shown in Figure 2.3.

Occupancy profiles comparison

The performance of riboSim predictions can be tested by comparing the simulated occupancy profiles with the experimental ones obtained by ribosome profiling (Figure 2.4). The comparison between the predicted and the experimental occupancy profile associated to the same transcript is performed codon by codon. After computing the ribosome coverage of each codon (i.e. the number of footprint per codon), a Pearson correlation of the codon coverage in the two profiles can be calculated for every mRNA.

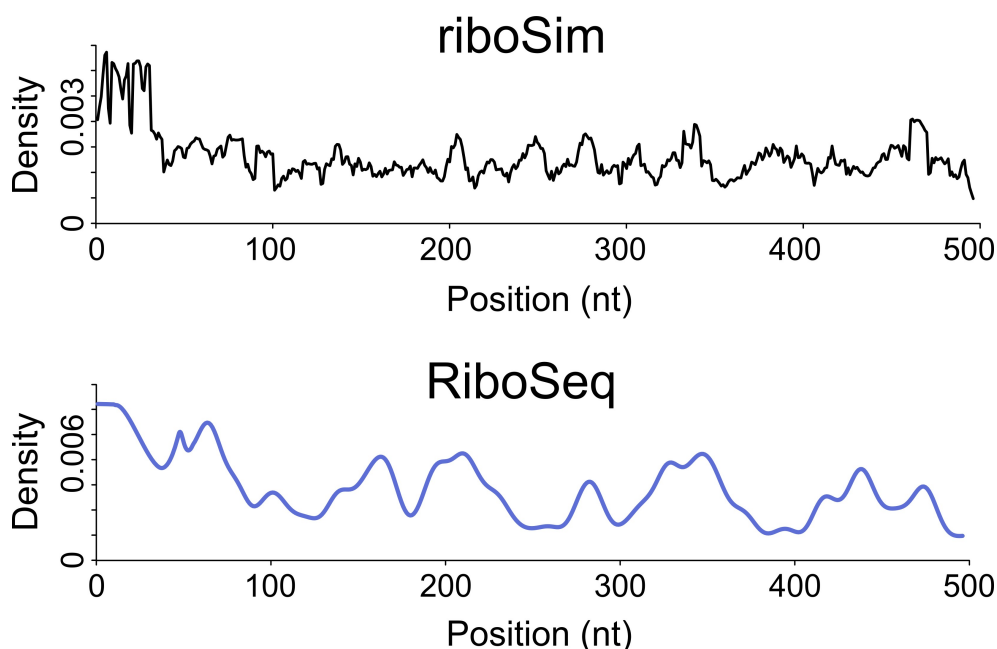


Figure 2.4. **Ribosome occupancy profiles.** Example of predicted (upper panel) and experimental (lower panel) ribosome occupancy profile of yeast APM4 transcript. Experimental data from Nedialkova and Leidel²⁶².

The resulting distribution of the correlation values for the whole transcriptome explains how the predicted profiles are consistent with the experimental data and the greater the mean value, the higher the ability of riboSim to predict ribosome localization. Finally, the one-sided Wilcoxon–Mann–Whitney test is used to compare the correlation distributions generated after the addition of a new feature (CDS length, codon usage bias and ramp). I also built a baseline correlation distribution by comparing randomly generated profiles with the experimental ones.

Use of riboSim

riboSim requires two input files: a list of transcript with their coding sequence and a list of organism-specific codon usage bias values (downloaded from <http://www.kazusa.or.jp/codon/> for my simulations). The transcript file must contain for each transcript two lines: the first reporting general information about the transcripts such as gene ID and transcript ID and the second reporting the nucleotide sequences. The codon usage file must contain the list of codons and the corresponding codon usage bias values, arranged in two columns. An additional file is required for introducing the slowdown of the ribosomes on specific regions of the CDS. Each line must contain the transcript ID, the start and the stop of the CDS region of interest and the ribosome slowdown rate chosen by the user.

The outputs of riboSim are a BED file containing the data for building the predicted occupancy profiles (i.e. the predicted number of footprint per codon) and a file reporting the transcript ID, the CDS length, the number of ribosomes bound to the transcript at the end of the simulation and the associated mRNA coverage. For the comparison with ribosome profiling data, a BED file of the transcriptome (CDS) alignment is needed. riboSim is coded in C.

2.2.2 Results

Datasets

I run riboSim on the transcriptome of three different organisms: *Saccharomyces cerevisiae*, *Mus musculus* and *Homo sapiens*. I chose these organisms to study ribosome localization along mRNAs in three evolutionarily distant species and to verify if each step of the progressive approach (i.e. the addition of CDS length, codon usage bias and ramp) to the model can differently affect the results, depending on the complexity of the organism under consideration. For yeast, the experimental ribosome profiling data (RiboSeq) were obtained from Nedialkova and Leidel²⁶², for mouse I used RiboSeq data from whole mouse brains obtained in the Laboratory of Translational

Architectomics (IBF-CNR, Trento)¹, for human the experimental data coming from Hek-293 cells were published by Gao and collaborators²⁶³. The number of transcripts with sufficient coverage used in the following analyses for yeast, mouse and human are 4887, 4403 and 14423, respectively.

Basic model: CDS length

The basic version of riboSim performs the simulation assigning the same probability p to each feasible reaction (initiation, elongation and termination). As already explained, the probability p becomes zero for unfeasible reactions, for example during the elongation stage of ribosomes when the first codon downstream its head is already occupied by the footprint of another ribosome. In this basic case, the only feature that differentiates mRNAs is the length of their coding sequence.

Running this basic model on the whole transcriptomes, I obtained three sets of transcript-specific profiles representing mRNAs ribosome occupancies in the three organisms. As discussed in the Method section, the predicted profiles are comparable with the experimental ones produced by RiboSeq (Figure 2.5A) by computing a Pearson correlation between them, with codon resolution. I compared the resulting correlation distribution with a random distribution specifically built to be employed as a lower baseline (Figure 2.5B). To understand if the choice of the random values employed in the basic model affects the predicted profiles, for each organism I also compared the results of two simulations with different seeds (i.e. the value used to initialize the random number generator). I will refer to the resulting correlation distribution as upper baseline that should be interpreted as follows: the greater the mean value, the lower the impact of the choice of random values on the simulation. For the basic model the upper baseline is almost centred at 1 for the three organisms (Figure 2.5B, dark gray boxes), implying that the chosen random values have basically no impact in the outcome of the predicted profiles.

The comparison between experimental and predicted profiles show low correlations and almost all three simulations are identical to the lower baseline, meaning that riboSim with no feature except for the CDS length does not provide good forecasting of ribosome occupancy profiles, i.e. of ribosome localization. This result reinforces the findings of riboAbacus about the inadequacy of the mRNA length to account for ribosome dynamics and number of ribosomes per transcript.

¹ Please refer to the Appendix at the end of the elaborate for the experimental protocol of ribosome profiling, the pre-processing of the data and the alignment steps.

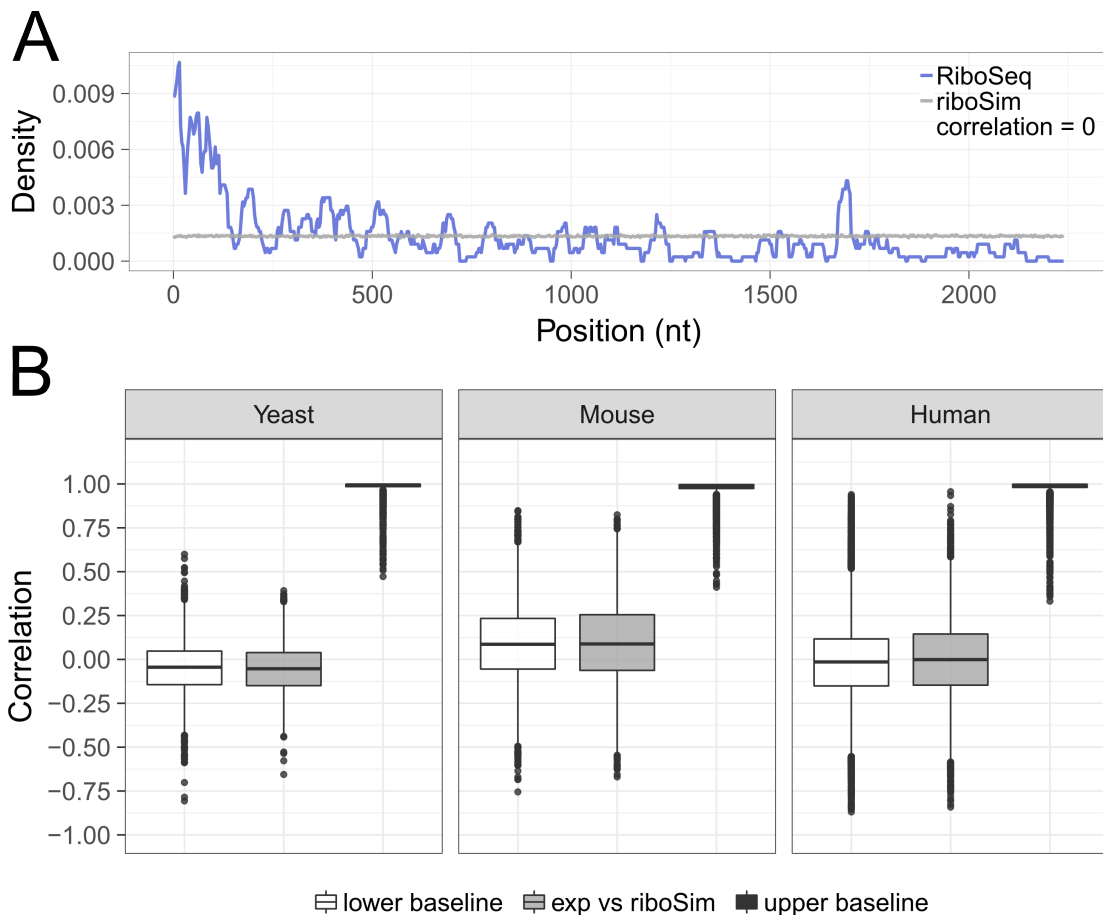


Figure 2.5. **Performance of basic riboSim.** (A) Example of experimental ribosome occupancy profile (blue line) and predicted profile from the basic model (gray line), associated to *YOR299W*. The correlation between the profiles is also reported. (B) Distributions of correlation values comparing experimental and predicted ribosome occupancy profiles for yeast, mouse brain and human Hek-293 using the corresponding transcriptomes. Predicted profiles were generated using riboSim (in grey) using as transcript-specific feature only the CDS length. Lower baseline correlation distributions were generated comparing experimental profiles with randomly generated profiles (in white). Upper baseline correlation distributions were generated comparing two sets of predicted profiles produced by the basic model with different seeds (in black).

Codon usage bias model

Given the above-mentioned results, I included in the model a second feature, the codon usage bias. To do this, I modified the probabilities associated to the movement of ribosomes along each transcript. In particular, I substituted the values that determine the choice of the ribosome to move forward changing the probability to move such that it would have been proportional to the codon usage bias (*CU*) values of the codon where the ribosome is located. Thus, the elongation probability p_i for codon i for a CDS of length N is defined as follows:

$$p_i = \begin{cases} 0 & \text{if ribosomes are colliding} \\ CU_i & \text{if } 1 \leq i < N \end{cases}$$

Basically, this formula ensures that the ribosomes positioned on non-optimal codons, compare to those on optimal codons¹¹³ would spend more time before moving forward, to those. The probability associated to the binding of new ribosomes p_I (probability of initiation) and to their detachment from the stop codon p_T (probability of termination) were set to the average codon usage value of the transcript. In fact, conclusive experimental measurement of initiation and termination rates is lacking. This choice of the initiation and the termination probability assure the removal of potential biases during the simulation caused by either excessively high or low initiation and termination probabilities with respect to the elongations ones¹²⁹. For a CDS of length N I set:

$$p_I = p_T = \frac{\sum_{i \in Cod} CU_i}{N - 1}$$

where *Cod* is the set of all the codons of the CDS but the last one.

I applied this variant of riboSim to yeast, mouse and human transcriptomes as in the previous case, and parsed the correlation between the predicted and the experimental profiles (Figure 2.6). In parallel I computed a new set of upper baselines, confirming the low impact of the choice of random values on predicted profiles (Figure 2.6B). The result shows a clear and significant increase of the correlation distribution toward positive values in yeast (Figure 2.6B dark gray box), a slighter positive shift in mouse and no shift at all in human, where the correlation distribution is still centred around 0. These findings suggest that for a simple organism such as yeast the codon usage bias alone seems to be sufficient for good predictions of ribosome localization along the mRNAs, while more complex organisms likely require additional features.

It has to be noted that the codon usage bias values employed in the simulations are based on the frequency of the triples along the mRNA, which are proportional to the abundance of the corresponding tRNA in the cell^{109–112}. Nevertheless, an empirical codon usage can be computed starting from the experimental ribosome occupancy profiles. This experimental codon usage may take into account possible additional control of translation and were used as a possible better proxy for the usage of codons in translation. In fact, I can reasonably assume that the intensity of the signal on a specific codon is inversely proportional to the speed of ribosomes on the triplet. From this assumption, I computed the empirical codon usage from ribosome profiling in yeast, run riboSim in the same organism, employing this time the new codon usage

values, and repeated the comparisons with the experimental profiles. Since the resulting correlation distribution showed no significant shift towards higher correlation values with respect to the previous simulation (data not shown), I abandoned this line of investigation and proceeded with the progressive approach, including in the model also the ramp hypothesis.

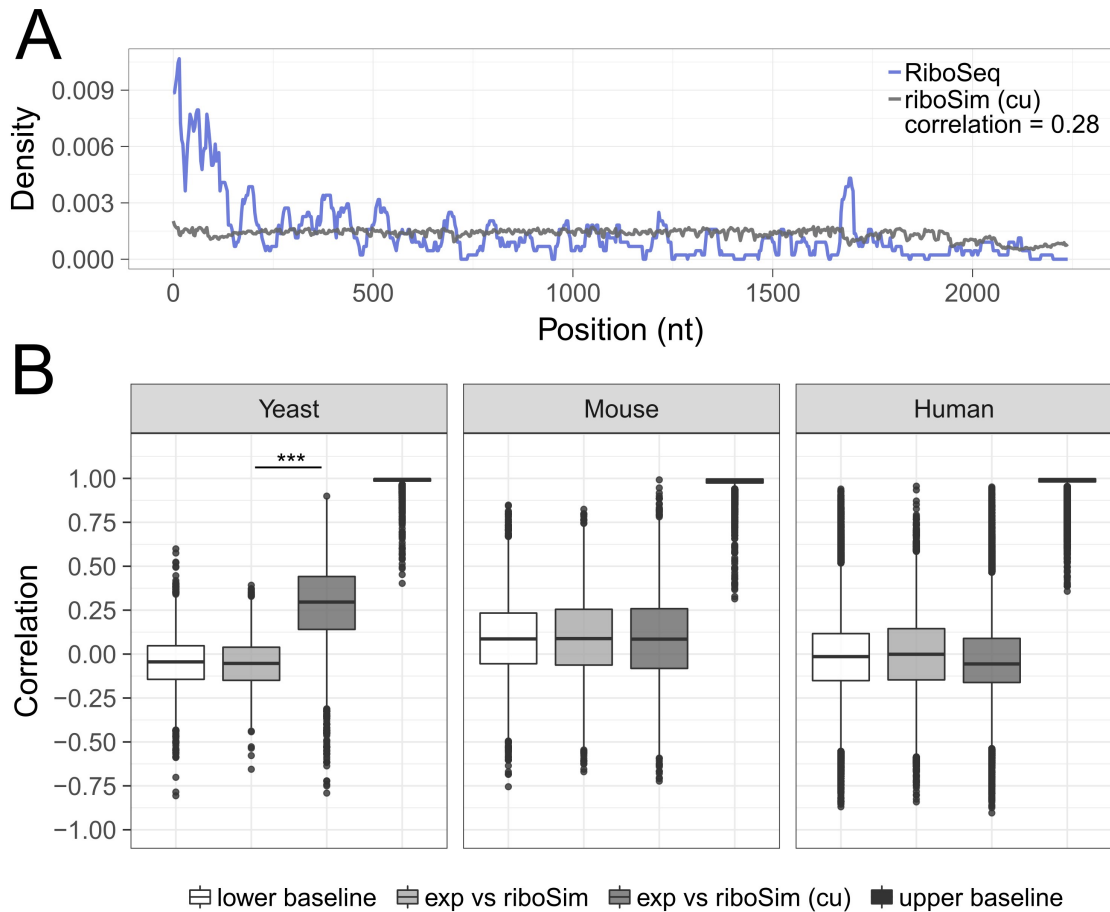


Figure 2.6. **Contribution of the codon usage bias in prediction performance.** (A) Example of experimental ribosome occupancy profile (blue line) and predicted profile from the codon usage bias model (gray line), associated to *YOR299W*. The correlation between the profiles is also reported. (B) Distributions of correlation values comparing experimental and predicted ribosome occupancy profiles for yeast, mouse brain and human Hek-293 using the corresponding transcriptomes. The predicted profiles were generated using riboSim integrated with the codon usage bias (dark grey). Statistical significance values from the Wilcoxon-Mann-Whitney test are shown (***) p-value < 0.001). Lower baseline correlation distributions were generated comparing experimental profiles with randomly generated profiles (in white). Upper baseline correlation distributions were generated comparing two sets of predicted profiles produced by the codon usage bias model with different seeds (in black).

Ramp hypothesis model

Results obtained with riboAbacus pointed to the ramp hypothesis as a major determinant for a correct estimation of the number of ribosomes per transcript, but this conclusion doesn't give any hint about the contribution of this feature in forecasting ribosome localization. For this reason, the third feature added to riboSim was the ramp, i.e. a slowdown of ribosomes at the beginning of the coding sequence¹²⁰⁻¹²⁵. The ramp hypothesis in riboSim was included by adding the two parameters that characterise the ramp: its length and the slow down rate of the ribosomes. Thus for ribosomes moving on the first L codons of the CDS, I introduced a correction factor that takes into account a slowdown rate SD. In this way I obtained a new set of elongation probabilities for the i^{th} codon:

$$p_i = \begin{cases} CU_i \cdot (1 - SD) & \text{if } 1 \leq i \leq L \\ CU_i & \text{if } L < i < N \end{cases}$$

where N is the length of the CDS and L is the length of the ramp region. Initiation and termination probabilities were calculated as in the previous model.

I decided to run riboSim employing the parameters for the ramp optimized during the training step of RiboAbacus (50 codons of ramp length and 70% of ribosome slowdown). Importantly, this values is very close to that observed experimentally^{82,115,138}. I then computed the new correlations between predicted and experimental ribosome occupancy profiles (Figure 2.7A), comparing the resulting distributions with the data generated without the ramp hypothesis (Figure 2.7B). I also produced new upper baselines based on this variant of riboSim, always obtaining high correlations.

Similarly to what observed with the previous version of the model, the correlation distribution associated to yeast shows also in this case a clear increase in correlation in yeast, with even more positive values (Figure 2.7B, darker gray bar) than those observed for the codon usage model. The improvement produced by the ramp in yeast was found also in mouse, even if the correlation is lower than in yeast. Importantly, these improvements are completely absent in human. Unexpectedly, in human the trend of the correlation goes in the opposite directions with respect to the simpler species, moving toward negative values.

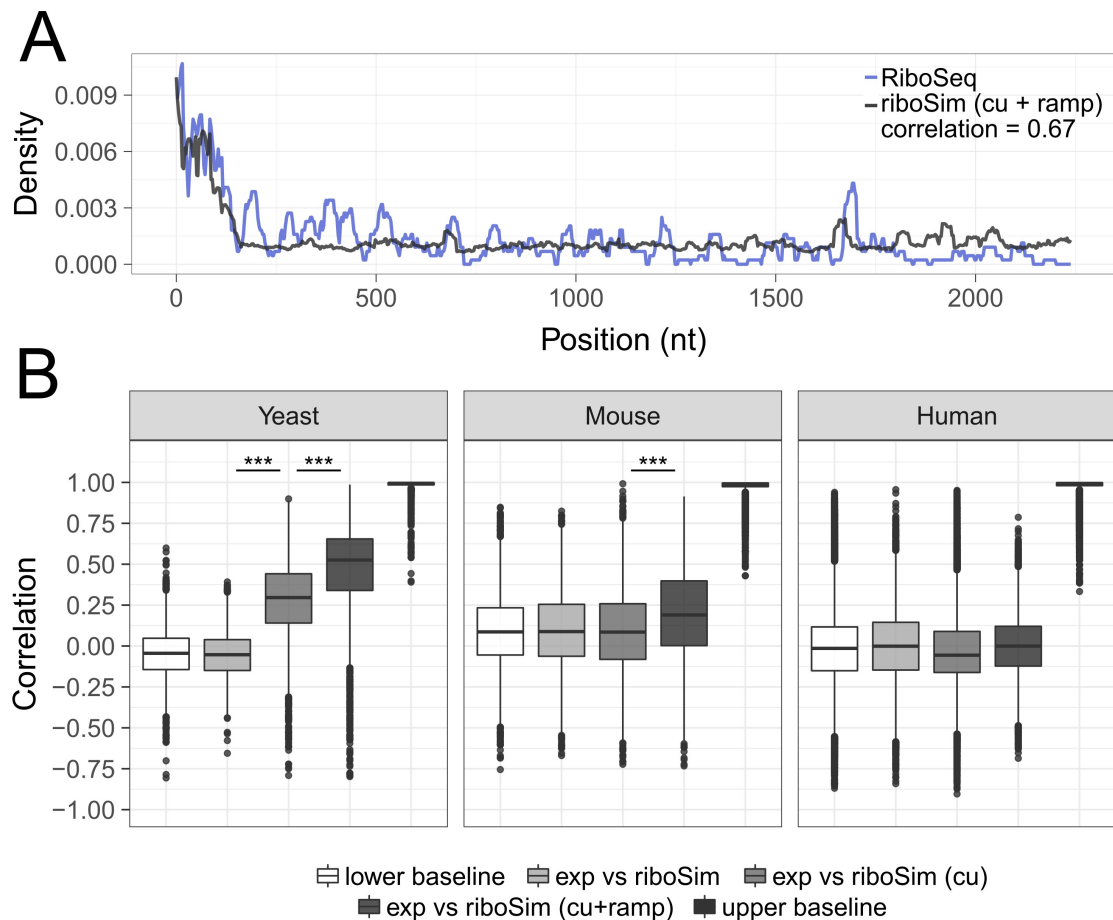


Figure 2.7. **Contribution of the ramp hypothesis in prediction performance.** (A) Example of experimental ribosome occupancy profile (blue line) and predicted profile from the ramp hypothesis model (gray line), associated to *YOR299W*. The correlation between the profiles is also reported. (B) Distributions of correlation values comparing experimental and predicted ribosome occupancy profiles for yeast, mouse brain and human Hek-293 using the corresponding transcriptomes. The predicted profiles were generated using riboSim integrated with the codon usage bias and the ramp hypothesis. The statistical significance values from the Wilcoxon-Mann-Whitney test are shown (***) p-value < 0.001). Lower baseline correlation distributions were generated comparing experimental profiles with randomly generated profiles (in white). Upper baseline correlation distributions were generated comparing two sets of predicted profiles produced by the ramp hypothesis model with different seeds (in black).

2.2.3 Conclusions

Summarising, the discussed findings point out the benefit of a stochastic approach in obtaining good forecasts of ribosome positions along mRNAs in yeast. Firstly, I showed that CDS length alone is not sufficient for predicting ribosome localization in any of the three species under consideration. Secondly, I demonstrated that only in yeast the

codon usage bias alone is a major determinant for increasing the correlation between predicted and experimental profiles with respect to the baseline. This similarity is further improved adding the ramp hypothesis. Thirdly, I proved that in biological systems with higher complexity, the performance of riboSim in predicting ribosome localization is low and that in this case neither the codon usage bias nor the ramp contribute positively to the prediction of ribosome position. In fact, in mouse, only the combination of codon usage bias and the ramp positively affect the predicted ribosome occupancy profiles, while in human neither the codon usage nor the ramp leads to positive correlations. An insightful conclusion can be drawn: additional features, more refined modelling, complex and less known mechanisms of translation control, seem to be necessary to potentially increase prediction performances in human. Notably, the results of riboSim are not in agreement with the findings of RiboAbacus, where the ramp hypothesis provided good predictions of the number of ribosomes per transcript in human.

3 Development of tools for analysing ribosome profiling data

3.1 riboWaltz

Ribosome profiling (RiboSeq) is an experimental technique designed to investigate translation at single nucleotide resolution and genome-wide scale^{123,150}. Following the rapid diffusion of ribosome profiling assays, many computational tools and pipelines dedicated to the analysis of RiboSeq data have been developed in the last years^{140,141,163,172,173,175,178,179,264}. Typically, these tools are aimed at performing differential expression analysis^{163,172,173}, identifying new open reading frames^{175,178,179,264} and, in very few cases, extracting positional information describing fluxes of ribosomes along the RNA^{140,141}.

The extraction of positional information relies on the ability to determine the exact localization of the P-site within ribosome protected fragments (reads), i.e. the site holding the t-RNA which is linked to the growing polypeptide chain during translation. The P-site offset is of crucial importance for a wide range of RiboSeq downstream analyses such as verifying the trinucleotide periodicity of ribosome along the coding sequence^{123,180} and identifying new open reading frames^{165–167}.

Here, I describe the development of riboWaltz, an R package aimed at computing the P-site-offset for all reads from single or multiple samples taking advantage of a two-step algorithm. riboWaltz computes the P-site offset after stratifying the reads in bins according to their length. It showed higher accuracy and specificity in ribosome localization than other existing tools based on a similar approach. riboWaltz provides the user with a variety of graphical representations, laying the foundations for further accurate RiboSeq analyses and better interpretation of positional information.

The paper “*riboWaltz: optimization of ribosome P-site positioning in ribosome profiling data*” reported below is a revised version of the manuscript previously uploaded to BioRxiv and it is going to be submitted to a convenient journal. My contribution in this paper consisted in all refinements of the computational procedures for the identification of the P-site and in the development of the whole R package, from the scripts of the functions to their documentation. riboWaltz is available at <https://github.com/LabTranslationalArchitectomics/riboWaltz>.

riboWaltz: optimization of ribosome P-site positioning in ribosome profiling data

Fabio Lauria^{1§*}, Toma Tebaldi^{2§}, Paola Bernabò¹, Ewout J.N. Groen³, Thomas H. Gillingwater³, Gabriella Viero^{1*}

¹Institute of Biophysics, CNR Unit at Trento, Italy

²Centre for Integrative Biology, University of Trento, Italy

³Euan MacDonald Centre for Motor Neurone Disease Research, University of Edinburgh, UK

⁴Centre for Integrative Physiology, University of Edinburgh, Edinburgh, UK

*Co-corresponding authors

§These authors equally contributed to this work

ABSTRACT

Ribosome profiling is a powerful technique used to study translation at the genome-wide level, generating unique information concerning ribosome positions along RNAs. Optimal localization of ribosomes requires the proper identification of the ribosome P-site in each ribosome protected fragment, a crucial step to determine trinucleotide periodicity of translating ribosomes, and draw correct conclusions concerning where ribosomes are located. To determine the P-site within ribosome footprints at nucleotide resolution, the precise estimation of its offset with respect to the protected fragment is necessary. Here we present riboWaltz, an R package for calculation of optimal P-site offsets, diagnostic analysis and visual inspection of ribosome profiling data. Compared to existing tools, riboWaltz shows improved accuracies for P-site estimation and neat ribosome positioning in multiple case studies. riboWaltz was implemented in R and is available as an R package at <https://github.com/LabTranslationalArchitectomics/RiboWaltz>.

Contact: gabriella.viero@cnr.it or fabio.lauria@unitn.it

Introduction

Ribosome profiling (RiboSeq) is an experimental technique used to investigate translation at single nucleotide resolution and genome-wide scale (Ingolia et al., 2009; Ingolia et al., 2012), through the identification of short RNA fragments protected by ribosomes from nuclease digestion (Steitz et al., 1969; Wolin et al., 1988). The last few years have witnessed a rapid adoption of this technique and a consequent explosion in the volume of RiboSeq data (Michel and Baranov 2013; Brar and Weissman, 2015). In parallel, a number of dedicated computational algorithms were developed for extracting transcript-level information, including novel translation initiation sites, coding regions and differentially translated genes (Xiao et al., 2016; Zhong et al., 2017), as well as positional information describing fluxes of ribosomes along the RNA at sub-codon resolution (Martens et al., 2015, Legendre et al., 2016) and conformational changes in ribosomes during the elongation step of translation (Lareau et al., 2014).

Much of this information relies on the ability to determine, within ribosome protected fragments (reads), the exact localization of the P-site, i.e. the site holding the t-RNA, which is linked to the growing polypeptide chain during translation. This position can be specified by the distance of the P-site from both 5' and 3' ends of the reads, the so-called P-site Offset, PO (**Figure 1A**). Accurate determination of the PO is a crucial step to verify the trinucleotide periodicity of ribosomes along coding regions (Ingolia et al., 2009, Guo et al., 2010), derive reliable translation initiation and elongation rates (Gritsenko et al., 2015; Michel et al., 2014), accurately estimate codon usage bias and translation pauses (Pop et al., 2014, Weinberg et al., 2016), and reveal novel translated regions in known protein coding transcripts or ncRNAs (Hsu et al., 2016; Kochetov et al., 2016; Raj et al., 2016).

Typically the PO is defined as a constant number of nucleotides from either the 3' or 5' end of ribosome protected fragments, independently from their length (**Figure 1A**) (Gao et al., 2015). This approach may lead to an inaccurate detection of the P-site's position owing to potential offset variations associated with the length of the reads. This problem is frequently resolved by selecting subsets of reads with defined length (Bazzini et al., 2014; Han et al., 2014). As such, this procedure removes from the analysis reads that are potentially derived from fragments associated to alternative conformations of the ribosome (Chen et al., 2012; Budkevich et al., 2014) and characterized by shorter or longer lengths (Lareau et al., 2014). Recently, computational tools have been developed to assist with RiboSeq analysis and P-site localization, for example Plastid (Dunn and Weissman, 2016) and RiboProfiling (Popa et al., 2016). Both tools compute the PO after stratifying the reads in bins, according to their length. However, each bin is treated independently, possibly leading to excessive variability of the offsets across bins.

Here, we describe the development of riboWaltz, an R package aimed at computing the PO for all reads from single or multiple RiboSeq samples. Taking advantage of a two-step algorithm where offset information is passed through populations of reads with different length in order to maximize offset coherence, riboWaltz computes with extraordinary precision the PO, showing higher accuracy and specificity of P-site positions than the other methods. riboWaltz provides the user with a variety of graphical representations, laying the foundations for further accurate RiboSeq analyses and better interpretation of positional information.

Design and Implementation

Input acquisition and processing

riboWaltz requires two mandatory input data: 1) alignment files, in BAM format, ideally from transcriptome alignment of RiboSeq reads; 2) transcript annotation files, in GTF/GFF3 format. Alternatively, annotation can be provided as a tab separated text file containing minimal transcript annotation: the length of the transcripts and of their annotated coding sequences and UTRs (**Figure 1B**). Optionally, a third file containing transcript sequence information in FASTA format can be provided as input to perform P-site specific codon sequence analysis (**Figure 1B**).

riboWaltz acquires BAM files and converts them into BED files utilizing the *bamtobed* function of the BEDTools suite (Quinlan and Hall, 2010).

Identification of the P-site position

The identification of the P-site, defined by the position of its first nucleotide within the reads, is based on reads aligning across annotated translation initiation sites (TIS or start codon), and in particular on the distance between their extremities and the start codon itself, as proposed by Ingolia et al., 2009.

riboWaltz specifically infers the PO for each sample in two-steps. At first, riboWaltz groups by length (L) the reads mapping on TIS. To avoid biases in PO calculation, reads whose extremities are too close to the start codon, identified by a parameter called “flanking length” (FL), are discarded from further analysis. Then, for each length group, riboWaltz generates the occupancy profiles of read extremities, i.e. the number of 5' and 3' read ends in the region around the start codon (**Figure 1C**). For each length group, we defined temporary 5' and 3' POs (tPO) as the distance between the first nucleotide of the TIS and the nucleotide corresponding to the global maximum found in the profiles of the 5' and the 3' end at the left and at the right of the start codon, respectively (**Figure 1C**). Therefore, considering the

occupancy profiles as a function f of the nucleotide position x with respect to the TIS, the temporary 5' and 3' PO for reads of length (L) are such that:

$$f(-5'tPO_L) = \max_{x \in [-L+FL, -FL]} f(x)$$

$$f(3'tPO_L) = \max_{x \in [FL-1, L-FL-1]} f(x)$$

The two sets of length-specific temporary POs are defined as:

$$5'tPO = \{5'tPO_{L_{min}}, \dots, 5'tPO_{L_{max}}\}$$

$$3'tPO = \{3'tPO_{L_{min}}, \dots, 3'tPO_{L_{max}}\}$$

where L_{min} and L_{max} are respectively the minimum and the maximum length of the reads.

At the end of this first step, the temporary POs are applied to all the reads (R), obtaining two sets of read-specific tPOs:

$$5'tPO_R = \{5'tPO_{R_1}, \dots, 5'tPO_{R_N}\}$$

$$3'tPO_R = \{3'tPO_{R_1}, \dots, 3'tPO_{R_N}\}$$

where N is the number of reads.

Despite good estimation of P-site positions, artifacts may arise from either the small number of reads with a specific length or the presence of noise in the signal, potentially producing inaccurate results. In other words, the offset estimated independently from the global maximum of each read length is not necessarily the best choice. This approach can produce high variability in PO values of reads differing for only one nucleotide in length (See **Supplementary Tables 1-7**) To minimize this problem, riboWaltz performs a second step for correcting the temporary POs.

The most frequent PO (called optimal PO, oPO) and the associated extremity (optimal extremity) are chosen as reference points to adjust the other values. The optimal PO is selected between the two modes of read specific tPO sets ($Mode(5'tPO_R)$ and $Mode(3'tPO_R)$) as the one with the highest frequency.

$$oPO := \begin{cases} Mode(5'tPO_R) & \text{if } frequency(Mode(5'tPO_R)) \geq frequency(Mode(3'tPO_R)) \\ Mode(3'tPO_R) & \text{if } frequency(Mode(5'tPO_R)) < frequency(Mode(3'tPO_R)) \end{cases}$$

where the notation $| \cdot |$ indicates the cardinality of a set.

Note that this step also selects the optimal extremity to calculate the corrected PO. The correction step is read-length-specific and works as follows: if the offset associated to a length bin is equal to the optimal PO no changes are made. Otherwise, i) the local maxima of the occupancy profiles are extracted; ii) the distances between the first nucleotide of the

TIS and each local maxima is computed; iii) the new PO is defined as the distance in point ii) that is closest to the optimal PO. Summarizing, given the set of local maxima positions (LMP) of the occupancy profile for the optimal extremity, the corrected PO for reads of length L (cPO_L) is such that

$$cPO_L - oPO = \min_{x \in LMP} (x - oPO)$$

Finally, the corrected POs are applied to all the reads.

Output

riboWaltz returns three data structures that can be used in multiple downstream analysis workflows (**Figure 1B**). The first is a list of sample-specific data frames containing for each read i) the position of the P-site (identified by the first nucleotide of the codon) with respect to the beginning of the transcript; ii) the distance between the P-site and both the start and the stop codon of the coding sequence; iii) the region of the transcript (5' UTR, CDS, 3' UTR) where the P-site is located iv) optionally, if a sequence file is provided as input, the sequence of the triplet covered by the P-site. The second data structure is a data frame reporting the percentage of reads aligning across the start codon (if any) and on the whole transcriptome, stratified by sample and read length. Moreover, this file includes the P-site offsets before and after the optimization (tPO and cPO values). The third data structure is a data frame containing, for each transcript, the number of ribosome protected fragments with in-frame P-site mapping on the CDS. This data frame can be used to estimate transcript-specific translation levels and perform differential analysis comparing multiple conditions.

riboWaltz also provides several graphical outputs using the popular “ggplot2” package. riboWaltz plots are described in more detail in the Results section. Any graphical output is returned as a list containing an object of class “ggplot”, further customizable by the user, and a data frame containing the source data for the plot.

Results

riboWaltz overview

In order to show riboWaltz functionalities, we analyzed authors' data obtained from mouse brain samples (GSE102318, see **Supplementary Methods**).

riboWaltz integrates several graphical functions that provide multiple types of output results. First, the distribution of the length of the reads (**Figure 2A**): this is a useful preliminary inspection tool to understand the contribution of each length to the final P-site determination, and possibly decide to remove certain lengths from further analyses. Second, the percentage of P-sites located in the 5' UTR, CDS and 3' UTR regions of mRNAs compared

with a uniform distribution weighted on region lengths, simulating random P-site positioning along mRNAs (**Figure 2B**). This analysis is a good way to verify the expected enrichment of ribosome signal in the CDS. Third, to understand if, and to which extent, P-site determination results in codon periodicity in the CDS, riboWaltz produces a plot with the percentage of P-sites matching one of the three possible translation reading frames for 5' UTR, CDS and 3' UTR, stratifying reads by length (**Figure 2C**). Fourth, the meta-gene read density heatmap, based on the position of read extremities and stratifying reads by length (**Figure 2D**). This plot provides an overview of the occupancy profiles used for P-site determination and allows to check by visual inspection if PO values are reasonable and possibly proceed with manual modification. Fifth, to understand which codons, if any, present higher or lower density of ribosome protected fragments, riboWaltz provide the user with the analysis of the empirical codon usage, i.e. the frequency of in-frame P-sites along the coding sequence associated to each codon, normalized for codon frequency in sequences (**Figure 2E**). Indeed, the comparison of these values in different biological conditions can be of great help to unravel possible defects in aa-tRNAs use or ribosome elongation at specific codons. Finally, single transcripts profiles and meta-gene profiles based on P-site position can be generated (**Figure 3B, top row**) with multiple options: i) combining multiple replicates applying convenient scale factors provided by the user, ii) considering each replicate separately, or iii) selecting a subsets of reads with defined length.

Comparison with other tools

We tested riboWaltz on multiple ribosome profiling datasets in different model organisms: yeast (*S. cerevisiae*, Beaupere et al, 2017; Lareau et al., 2014), mouse (authors' data GSE102318; Shi et al. 2017) and human samples (MCF-7, authors' unpublished data; Hek-293, Gao et al., 2015) and compared our results to those obtained using RiboProfiling (v1.2.2, Popa et al., 2016) and Plastid (v0.4.5, Dunn and Weissman, 2016) (**Table 1** and **Supplementary Tables 1-6**). Comparisons for single datasets are displayed in **Figure 3** and in **Supplementary Figures 1-6**. A summary of overall performances for all the datasets we tested is provided in **Figure 4**.

The first performance measure we considered is the percentage of P-sites with correct frame within the CDS region. For RiboWaltz and RiboProfiling, this measure was comparable in almost all datasets, while Plastid showed lower performances (**Figure 3A, Supplementary Figure 1-6A and Figure 4A**).

Remarkably, meta-profiles produced by riboWaltz displayed a neat periodicity uniquely in the CDS (**Figure 3B** and **Supplementary Figure 1-6B**), with almost no signal along UTRs, neither in the proximity of the start nor of the stop codon. By contrast, both Plastid and

RiboProfiling generated a shift of the start of the periodic region toward the 5' UTR (**Figure 3B** and **Supplementary Figure 1-6B**), suggesting a possible mislocalization of ribosomes before the start and stop codons, an issue that has the potential to generate inaccurate biological conclusions. In order to quantify this effect, we determined a “TIS accuracy score”, comparing the amount of periodic signal before and after the translation initiation site. In the ideal scenario this score is 1, meaning that all the periodicity is restricted in the CDS region. Lower scores are associated with a progressive increase of periodicity in the 5'UTR, indicative of ribosome mislocalization. riboWaltz shows higher TIS accuracy scores with respect to both RiboProfiling and Plastid (**Figure 4B**).

Correct localization of ribosomes is crucial for all downstream positional analyses. Empirical codon usage determination is a popular analysis for ribosome profiling data, equally important for the biological interpretation of results and for building reliable mathematical models of translation. We compared codon usage values based on riboWaltz, RiboProfiling and Plastid (**Figure 3C** and **Supplementary Figures 1-6C**). Correlation values ranged from 0.118 to 0.999, emphasizing that an optional strategy for P-site positioning has a huge impact on downstream analyses.

Availability and future directions

riboWaltz identifies with high precision the position of ribosome P-sites from ribosome profiling data. By improving on other currently-available approaches, riboWaltz can assist with the detailed interrogation of RiboSeq data at single nucleotide resolution, providing precise information that may lay the groundwork for further positional analyses and new biological discoveries.

riboWaltz is written in the R programming language, and can run on Linux, Macs, or Windows PCs. riboWaltz requires several R modules (like GenomicFeatures for handling the GTF/GFF3 files, Biostrings for dealing with FASTA objects and ggplot2 for data visualization), and these dependencies must also be installed. Installation instructions are provided in the manual.

riboWaltz is an Open-Source software package that can be extended in future releases to include other analysis methods as they are developed. Source code for riboWaltz is available is distributed under the MIT license, and available at the following GitHub repository: <https://github.com/LabTranslationalArchitectomics/riboWaltz>. The package includes the R implementation of riboWaltz, data used in this article, and extensive documentation.

Funding

This work was supported by the Autonomous Province of Trento through the Axonomix project (to FB, TT, PB and GV), and the Wellcome Trust (106098/Z/14/Z; to EJNG and THG).

Acknowledgements

We thank the Core Facility, Next Generation Sequencing Facility (HTS) CIBIO, University of Trento (Italy) for technical support.

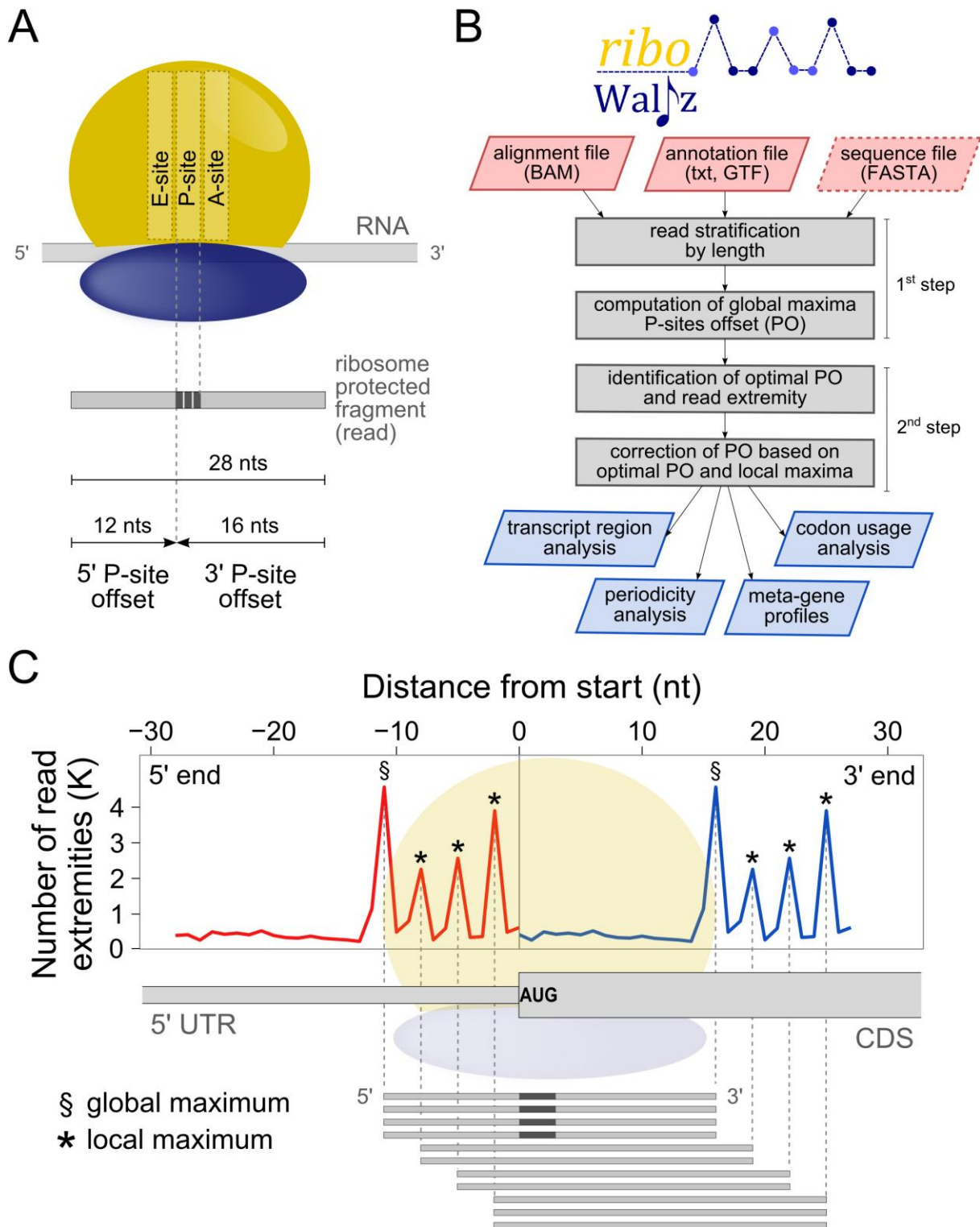


Figure 1. (A) Schematic representation of the P-site offset. Two offsets can be defined, one for each extremity of the read. (B) Flowchart representing the basic steps of riboWaltz, its inputs requirements and outputs. (C) An example of ribosome occupancy profile obtained from the alignment of the 5' and the 3' end of reads around the start codon (reads length, 28 nucleotides) is superimposed to the schematic representation of a transcript, a ribosome positioned on the TIS and a set of reads used for generating the profiles.

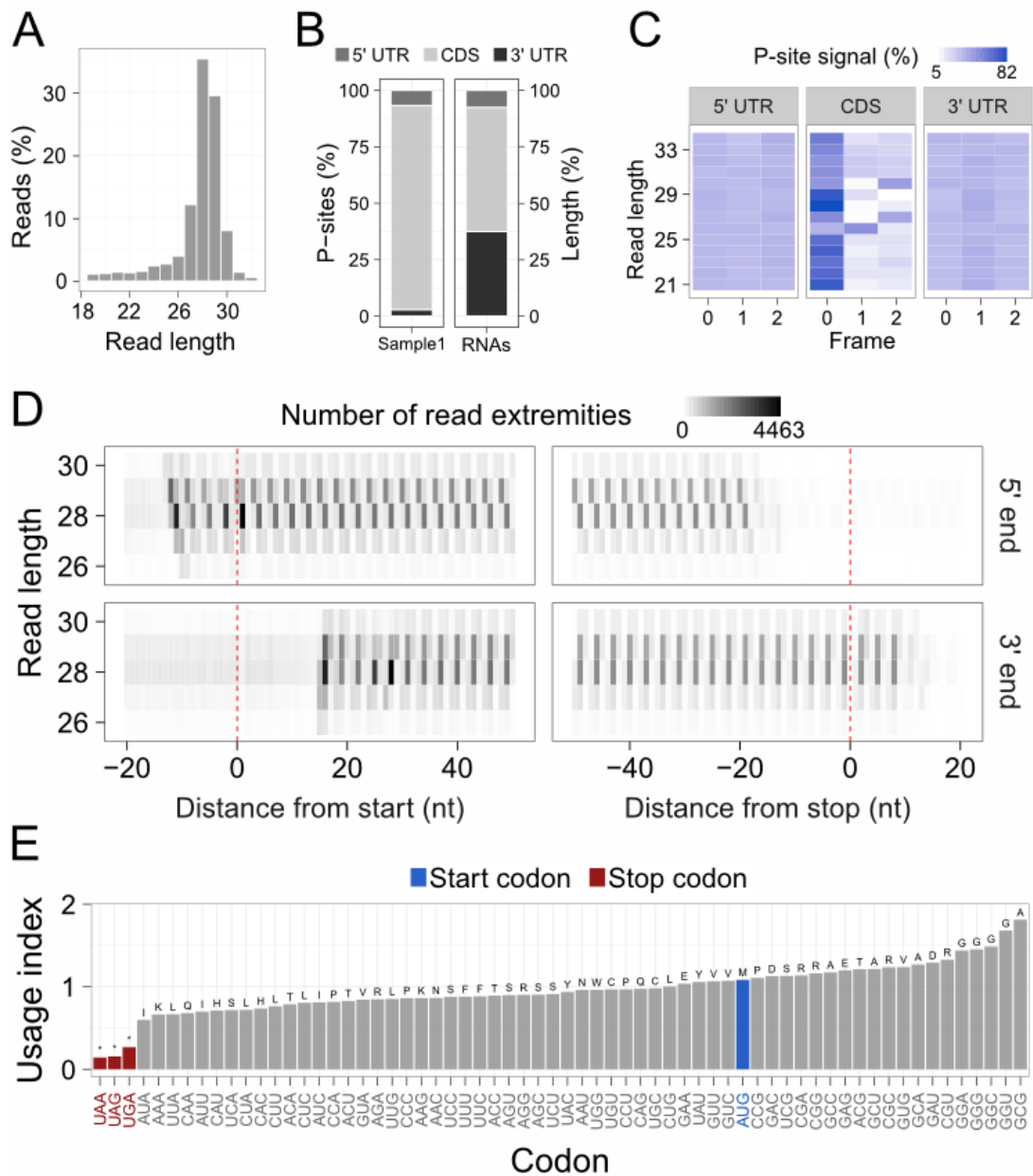


Figure 2. (A) Distribution of the read lengths. (B) Left, percentage of P-sites in the 5' UTR, CDS and 3' UTR of mRNAs from ribosome profiling data. Right, percentage of region lengths in mRNAs sequences. (C) Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR, stratified for read length. (D) Example of meta-gene heatmap reporting the signal associated to the 5' end (upper panel) and 3' end (lower panel) of the reads aligning around the start and the stop codon for different read lengths. (E) Codon usage analysis based on in-frame P-sites. Codon usage index is calculated as the frequency of in-frame P-sites along the coding sequence associated to each codon, normalized for codon frequency in sequences. Aminoacids corresponding to each codon are displayed above each bar. All panels were obtained from ribosome profiling of whole mouse brain.

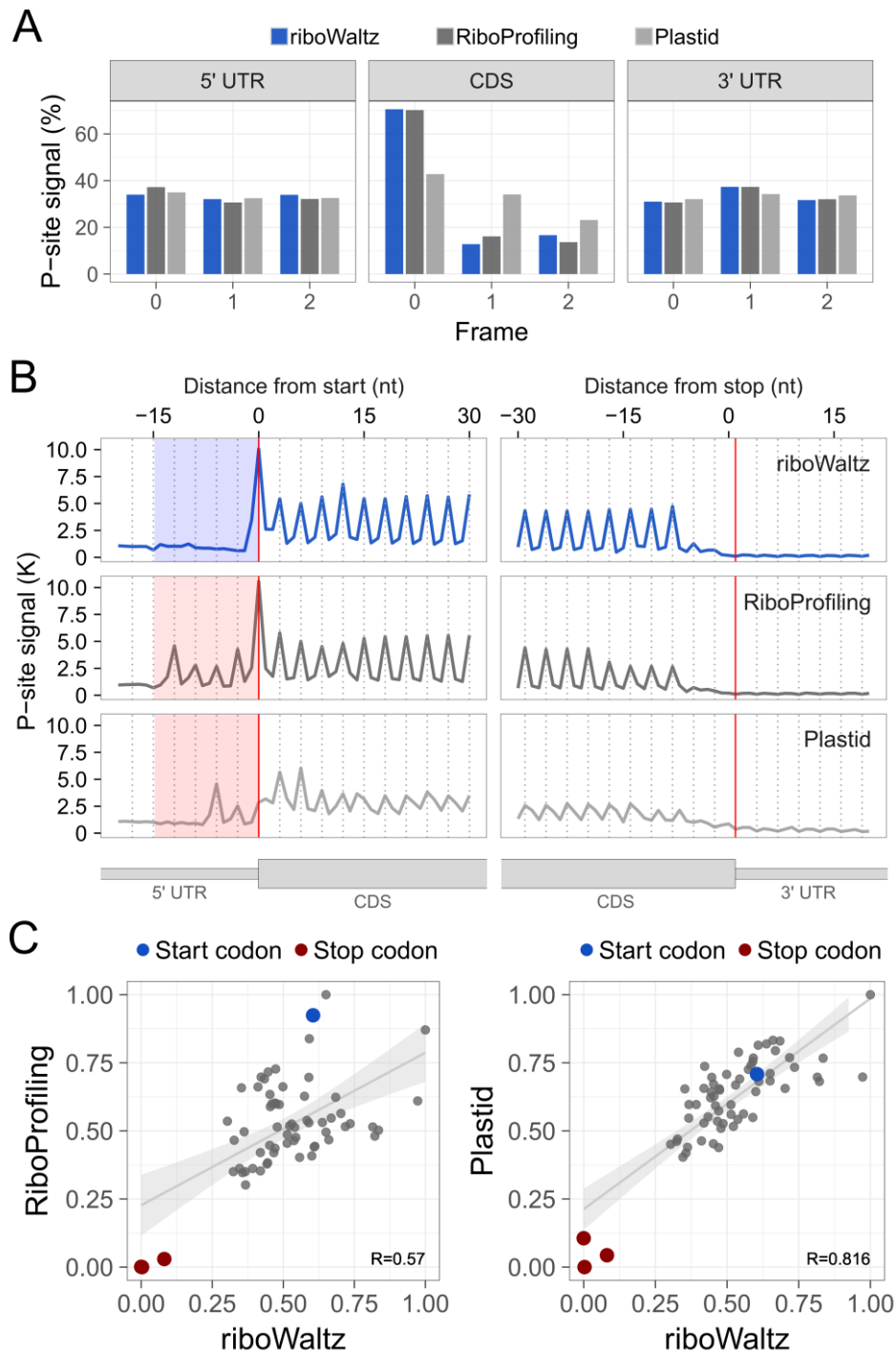


Figure 3. (A) Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR from ribosome profiling performed in mouse brain and (B) meta-profiles showing the periodicity of ribosomes along transcripts at genome-wide scale, based on P-site identification by riboWaltz, RiboProfiling and Plastid. The shaded areas to the left of the start codon highlight the shift of the periodicity toward the 5' UTR that is absent in the case of data analysed using riboWaltz. (C) Comparison between the codon usage index based on in-frame P-sites from riboWaltz and RiboProfiling (left panel) and between the codon usage index based on in-frame P-sites from riboWaltz and Plastid (right panel).

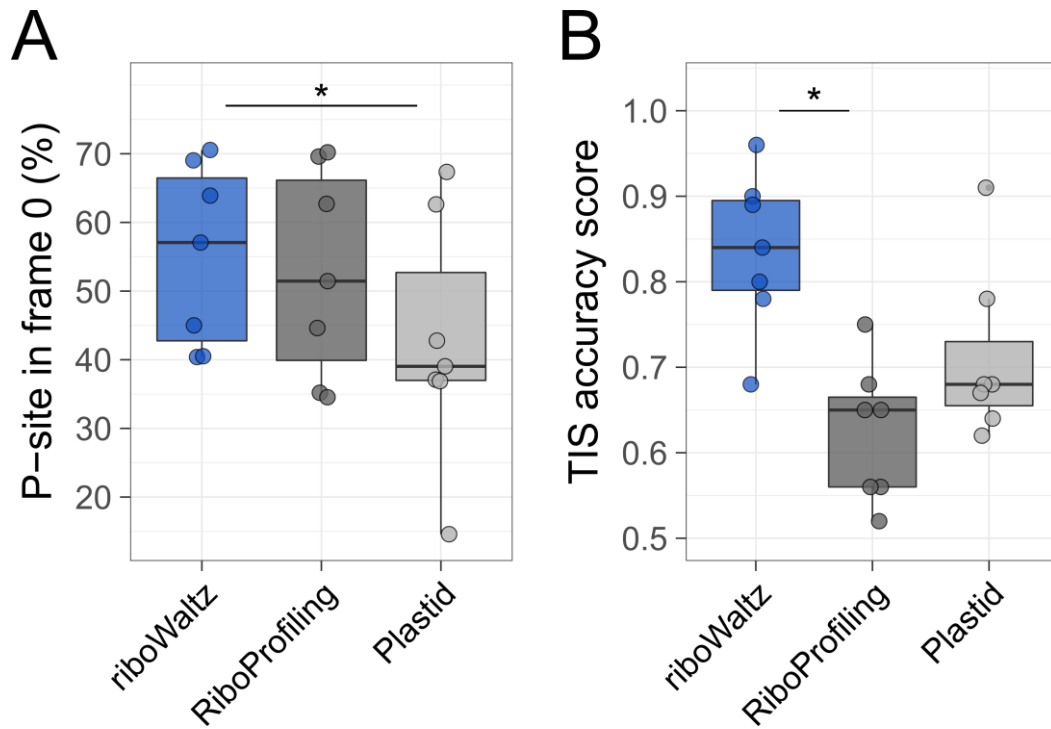


Figure 4. (A) Comparison of the percentage of P-sites in frame 0 along the coding sequence and (B) comparison of the average TIS accuracy score based on P-sites identification by riboWaltz, RiboProfiling and Plastid. Both panels display the results obtained from 7 datasets (2 yeast, 3 mouse and 2 human), each dataset represented by a dot. Statistical significances from Wilcoxon–Mann–Whitney test are shown: (* P-value < 0.05).

read length	riboWaltz		RiboProfiling		Plastid	
	from 5'	from 3'	from 5'	from 3'	from 5'	from 3'
19	2	16	2	16	13	5
20	4	15	4	15	13	6
21	4	16	4	16	13	7
22	5	16	5	16	13	8
23	6	16	6	16	13	9
24	7	16	7	16	13	10
25	8	16	1	25	13	11
26	10	15	10	15	13	12
27	10	16	10	16	13	13
28	11	16	1	28	5	22
29	12	16	12	16	13	15
30	12	17	10	19	35	6
31	13	17	20	50	13	17
32	15	16	15	16	13	18
33	16	16	17	15	13	19
34	17	16	17	16	13	20
35	18	16	18	16	13	21
36	16	19	19	16	13	22
37	20	16	22	58	13	23
38	21	16	15	22	13	24

Table 1: Comparison of the P-site offsets identified for each read length by riboWaltz, RiboProfiling and Plastid in mouse (GSE102318). The PO computed from both read extremities are reported.

References

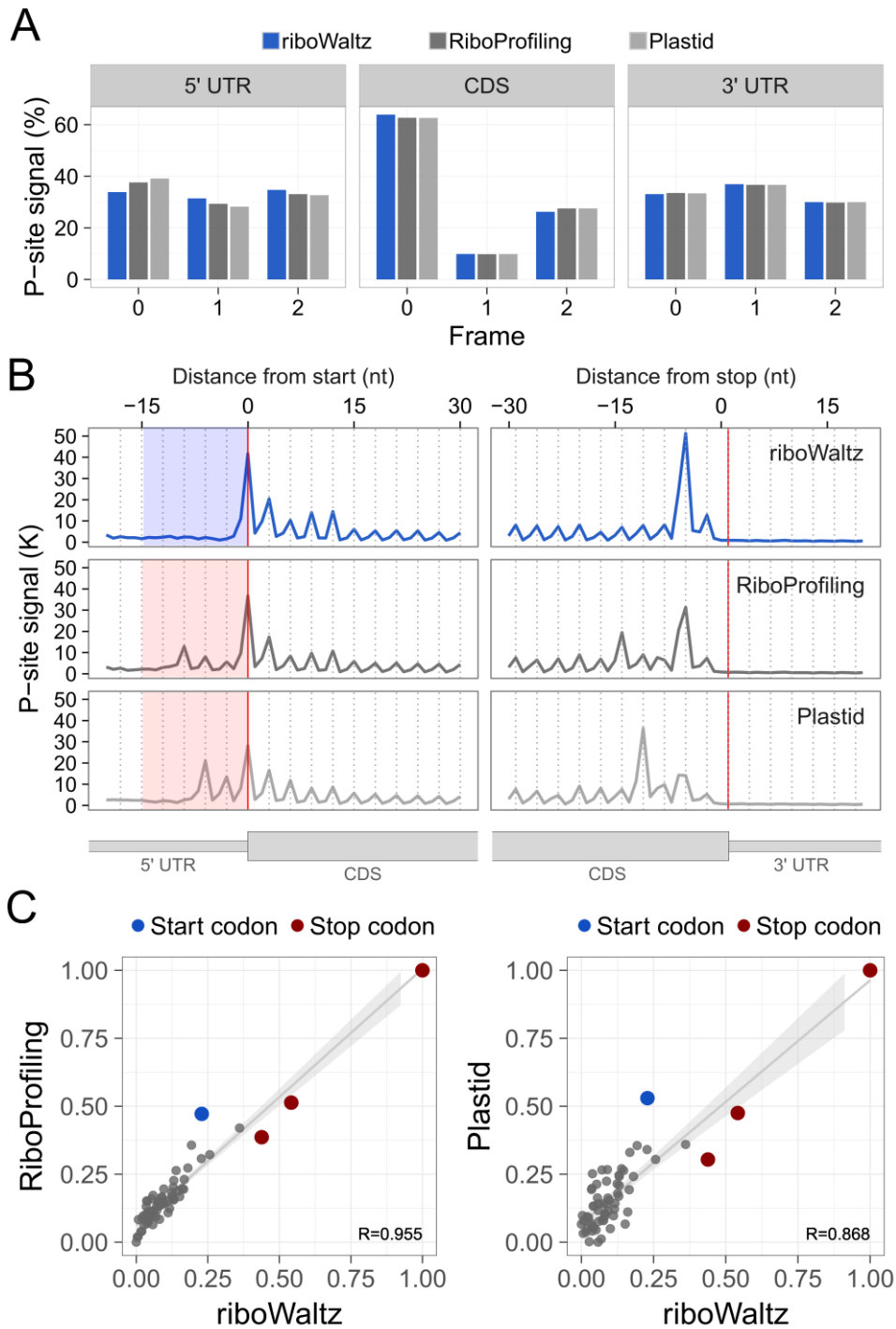
- Beaupere, C., Wasko, B. M., Lorusso, J., Kennedy, B. K., Kaeberlein, M., & Labunsky, V. M. (2017). CAN1 Arginine Permease Deficiency Extends Yeast Replicative Lifespan via Translational Activation of Stress Response Genes. *Cell reports*, 18(8), 1884-1892.
- Bazzini, A. A., Johnstone, T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., ... & Giraldez, A. J. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal*, e201488411.
- Brar, G. A., & Weissman, J. S. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology*.
- Budkevich, T. V., Giesebrecht, J., Behrmann, E., Loerke, J., Ramrath, D. J., Mielke, T., ... & Sanbonmatsu, K. Y. (2014). Regulation of the mammalian elongation cycle by subunit rolling: a eukaryotic-specific ribosome rearrangement. *Cell*, 158(1), 121-131.
- Chen, J., Tsai, A., O'Leary, S. E., Petrov, A., & Puglisi, J. D. (2012). Unraveling the dynamics of ribosome translocation. *Current opinion in structural biology*, 22(6), 804-814.
- Dunn, J. G., & Weissman, J. S. (2016). Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC genomics*, 17(1), 958.
- Gao, X., Wan, J., Liu, B., Ma, M., Shen, B., & Qian, S. B. (2015). Quantitative profiling of initiating ribosomes in vivo. *Nature methods*, 12(2), 147-153.
- Gritsenko, A. A., Hulsman, M., Reinders, M. J., & de Ridder, D. (2015). Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. *PLoS Comput Biol*, 11(8), e1004336.
- Guo, H., Ingolia, N. T., Weissman, J. S., & Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308), 835-840.
- Han, Y., Gao, X., Liu, B., Wan, J., Zhang, X., & Qian, S. B. (2014). Ribosome profiling reveals sequence-independent post-initiation pausing as a signature of translation. *Cell research*, 24(7), 842-851.
- Hsu, P. Y., Calviello, L., Wu, H. Y. L., Li, F. W., Rothfels, C. J., Ohler, U., & Benfey, P. N. (2016). Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. *Proceedings of the National Academy of Sciences*, 113(45), E7126-E7135.
- Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M., & Weissman, J. S. (2012). The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature protocols*, 7(8), 1534-1550.
- Ingolia, N. T., Ghaemmghami, S., Newman, J. R., & Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924), 218-223.
- Kochetov, A. V., Allmer, J., Klimenko, A. I., Zuraev, B. S., Matushkin, Y. G., & Lashin, S. A. (2016). AltORFev facilitates the prediction of alternative open reading frames in eukaryotic mRNAs. *Bioinformatics*, btw736.

- Legendre, R., Baudin-Baillieu, A., Hatin, I., & Namy, O. (2015). RiboTools: a Galaxy toolbox for qualitative ribosome profiling analysis. *Bioinformatics*, 31(15), 2586-2588.
- Lareau, L. F., Hite, D. H., Hogan, G. J., & Brown, P. O. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife*, 3, e01257.
- Martens, A. T., Taylor, J., & Hilser, V. J. (2015). Ribosome A and P sites revealed by length analysis of ribosome profiling data. *Nucleic acids research*, gkv200
- Michel, A. M., & Baranov, P. V. (2013). Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdisciplinary Reviews: RNA*, 4(5), 473-490.
- Michel, A. M., Andreev, D. E., & Baranov, P. V. (2014). Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. *BMC bioinformatics*, 15(1), 380.
- Pop, C., Rouskin, S., Ingolia, N. T., Han, L., Phizicky, E. M., Weissman, J. S., & Koller, D. (2014). Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Molecular systems biology*, 10(12), 770.
- Popa, A., Lebrigand, K., Paquet, A., Nottet, N., Robbe-Sermesant, K., Waldmann, R., & Barbry, P. (2016). RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Research*, 5.
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842.
- Raj, A., Wang, S. H., Shim, H., Harpak, A., Li, Y. I., Engelmann, B., ... & Pritchard, J. K. (2016). Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife*, 5, e13328.
- Shi, Z., Fujii, K., Kovary, K. M., Genuth, N. R., Röst, H. L., Teruel, M. N., & Barna, M. (2017). Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. *Molecular Cell*.
- Steitz, J. A. (1969). Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature*, 224, 957-964.
- Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., & Bartel, D. P. (2016). Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. *Cell reports*, 14(7), 1787-1799.
- Wolin, S. L., & Walter, P. (1988). Ribosome pausing and stacking during translation of a eukaryotic mRNA. *The EMBO journal*, 7(11), 3559.
- Xiao, Z., Zou, Q., Liu, Y., & Yang, X. (2016). Genome-wide assessment of differential translations with ribosome profiling data. *Nature communications*, 7.
- Zhong, Y., Karaletsos, T., Drewe, P., Sreedharan, V. T., Kuo, D., Singh, K., ... & Räscht, G. (2017). RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics*, 33(1), 139-141.

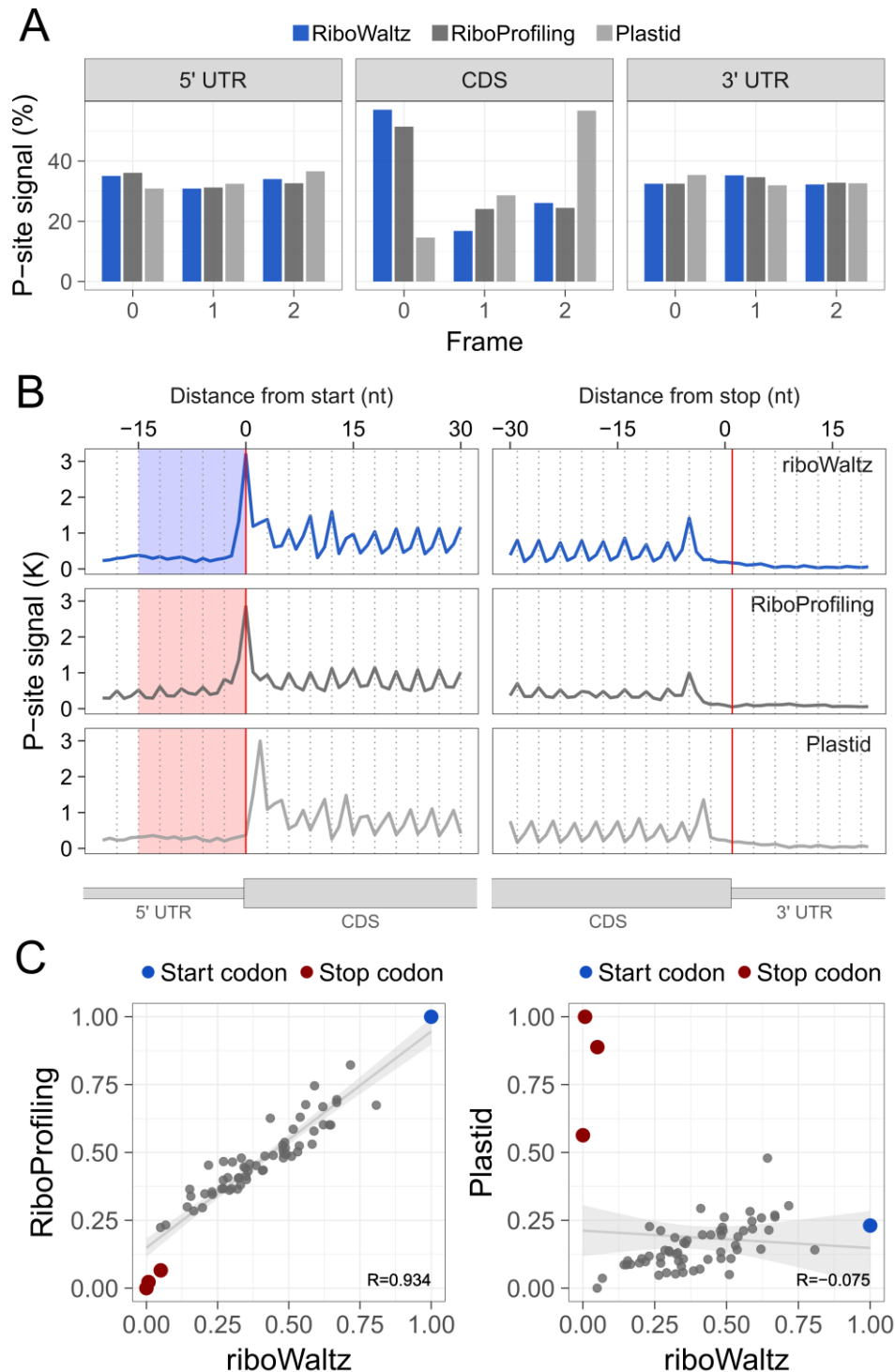
**riboWaltz: optimization of ribosome P-site positioning in ribosome profiling
data**

Fabio Lauria^{1§*}, Toma Tebaldi^{2§}, Paola Bernabò¹, Ewout J.N. Groen³, Thomas H. Gillingwater³, Gabriella Viero^{1*}

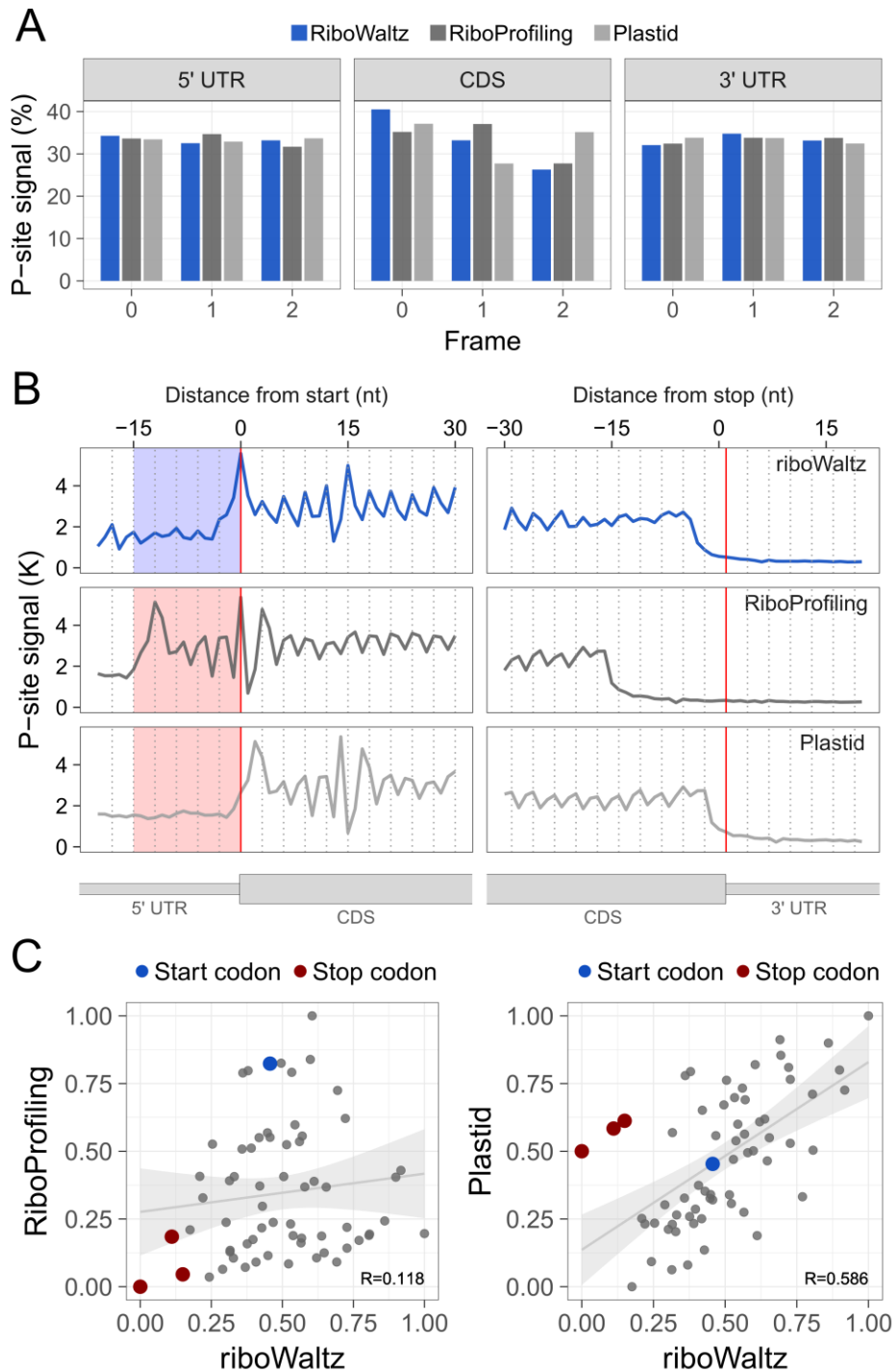
Supplementary Information



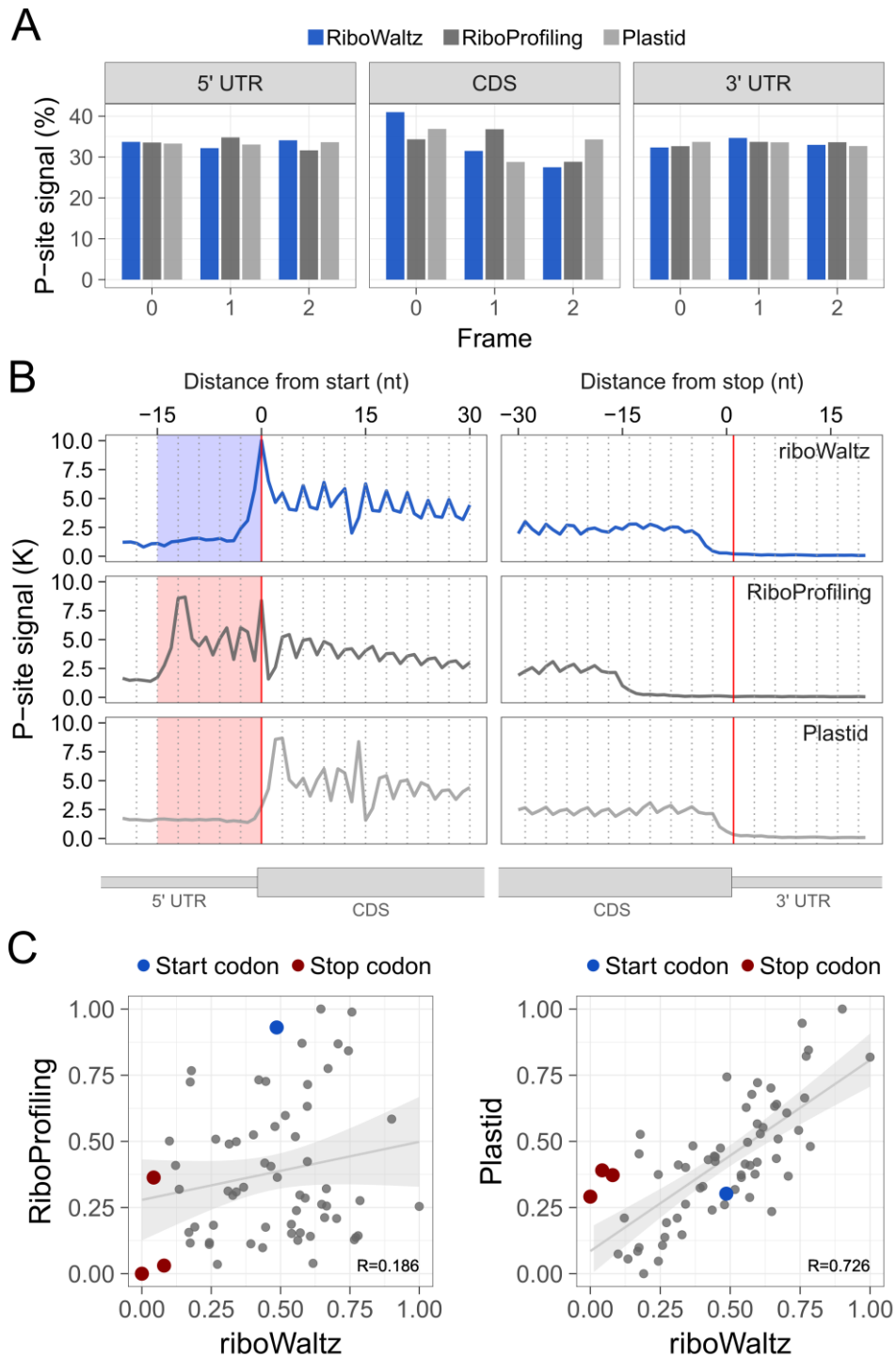
Supplementary Figure 1. (A) Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR from ribosome profiling in Hek-293 (Gao et al., 2015) and **(B)** meta-profiles showing the periodicity of ribosomes along transcripts at genome-wide scale, based on P-site identification by riboWaltz, RiboProfiling and Plastid. The shaded areas to the left of the start codon highlight the shift of the periodicity toward the 5' UTR. **(C)** Comparison between the codon usage index based on in-frame P-sites from riboWaltz and RiboProfiling (left panel) and between the codon usage index based on in-frame P-sites from riboWaltz and Plastid (right panel).



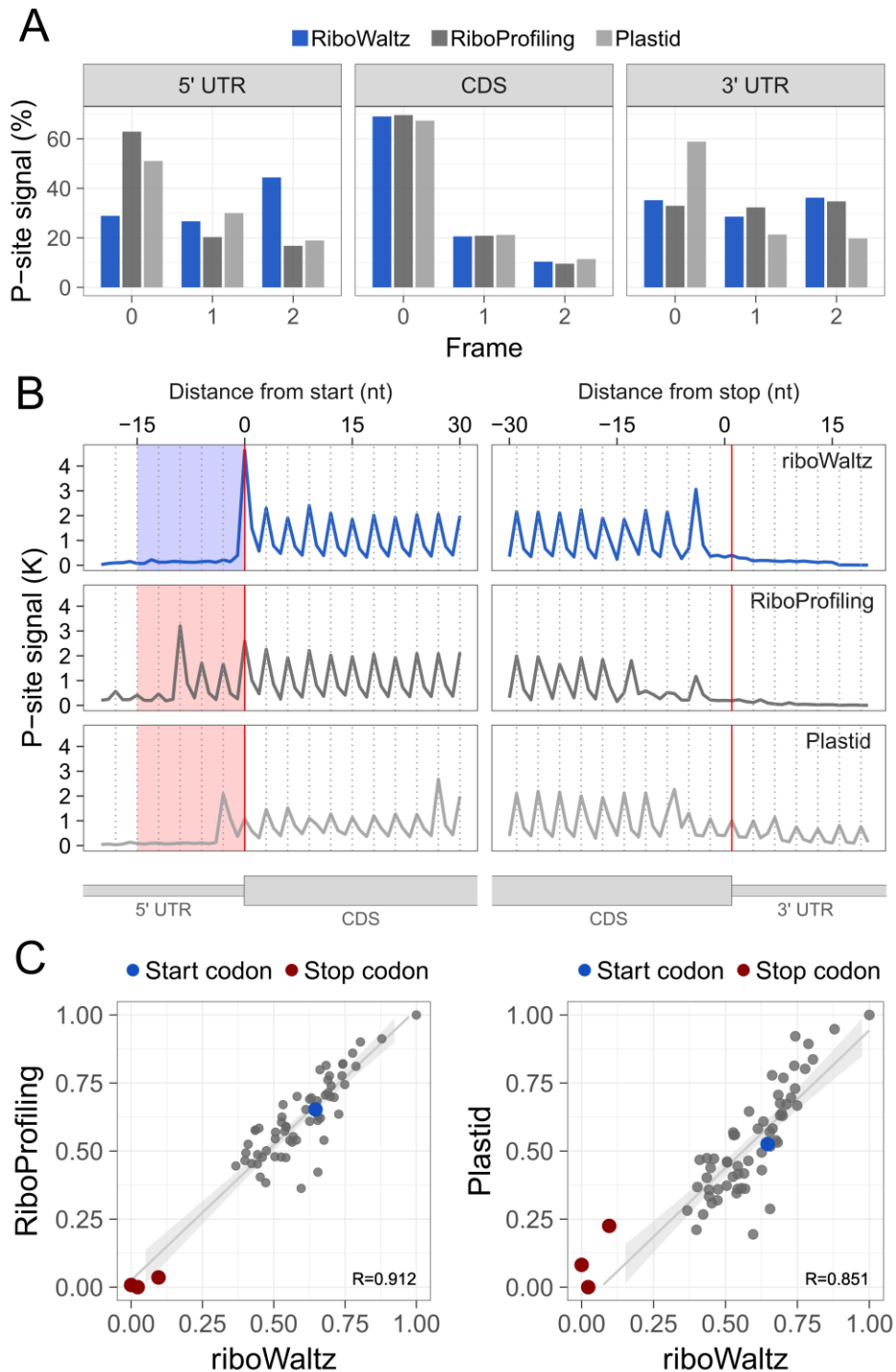
Supplementary Figure 2. (A) Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR from ribosome profiling in MCF-7 (unpublished data) and **(B)** meta-profiles showing the periodicity of ribosomes along transcripts at genome-wide scale, based on P-site identification by riboWaltz, RiboProfiling and Plastid. The shaded areas to the left of the start codon highlight the shift of the periodicity toward the 5' UTR. **(C)** Comparison between the codon usage index based on in-frame P-sites from riboWaltz and RiboProfiling (left panel) and between the codon usage index based on in-frame P-sites from riboWaltz and Plastid (right panel).



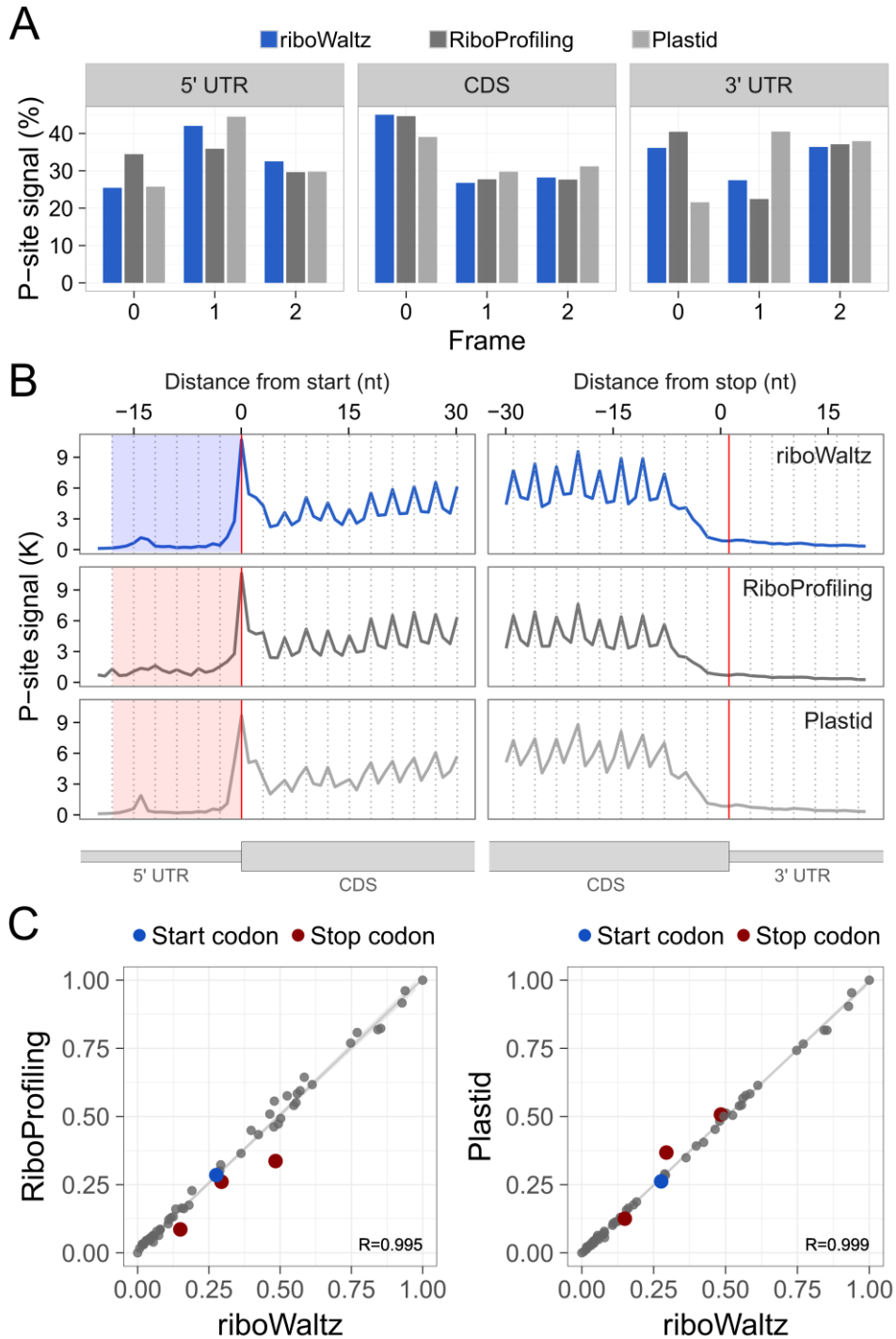
Supplementary Figure 3. (A) Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR from ribosome profiling in mouse after pull-down of RPL10 (Shi et al. 2017) and **(B)** meta-profiles showing the periodicity of ribosomes along transcripts at genome-wide scale, based on P-site identification by riboWaltz, RiboProfiling and Plastid. The shaded areas to the left of the start codon highlight the shift of the periodicity toward the 5' UTR. **(C)** Comparison between the codon usage index based on in-frame P-sites from riboWaltz and RiboProfiling (left panel) and between the codon usage index based on in-frame P-sites from riboWaltz and Plastid (right panel).



Supplementary Figure 4. (A) Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR from ribosome profiling in mouse after pull-down of RPL22 (Shi et al. 2017) and **(B)** meta-profiles showing the periodicity of ribosomes along transcripts at genome-wide scale, based on P-site identification by riboWaltz, RiboProfiling and Plastid. The shaded areas to the left of the start codon highlight the shift of the periodicity toward the 5' UTR. **(C)** Comparison between the codon usage index based on in-frame P-sites from riboWaltz and RiboProfiling (left panel) and between the codon usage index based on in-frame P-sites from riboWaltz and Plastid (right panel).



Supplementary Figure 5. (A) Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR from ribosome profiling in yeast (Beaupere et al., 2017) and **(B)** meta-profiles showing the periodicity of ribosomes along transcripts at genome-wide scale, based on P-site identification by riboWaltz, RiboProfiling and Plastid. The shaded areas to the left of the start codon highlight the shift of the periodicity toward the 5' UTR. **(C)** Comparison between the codon usage index based on in-frame P-sites from riboWaltz and RiboProfiling (left panel) and between the codon usage index based on in-frame P-sites from riboWaltz and Plastid (right panel).



Supplementary Figure 6. (A) Percentage of P-sites in the three frames along the 5' UTR, CDS and 3' UTR from ribosome profiling in yeast (Lareau et al., 2014) and (B) meta-profiles showing the periodicity of ribosomes along transcripts at genome-wide scale, based on P-site identification by riboWaltz, RiboProfiling and Plastid. The shaded areas to the left of the start codon highlight the shift of the periodicity toward the 5' UTR. (C) Comparison between the codon usage index based on in-frame P-sites from riboWaltz and RiboProfiling (left panel) and between the codon usage index based on in-frame P-sites from riboWaltz and Plastid (right panel).

read length	riboWaltz		RiboProfiling		Plastid	
	from 5'	from 3'	from 5'	from 3'	from 5'	from 3'
25	12	12	0	24	3	21
26	12	13	12	13	12	13
27	12	14	12	14	12	14
28	12	15	3	24	6	21
29	12	16	12	16	6	22
30	12	17	12	17	12	17
31	13	17	9	21	9	21
32	13	18	10	21	7	24
33	12	20	12	20	50	18
34	15	18	17	16	13	20

Supplementary Table 1: Comparison of the P-site offsets identified for each read length by riboWaltz, RiboProfiling and Plastid in human (Hek-293, Gao et al., 2015). The PO computed from both read extremities are reported.

read length	riboWaltz		RiboProfiling		Plastid	
	from 5'	from 3'	from 5'	from 3'	from 5'	from 3'
20	11	8	-1	20	13	6
21	11	9	3	17	13	7
22	11	10	4	17	13	8
23	11	11	23	-1	13	9
24	11	12	6	17	13	10
25	11	13	25	-1	13	11
26	11	14	9	16	13	12
27	11	15	9	17	13	13
28	11	16	-13	40	13	14
29	11	17	11	17	13	15
30	11	18	11	18	13	16
31	12	18	12	18	13	17
32	12	19	12	19	13	18
33	12	20	-10	42	13	19
34	11	22	17	16	13	20
35	10	24	4	30	13	21
36	12	23	12	23	13	22
37	10	26	35	1	13	23
38	12	25	-7	44	13	24
39	10	28	20	18	13	25
41	23	17	-14	54	13	27
42	17	24	37	4	13	28
43	11	31	0	42	13	29
45	14	30	48	-4	13	31

Supplementary Table 2: Comparison of the P-site offsets identified for each read length by riboWaltz, RiboProfiling and Plastid in human (MCF-7, unpublished data). The PO computed from both read extremities are reported.

read length	riboWaltz		RiboProfiling		Plastid	
	from 5'	from 3'	from 5'	from 3'	from 5'	from 3'
19	10	8	-1	19	13	5
20	11	8	-1	20	13	6
21	11	9	-1	21	13	7
22	10	11	-1	22	13	8
23	11	11	-1	23	13	9
24	11	12	-1	24	13	10
25	10	14	-1	25	13	11
26	11	14	-1	26	13	12
27	11	15	-1	27	13	13
28	10	17	-1	28	13	14
29	11	17	-1	29	13	15
30	11	18	-1	30	13	16
31	10	20	-1	31	13	17
32	11	20	-1	32	13	18
33	12	20	-1	33	13	19
34	10	23	-1	34	13	20
35	11	23	-1	35	13	21
36	12	23	-1	36	13	22
37	11	25	-1	37	13	23
38	11	26	-1	38	13	24
39	10	28	-1	39	13	25
40	11	28	-1	40	13	26
41	13	27	-1	41	13	27
42	11	30	-1	42	13	28
43	14	28	-1	43	13	29
44	11	32	-1	44	13	30
45	12	32	-1	45	13	31
46	13	32	-1	46	13	32
47	11	35	-1	47	13	33
48	11	36	-1	48	13	34
49	12	36	-1	49	13	35
50	12	37	-1	50	13	36

Supplementary Table 3: Comparison of the P-site offsets identified for each read length by riboWaltz, RiboProfiling and Plastid in mouse (after pull-down of RLP10, Shi et al. 2017). The PO computed from both read extremities are reported.

read length	riboWaltz		RiboProfiling		Plastid	
	from 5'	from 3'	from 5'	from 3'	from 5'	from 3'
19	12	6	-1	19	13	5
20	11	8	-1	20	13	6
21	12	8	-1	21	13	7
22	11	10	-1	22	13	8
23	10	12	-1	23	13	9
24	9	14	-1	24	13	10
25	10	14	-1	25	13	11
26	10	15	-1	26	13	12
27	11	15	-1	27	13	13
28	10	17	-1	28	13	14
29	11	17	-1	29	13	15
30	11	18	-1	30	13	16
31	10	20	-1	31	13	17
32	11	20	-1	32	13	18
33	12	20	-1	33	13	19
34	10	23	-1	34	13	20
35	10	24	-1	35	13	21
36	10	25	-1	36	13	22
37	10	26	-1	37	13	23
38	10	27	-1	38	13	24
39	11	27	-1	39	13	25
40	10	29	-1	40	13	26
41	11	29	-1	41	13	27
42	11	30	-1	42	13	28
43	7	35	-1	43	13	29
44	10	33	-1	44	13	30
45	16	28	-1	45	13	31
46	11	34	-1	46	13	32
47	11	35	-1	47	13	33
48	11	36	-1	48	13	34
49	11	37	-1	49	13	35
50	11	38	-1	50	13	36

Supplementary Table 4: Comparison of the P-site offsets identified for each read length by riboWaltz, RiboProfiling and Plastid in mouse (after pull-down of RLP22, Shi et al. 2017). The PO computed from both read extremities are reported.

read length	riboWaltz		RiboProfiling		Plastid	
	from 5'	from 3'	from 5'	from 3'	from 5'	from 3'
20	11	8	-23	42	13	6
21	8	12	-10	30	13	7
22	11	10	19	2	13	8
23	7	15	-29	51	13	9
24	8	15	-10	33	13	10
25	9	15	18	6	13	11
26	10	15	-17	42	13	12
27	11	15	2	24	38	-12
28	12	15	3	24	9	18
29	13	15	13	15	10	18
30	13	16	-8	37	24	5
31	15	15	-22	52	13	17
32	16	15	-27	58	13	18
33	14	18	11	21	13	19
34	18	15	-19	52	13	20
35	16	18	-47	81	13	21
37	12	24	-34	70	13	23
38	20	17	-24	61	13	24
40	22	17	20	19	13	26
41	15	25	27	13	13	27
42	23	18	-1	42	13	28
43	23	19	-31	73	13	29
44	21	22	6	37	13	30
46	30	15	-15	60	13	32

Supplementary Table 5: Comparison of the P-site offsets identified for each read length by riboWaltz, RiboProfiling and Plastid in yeast (Beaupere et al., 2017). The PO computed from both read extremities are reported.

read length	riboWaltz		RiboProfiling		Plastid	
	from 5'	from 3'	from 5'	from 3'	from 5'	from 3'
21	12	8	12	8	12	8
22	13	8	50	71	13	8
23	13	9	2	20	13	9
24	13	10	22	45	13	10
25	13	11	9	15	13	11
26	12	13	44	69	13	12
27	13	13	10	36	13	13
28	12	15	12	15	12	15
29	13	15	13	15	12	16
30	12	17	12	17	12	17
31	13	17	13	17	13	17
32	14	17	14	17	13	18
33	14	18	43	75	13	19
34	15	18	3	36	13	20
35	10	24	5	39	13	21
36	13	22	11	24	13	22
37	15	21	12	48	13	23
38	14	23	23	60	13	24
39	22	16	12	26	13	25
40	7	32	7	32	13	26

Supplementary Table 6: Comparison of the P-site offsets identified for each read length by riboWaltz, RiboProfiling and Plastid in yeast (Lareau et al., 2014). The PO computed from both read extremities are reported.

Supplementary methods

RiboSeq data processing

Raw reads were processed by removing 5' adapters, discarding reads shorter than 20 nucleotides and trimming the first nucleotide (using Trimmomatic v0.36). Reads mapping on rRNAs and tRNAs (downloaded from the SILVA rRNA and the Genomic tRNA databases respectively) were removed. The remaining reads were aligned to the organism transcriptome with Bowtie2 (v2.2.6) employing the default settings. All reads aligning to the very same region were collapsed to avoid potential PCR duplicates, and only strand-specific reads were kept.

3.2 riboScan

As mentioned in the introduction to riboWaltz, the development of many pipelines dealing with ribosome profiling (RiboSeq) data, typically aimed at extracting typically translation efficiencies from entire transcripts performing differential expression analysis^{163,172,173}, identifying new open reading frames^{175,178,179,264} and more rarely for obtaining positional information describing fluxes of ribosomes along the RNA^{140,141}. Nevertheless, some aspects such as statistical procedures for the extraction of meaningful positional information still need to be computationally addressed.

It's known that high signal from reads obtained from ribosome protected fragments can be related to ribosome slowdown^{82,253} and ribosome stalling^{97,254}, two scenarios connected to many pathologies such as neurodegenerative diseases^{47,255}, diabetes and multi-systemic failure²⁵⁶. This highlights the importance in identifying mRNA regions showing significant signal enrichments (a problem more generally known as peak calling) along ribosome occupancy profiles to achieve a better understanding of translation regulation through the characterization of polysome organization. This issue has not yet been exhaustively examined. In fact, peak calling has been initially employed for the analysis of other positional data, such as ChIP-Seq data, then further extended to the analysis of RNA-RNA Binding Protein (RBP) interaction by CLIP-Seq and related approaches²⁶⁵⁻²⁶⁸. Typically, interaction profiles between RNA and RBPs are characterised by well distinct signal peaks surrounded by regions with no signal. These regions specifically identify the neat binding sites of RBPs. This type of clearness is not present in ribosome profiles, which are characterized by a noisy and continuous signal with fluctuations along the CDSs of translated transcripts. The average signal along the mRNA in ribosome profiling depends on the translation levels of the transcript, and can span many orders of magnitude. This difference with CLIP-seq data represents a challenge in the direct application of existing peak-calling algorithms to RiboSeq data. The only attempt of extracting positional information by individuating enriched regions along ribosome occupancy profiles has been proposed last year by Diamant and Tuller¹⁹⁶, and is based on transcript-specific standardization of the data. Despite the advantages of this approach, an accurate procedure for defining statistically significant ribosome peak has not been discussed. Thus, after extensive revision of available protocols, I decided to develop a dedicated approach, riboScan.

The flowchart of riboScan is shown in Figure 3.1 and a detailed description of whole procedure is reported in the following sections. After the acquisition of the input files a step of filtering and normalization of the data is performed. Then, a table reporting the

coverage of every codon of the transcriptome is generated and the transcripts are grouped (stratified) depending on their expression level. Two distributions based on the codon coverage are fitted, a pair of codon-specific p-values is computed and then blended in a single p-value by the Fisher's combined probability test. Statistically significant enriched regions are identified and eventually aggregated.

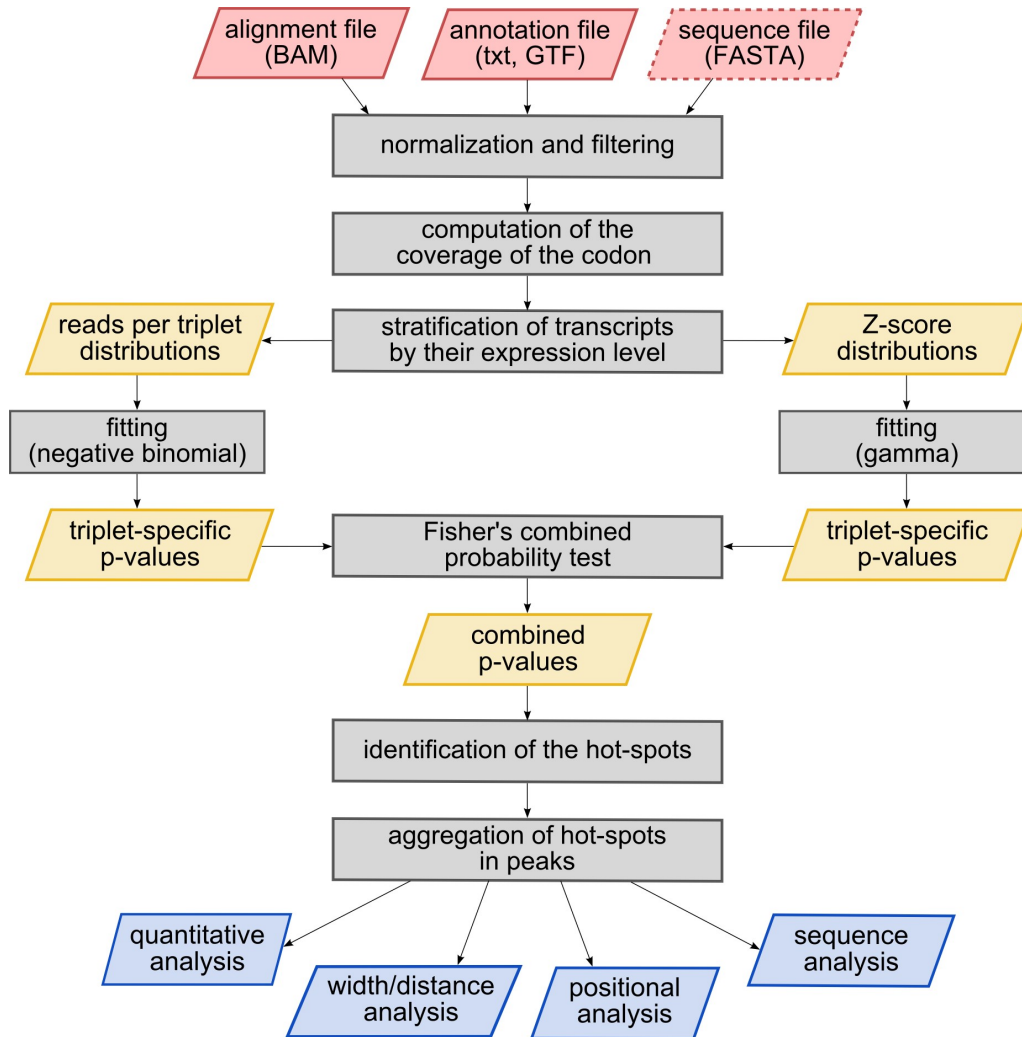


Figure 3.1. **riboScan**. Flowchart representing the basic steps of riboScan, its inputs requirements and outputs. The steps of the pipeline are represented by gray rectangles. The red parallelograms specify the mandatory (continuous perimeter) and optional (dotted perimeter) input files while the yellow parallelograms indicate the output of the pipeline employed as input data in the subsequent step. Finally, the blue parallelograms indicate some analyses based on the identification of enriched regions along the mRNA.

riboScan has been developed and tested on Poly-RiboSeq data from healthy mouse brains produced in my Lab (Laboratory of Translational Architectomics, IBF-CNR,

Trento). For details about the Poly-RiboSeq protocol, the pre-processing of the data and the alignment steps, please refer to the Appendix at the end of the elaborate.

3.2.1 Materials and methods

Filtering and normalization

Ribosome profiling data were processed by applying three steps of filtering and normalization. Firstly, only the protein coding transcripts with coding sequence length divisible by 3 and both 3' and 5' UTRs longer than 3 nucleotides were kept for further analyses. These restrictions guarantee the analysis of ribosome protected fragments from mRNAs with CDS that can be divided into codons and annotated UTRs. Secondly, the trimmed mean of M-values normalization method (TMM) was applied to remove possible size or compositional difference between libraries coming from multiple conditions and replicates. Thirdly, transcript-specific FPKMs (fragments per kilobase per million fragments mapped) were computed. All mRNAs with average FPKM values under the 80th percentile were discarded. Finally, a similar filtering approach, based on CPM values (counts of fragments per million fragments mapped), was applied and only mRNAs with a CPM (count per million) mean above the 80th percentile were kept. This last step ensures to work with transcripts with a number of mapped reads (transcript coverage) sufficient for further analysis.

Triplet coverage

A table containing the coverage of every codon (codon coverage) of the transcriptome was generated. The sequence of every transcript was divided in triplets starting from the annotated Translation Initiation Site (TIS) proceeding towards the UTRs extremities, and eventually discarding the exceeding 1 or 2 nucleotides at the extremities of the transcript. Codon-specific coverage was determined using the full coverage of ribosome protected fragments. Acting, *de facto*, as a smoothing factor, this choice enables an improved detection of peaks.

Hot-spots detection

To get positional information about regions with accumulation of ribosomes, that we called "hot-spots", I developed a dedicated pipeline.

Firstly, transcripts were stratified in 20 bins based on their mean FPKM (fragments per kilobase per million fragments mapped) values. This step was necessary to avoid potential biases due to large differences in the mean transcript coverage. Secondly, I used in parallel two different strata-specific methods aimed at assigning coverage p-

values to each triplet, where the null hypothesis states in both cases that there is no accumulation of reads on the triplet. Their outputs are finally combined using the Fisher's method. This double-check procedure may identify consistent and robust ribosome profiling peaks.

The first method takes the cue from a CLIP- and RIP-Seq peak-caller algorithm named Piranha²⁶⁶. This approach relies on dividing the transcripts in consecutive regions of a specified length and on fitting the distribution of the number of reads per region by a negative binomial probability density function (Figure 3.2A). This distribution, previously shown to be the best choice for fitting CLIP data²⁶⁶ and widely used in all NGS data analyses, was employed to compute the first set of codon-specific p-values. Note that transcripts with a small number of mapped reads are penalised by this procedure, since the coverage of all their codons is generally low and, as a consequence, the associated p-value is not significant.

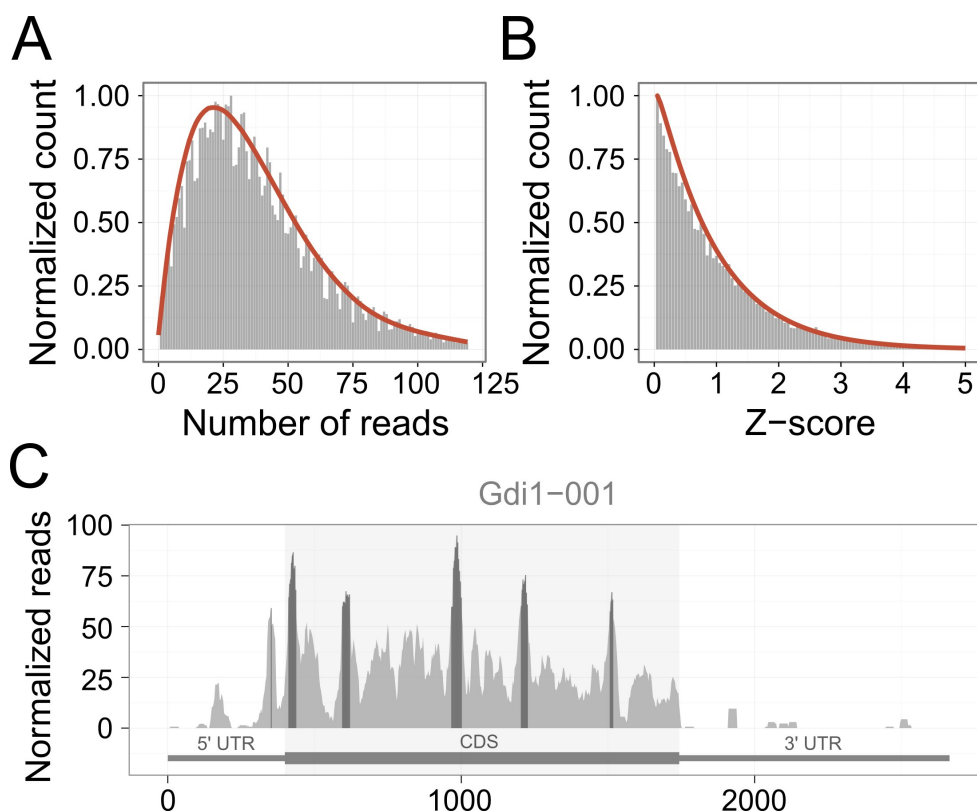


Figure 3.2. **Examples of riboScan fitting procedures and outcome.** (A) Empirical distribution of the number of reads per triplet (determined for transcripts in the 10th expression bin). The red curve represents the negative binomial fit of the empirical distribution. (B) Empirical distribution of z-scores per triplet (determined for transcripts in the 10th expression bin). The red curve represents the gamma fit. (C) Example of single-transcript RiboSeq density profile. The dark areas highlight the hot-spots detected by riboScan. All the plots in this panel were generated using ribosome profiling data performed on healthy mouse brains.

The second method relies on calculating the z-scores of the number of reads per triplet obtained after a transcript-specific standardization of the data, similarly to what was proposed by Diamant and Tuller¹⁹⁶. Differently from the previous method, this procedure does not penalize low coverage transcripts, minimizing the negative effects of divergent mRNA coverages within the bins. The z-score distribution was initially fitted to a lognormal or a gamma density function, that are continuous and asymmetric distributions suitable for the approximation of these experimental data. I then tested the goodness of the fits by the Akaike²⁶⁹ and Bayesian²⁷⁰ information criteria. The gamma distribution was selected as the best fitting curve (Figure 3.2B) and the second set of codon-specific p-values was computed accordingly.

Finally, the two sets of p-values calculated for each triplet were combined applying the Fisher's combined probability test²⁷¹. All triplets with at least 5 reads (minimum coverage threshold) were associated to a Fisher's p-value. When this value was lower than 0.05 were the triplets were tagged as statistically significant enriched regions and named "hot-spots" henceforward.

In Figure 3.2C, a typical example of RiboSeq profile for a transcript (light gray) with the detected hot-spots (dark gray) is shown.

Definition of peaks

Typically, many adjacent hot-spots can be identified to concentrate in specific larger region of the transcript. Therefore, I decided to consider an additional characteristic that aggregate close hot-spots present in a region, and named it "peak". In other words, a peak identifies a region of 1 or more hot-spots significantly enriched in ribosome coverage (Figure 3.3).

Therefore, a peak consists of either a single hot-spot or many hot-spots closer than a threshold L , implying that the minimum distance between two peaks is exactly L codons. This definition allows to detect ribosome enriched regions of different sizes, that may be associated to either single or multiple ribosomes stuck along the transcripts. If not otherwise specified, I used a threshold value of 5 codons, i.e. approximately half a ribosome footprint.

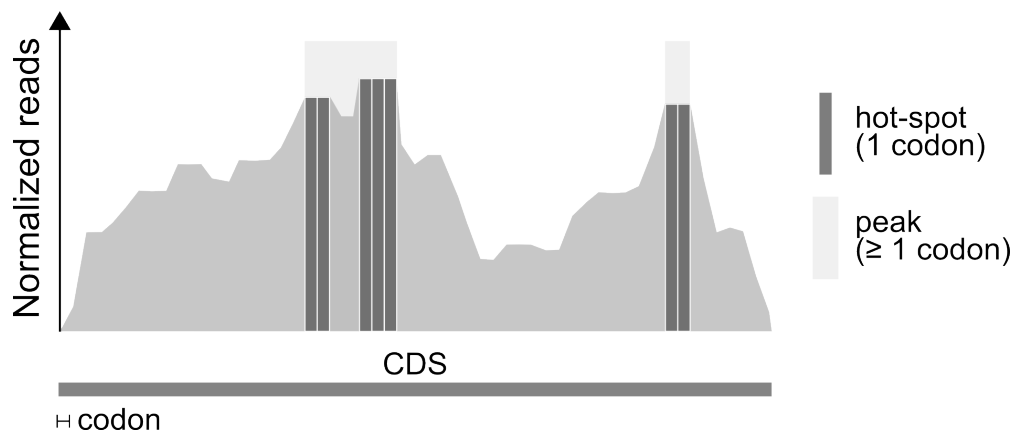


Figure 3.3. **Aggregation of hot-spots in peaks.** Example of ribosome occupancy profile containing 7 hot-spots (dark gray vertical bars) corresponding to 2 peaks (light gray areas). The dimension of a hot-spot is always of 1 codon, the width of a peak varies depending on the number of hot-spots it includes and on the spacing between them.

3.2.2 Conclusions

In conclusion, here I propose for the first time a pipeline dedicated to the extraction of meaningful positional information from ribosome profiling data at codon resolution. riboScan, based on the detection of statistically significant enriched regions within ribosome occupancy profiles, assist the identification of ribosome accumulations along the mRNAs. Consequently, riboScan may lay the groundwork for a better understanding of polysome organizational rules governing the number and the localization of ribosomes as well as their aggregation in clusters along the transcripts.

In the present work I apply riboScan to ribosome profiling data from healthy and Spinal Muscular Atrophy-affected mouse brains, revealing possible drop-off and mislocalization of ribosomes in the diseased samples. See Chapter 4.2 for more details.

4 Biological case study

Having developed tools dedicated to ribosome profiling analyses, I applied both riboWaltz and riboScan to investigate translation in different datasets obtained in the Laboratory of Translational Architectomics (IBF-CNR, Trento) and aimed at investigating possible translational defects in a motor neuron disease, Spinal Muscular Atrophy.

Spinal muscular atrophy (SMA) is a progressive neurological disorder characterized by degeneration of lower motor neurons²³⁸, caused by genetic alterations of the Survival of Motor Neuron (Smn) gene that induce the production of low levels of SMN protein²⁴³. Even though the genetic cause of SMN is well-established, the molecular mechanisms that link SMN depletion to the pathogenesis of SMA are yet unclear. Recent findings demonstrated a strict relationship between SMN and the translational machinery^{245,246,272} leading to pathological dysregulation of protein synthesis²⁷³. In the Laboratory of Translational Architectomics (IBF-CNR, Trento) it has been recently showed that low levels of SMN are connected to a reduction in the number of ribosomes in polysomes in tissues of the central nervous system in a mouse model of SMA at early and late symptomatic stages, a defect in translation that correlates with SMA disease progression²⁴⁹. Nevertheless, a clear mechanism connecting SMN and translation has not yet been obtained. Moreover, preliminary observation from my Lab showed that in brain and spinal cord from healthy mice SMN binds the 40S subunit and ribosomes through RNA-independent interactions. This finding reinforced the hypothesis that SMN protein could play a major role in regulation of translation and that its loss, in SMA conditions, might directly impact translation.

To investigate possible mislocalization of ribosomes caused by reduced level of SMN, I first applied riboWaltz and riboScan to compare Poly-RiboSeq and Active-RiboSeq data (Chapter 4.2) from brains of early-symptomatic SMA-affected mice and control littermates. Then, I took advantage of RiboSeq data obtained from SMN-specialized ribosomes in control mouse brains (Chapter 4.3 and 4.4) to identify mRNAs preferentially controlled by ribosomes associated to SMN protein and unravelling their positions along the transcripts.

4.1 Ribosome profiling datasets

To this aim, I took advantage of three ribosome profiling protocols: Poly-RiboSeq, Active-RiboSeq and SMN-specific RiboSeq (Figure 4.1 and section 1.4.2 for more details).

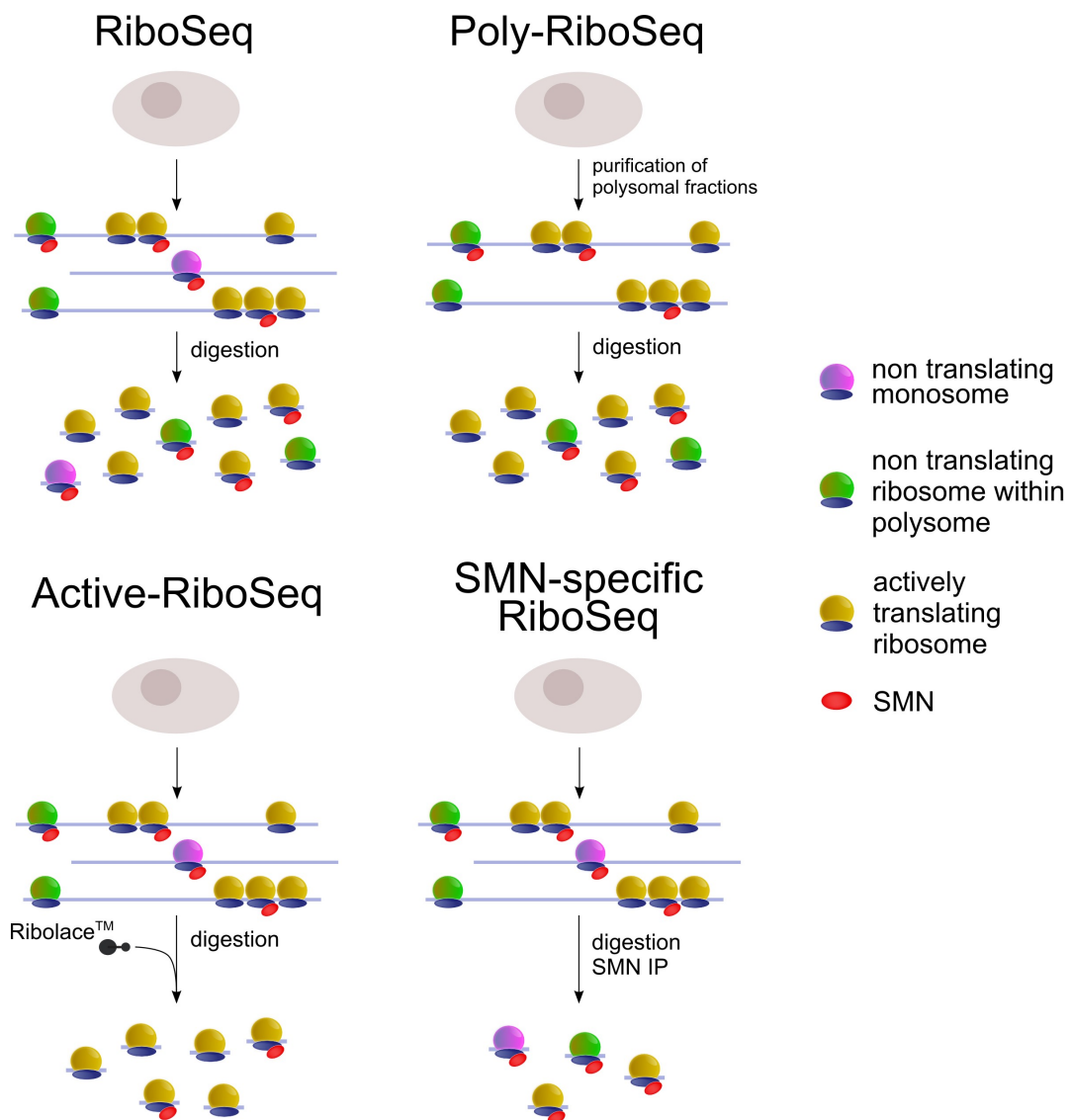


Figure 4.1. **Differences in ribosome profiling assays used in this thesis.** Diagram of the experimental protocols for Ribo-Seq (top left); Poly-RiboSeq (top right): pre-purification of polysomes removes any possible contamination associated to monosomes; Active-RiboSeq (bottom left): the RiboLace™ technology exclusively captures actively translating ribosomes, getting rid of non-translating ribosomes associated to polysomes; SMN-specific RiboSeq (bottom right): only fragments of transcripts protected by SMN-associated ribosomes are captured. After the extraction of ribosome protected fragments of mRNAs, library preparation and deep sequencing are performed.

With respect to the RiboSeq assay proposed by Ingolia and co-workers¹³⁸, Poly-RiboSeq removes any possible contamination associated to putative non-translating monosomes by applying the nuclease digestion step after pre-purification of polysomes (see Appendix for further details). Active-RiboSeq is a newly developed technique that uses a technology called RiboLace™. RiboLace™ has been developed by IMMAGNA Biotechnology (<http://www.immaginabiotech.com/products/ribo-lace-technology/>) and allows to obtain ribosome protected fragments only from actively translating ribosomes. Finally, SMN-specific RiboSeq captures ribosome protected fragments isolated from SMN-associated ribosomes, after sub-fractionation of ribosomes followed by immunoprecipitation (IP) of SMN.

Poly-RiboSeq and Active-RiboSeq were performed in brains from early-symptomatic and littermate control mice in biological duplicate (Figure 4.2). The mouse model of SMA was the Tawanese model²⁷⁴ and were collected at post-natal day 5 (P5) in the Laboratory of Prof. Thomas Gillingwater, Univeristy of Edinburgh. Poly-RiboSeq data were produced in my Lab (Laboratory of Translational Architectomics, IBF-CNR, Trento) by Dr. Paola Bernabò. Active-RiboSeq data were obtained in collaboration with IMMAGINA BioTechnology (<http://www.immaginabiotech.com/>).

SMN-specific RiboSeq data were produced in the Laboratory of Translational Architectomics (IBF-CNR, Trento) using control P5 brains and run in biological triplicate (Figure 4.2). As control for aspecific binding, immunoprecipitation using IgG was performed in parallel.



Condition	Technique	Replicas
 Healthy	Poly-RiboSeq	2
	Active-RiboSeq	2
	SMN-specific RiboSeq	3
 SMA - early symptomatic	Poly-RiboSeq	2
	Active-RiboSeq	2

Figure 4.2. **Ribosome profiling datasets.** Three ribosome profiling protocols were used: Poly-RiboSeq, Active-RiboSeq and SMN-specific RiboSeq. All experiments were performed in biological duplicate or triplicate for a total of 11 datasets.

Deep sequencing of polysomal RNAs (PolSeq) in the two conditions was also performed, in parallel.

Please refer to the Appendix at the end of the elaborate for the additional details about Poly-RiboSeq and SMN-specific RiboSeq, pre-processing and alignment steps for all the ribosome profiling data. A table displaying the number of reads left after each step of the alignment is also reported (Table A1). The Active-RiboSeq protocol is reported by Clamer et al.¹⁶².

4.2 Ribosome profiling of early-symptomatic SMA mouse brains

In this section I analyse Poly-RiboSeq and Active-RiboSeq performed in mouse from early symptomatic and control brains. First, I verify the ability of Active-RiboSeq in capturing ribosome protected fragments, looking for potential differences with respect to Poly-RiboSeq (section 4.2.1). Secondly, I investigate a decrease in mapping reads along the coding sequence with respect to the first 5 codons of the CDS emerging for Active-RiboSeq in the SMA condition. Thirdly, I explore the causes of quantitative differences between control and SMA-affected mouse brains in the amount and position of read enriched regions detectable along the 3' UTR.

4.2.1 Ribosome profiling of actively translating ribosomes

To verify the ability of Active-RiboSeq in capturing ribosome protected fragments and, more in general, to evaluate the quality of the four ribosome profiling datasets, I followed a four-step approach. For the two techniques (Poly-RiboSeq and Active-RiboSeq) and the two conditions (control and SMA) I i) assessed the reproducibility of the biological replicas; ii) generated the distribution of the length of the reads; iii) computed the percentage of P-sites falling in the three regions of the transcripts (5' UTR, CDS, and 3' UTR); iv) verified the trinucleotide periodicity along the coding sequence.

I assessed the reproducibility of the biological replicas by correlating the number of mapped reads per transcript between the replicas for each condition and each RiboSeq technique (Figure 4.3). The significance of the resulting correlations was measured with correlation tests, always obtaining statistically significant p-values (***) $p < 0.001$.

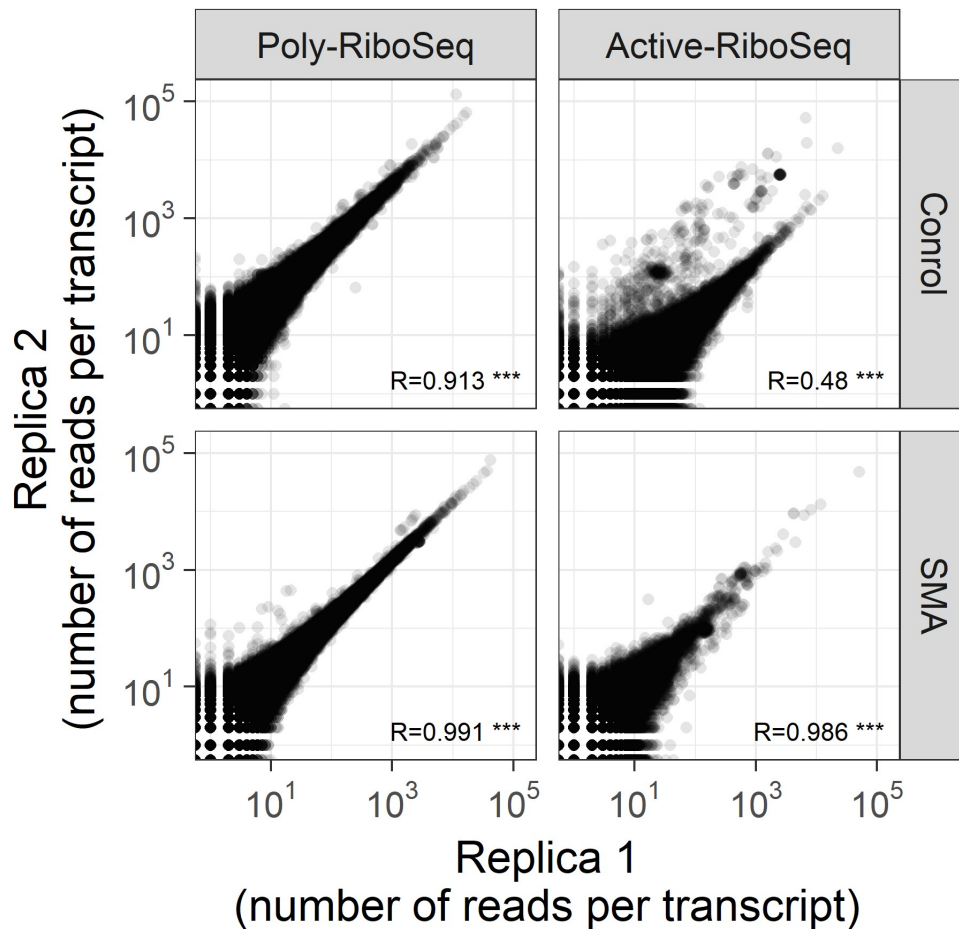


Figure 4.3. **Reproducibility of RiboSeq in biological replicas.** Scatterplots of the number of reads per transcript between the two replicas of Poly-RiboSeq and Active-RiboSeq for both the control and the SMA sample. The correlation coefficients and the statistical significances are shown (***) p-value < 0.001).

After the removal of the read duplicates and the acquisition of the alignment files I generated the distribution of the read length, temporarily combining the two replicas of each technique and condition (Figure 4.4). I observed a slight increase in the frequency of shorter reads (smaller than 25 nucleotides) in Active-RiboSeq, possibly due to alternative ribosome conformations^{275,276}. However, the most abundant populations always correspond to reads of 28 and 29 nucleotides, consistent with canonical eukaryotic ribosome footprints^{97,151}.

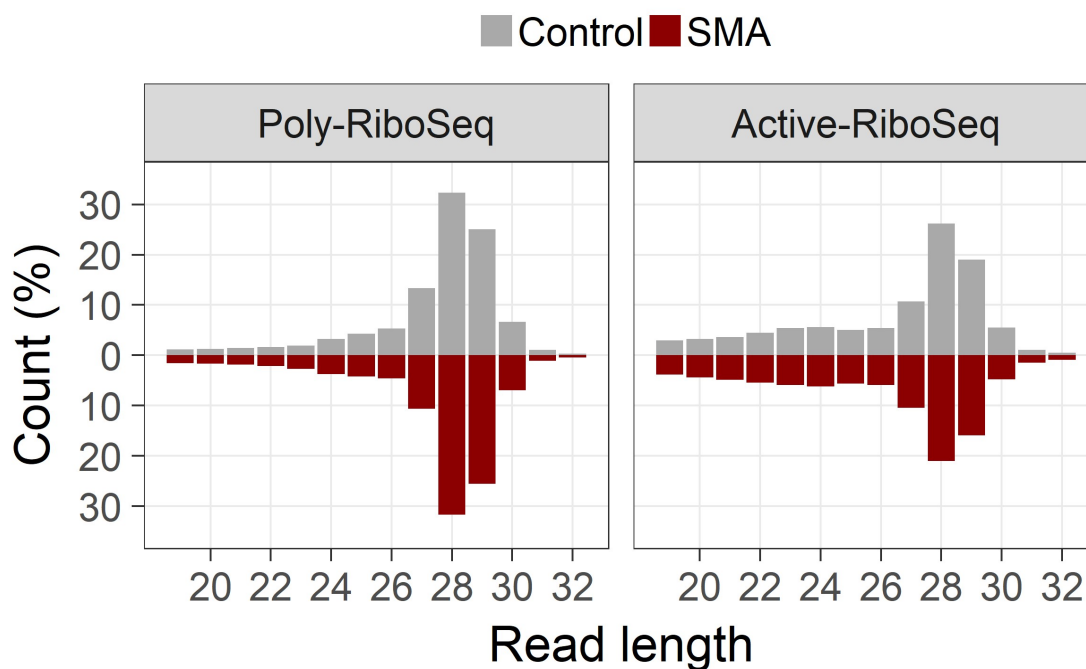


Figure 4.4. **Length of ribosome protected fragments.** Distribution of the length of the reads for Poly-RiboSeq and Active-RiboSeq in control SMA-affected mice. Poly-RiboSeq control: ~7,100,000 reads; Poly-RiboSeq SMA: ~8,400,000 reads; Active-RiboSeq control: ~ 900,000 reads; Active-RiboSeq SMA: ~ 200,000 reads.

In order to identify a subset of reliable and adequately covered transcripts, before proceeding with further analysis I performed the normalization and filtering steps provided by RiboScan. Poly-RiboSeq and Active-RiboSeq data were processed separately, due to the different protocols used for capturing mRNA ribosome protected fragments. A common set of 4324 mRNAs from both the techniques was identified and used for all subsequent analyses.

Next, I run riboWaltz on all samples to obtain further evidence about the ability of the RiboLace™ technology to extract actively translating ribosomes, based on their P-site localization. The identification of the P-site position within the reads was performed using riboWaltz with the automatic detection of the optimal extremity and P-site offsets (see *Identification of the P-site position* paragraph, method section of Chapter 3.1). For each sample the optimal extremity was identified as the 3' end and the optimal offset was set to 16 nucleotides.

Ribosome profiling data should highlight the CDS region of transcripts as the region with the higher percentage of reads. To see if this is the case in my data, I computed the percentage of P-sites falling in the three regions of the transcripts (5' UTR, CDS,

and 3' UTR) for both the RiboSeq techniques. These results were compared to the expected distribution of randomly mapped reads (based on the cumulative nucleotide size of 5'UTR, CDS and 3'UTR, respectively) and to the distribution obtained from deep sequencing of polysomal RNA (PolSeq), whose signal should also reflect the random fragmentation of RNA and the corresponding mapped reads along transcript regions (Figure 4.5). In the last case the position of the read was set to its central nucleotide, since no P-site can be identified from PolSeq. As expected, the results in the case of PolSeq assay are similar to the random distribution, while an enrichment in the coding sequence can be observed for both Poly-RiboSeq and Active-RiboSeq.

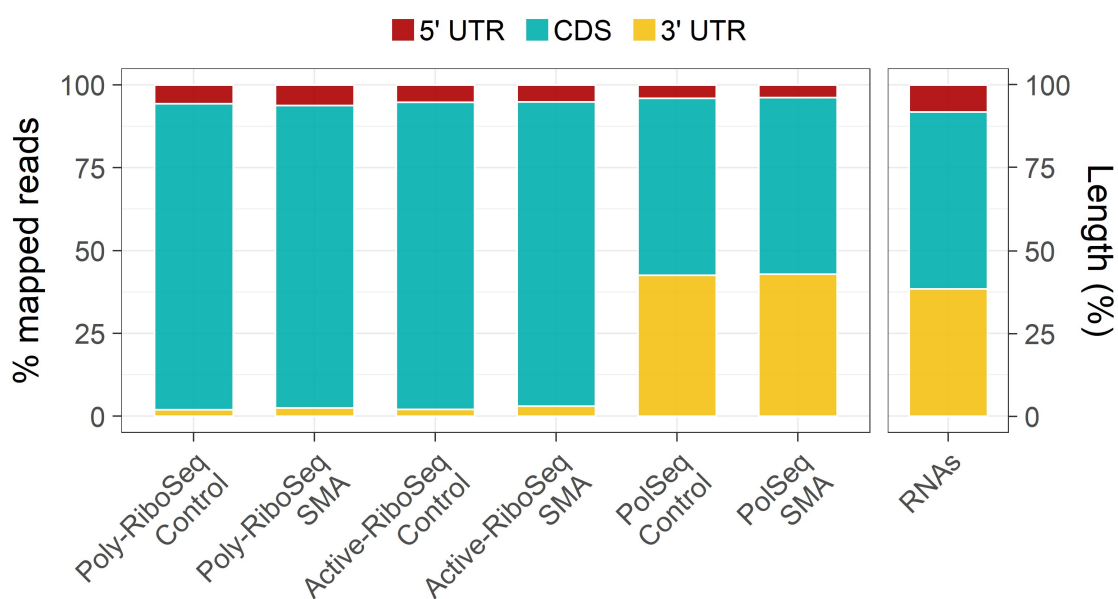


Figure 4.5. **Enrichment along the coding sequence of RiboSeq reads.** The bar plots displays the percentages of reads aligning on three mRNA regions (5' UTR, coding sequence and 3' UTR) for PolSeq and RiboSeq assays. The last bar represents the percentage of region length.

Then, I verified the presence of the trinucleotide periodicity of the ribosome footprints along the coding sequences. To do that, I first looked at the percentage of P-sites corresponding to the three reading frames for 5' UTR, CDS and 3' UTR, stratifying the reads by length (Figure 4.6). I observed an enrichment of P-sites in the first frame along the coding sequence but not along the UTRs, proving that both Poly-RiboSeq (Figure 4.6A) and Active-RiboSeq (Figure 4.6B) reads are in the correct frame, in agreement with ribosome protected fragments from coding mRNAs. I also showed the absence of strong differences in the trinucleotide periodicity in all types of ribosome profiling.

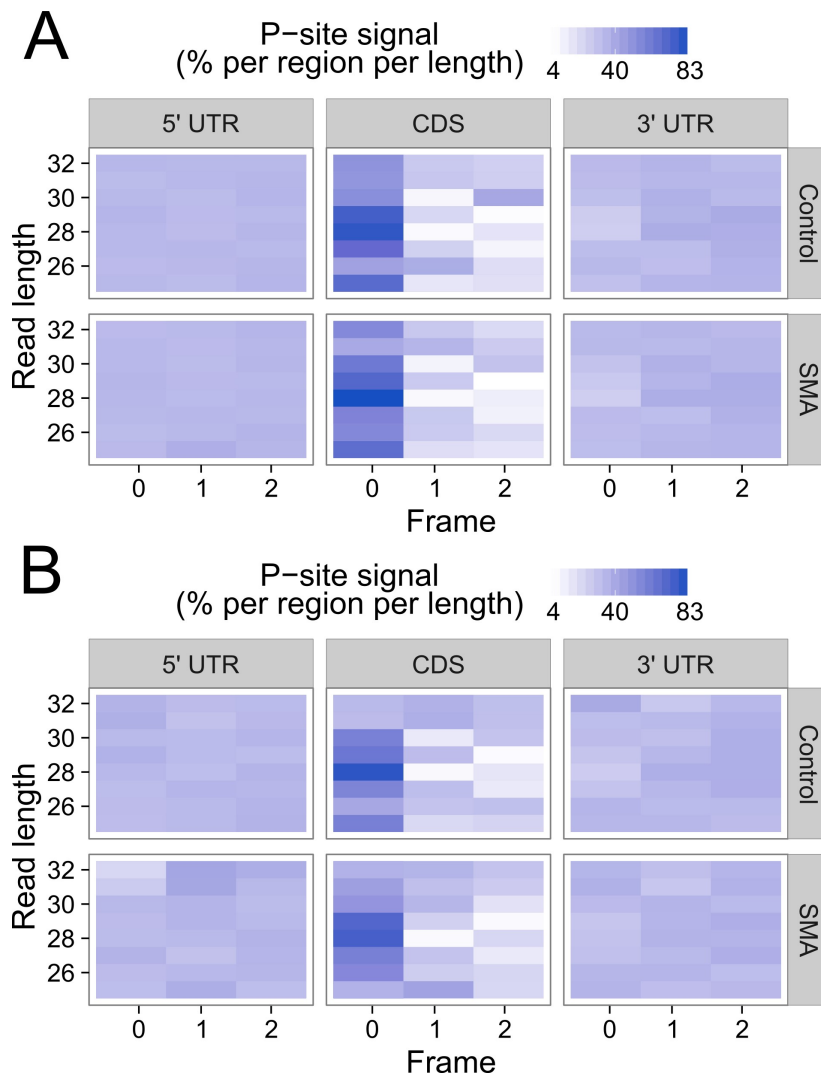


Figure 4.6. **Percentage of P-sites in frame.** Heatmap of the percentage of P-sites according to the three reading frames for 5' UTR, CDS and 3' UTR of mRNAs, stratifying the reads by length for (A) Poly-RiboSeq and (B) Active-RiboSeq.

To provide a visual representation of the trinucleotide periodicity along the coding sequence and look for potential differences in global profiles from control and SMA affected mice, I produced P-site-based meta-profiles (Figure 4.7). I overlay for each ribosome profiling dataset the profiles associated to the two conditions. To compare the meta-profiles, I displayed the density of the signal around the translation initiation and translation termination sites, so that the area under each meta-profile (composed by the portion around the start codon and the portion around the stop codon) is equal to 1. Both Poly-RiboSeq and Active-RiboSeq meta-profiles show a clear periodicity along the CDS and a lower, almost uniform signal along the UTRs, confirming the previous observations.

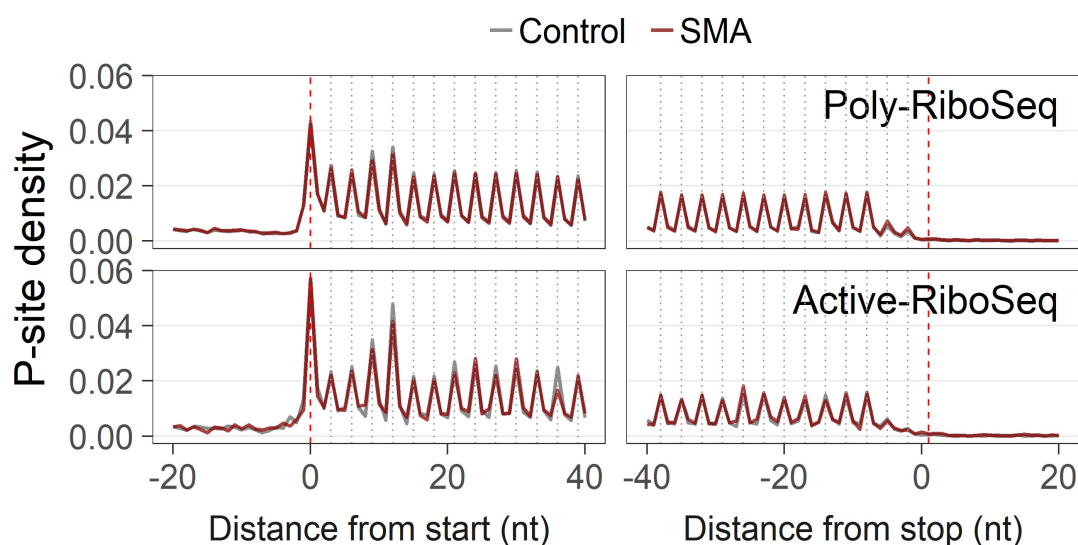


Figure 4.7. **Trinucleotide periodicity along the CDS.** Overlay meta-profiles based on the P-site position of the reads along transcripts for control and SMA samples from Poly-RiboSeq and Active-RiboSeq assays.

Concerning the results of Active Ribo-seq, the results obtained so far strongly support the ability of Active-RiboSeq technique and RiboLace™ to purify *bona fide* mRNA fragments protected by ribosomes. In fact, the distribution of the reads length (Figure 4.4), the accumulation of reads mapping on the coding sequence (Figure 4.5) and the presence of clear trinucleotide periodicity in the correct frame along the coding sequence but not on the UTRs (Figure 4.6 and Figure 4.7) stand in close agreement with the results obtained for Poly-RiboSeq, a well-established RiboSeq protocol. These data, supported by other validation (not shown) from the Laboratory of Translational Architectomics (IBF-CNR, Trento) and Immagina BioTechnology prove that RiboLace™ technology can capture active ribosomes and that it can be used to localised translating ribosomes.

4.2.2 Ribosome drop-off in SMA

Comparing Active-RiboSeq and Poly-RiboSeq meta-profiles (Figure 4.7) it's possible to observe that the signal of peaks at the beginning of the coding sequence is higher in Active-RiboSeq profiles than in Poly-RiboSeq ones. Starting from this observation, I further investigated the involvement of SMN in the alterations of ribosome localization along the transcripts of diseased mice, providing evidences of ribosome drop-off caused by the loss of SMN.

Han and collaborators²⁷⁷ deeply examined an analogous accumulation of reads on the 5th codon emerged in RiboSeq data of Hek-293. The authors attributed the post-initiation ribosome stalling, already observed in other studies^{263,278,279}, to the geometry of the peptide exit tunnel of the ribosome, suggesting the existence of a functional pause of the translation machinery for productive protein synthesis. Therefore, an accumulation of ribosomes at the 5th codon is a good indication of active translation.

Han and collaborators also excluded a connection between the accumulation of ribosomes at the 5th codon and the nucleotide composition of the mRNAs²⁷⁷. To corroborate this result, I performed a sequence enrichment analysis around the 5th codon for the Active-RiboSeq assay in control and SMA samples, separately. For both conditions I used as foreground the transcripts showing a signal on the 5th codon and as background the 4324 mRNAs, taking into consideration sequences ranging from 3 to 10 nucleotides. From this analysis no enriched sequence emerged, confirming that the nucleotide composition of the mRNAs do not account for accumulation of ribosomes at the 5th codon. a good indication of active translation.

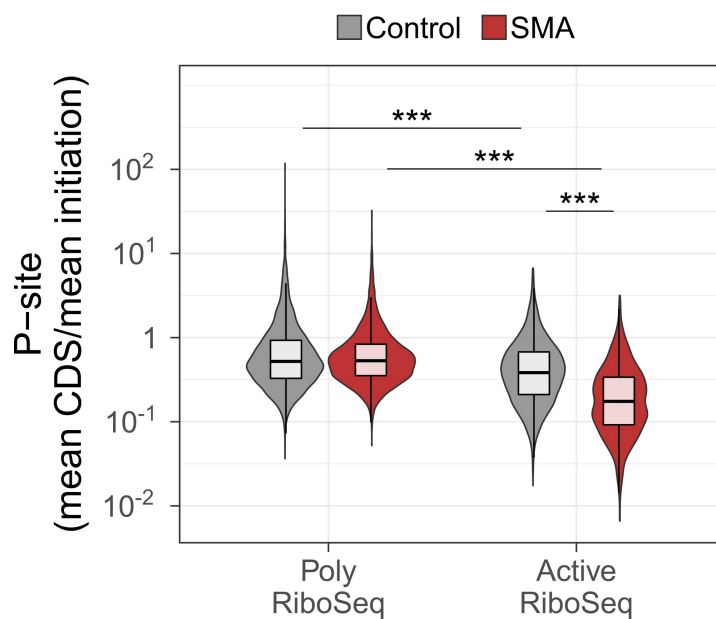


Figure 4.8. **P-site signal close to the start codon.** Violin plots showing the distribution of the ratios between the average number of P-sites on the coding sequence and the average number of P-sites on the first 5 codons for Poly-RiboSeq and Active-RiboSeq. The box plots associated to the distributions are also reported, along with the statistical significances from the Wilcoxon-Mann-Whitney test (***) p-value < 0.001).

To further investigate the presence of a ribosome accumulation at the beginning of the coding sequence and reveal potential differences between the conditions, I took the

cue from the analysis workflow proposed in Han and collaborators²⁷⁷ and split each transcript in two regions. The first region includes the nucleotides from 0 to 14 (corresponding to the first 5 codons of the transcript) and the second nucleotides from 15 to the end of the coding sequence. Then, I calculated the ratio between the average number of P-site falling in the second region and the average number of P-sites on the first one (Figure 4.8).

The results show a general and statistically significant decrease in the ratios calculated for Active-RiboSeq with respect to Poly-RiboSeq in both conditions. This result can be possibly explained by the detection of RNA protected fragment from of non-actively translating ribosomes that can be purified using with the standard RiboSeq procedures but not by RiboLaceTM. Interestingly, Active-RiboSeq shows a significant difference in ratio distributions in SMA and control, suggesting an accumulation of ribosomes around the start codon or, alternatively, a lower amount of ribosomes along the coding sequence, consistent with higher drop-off rates in SMA samples. Interestingly, the latter hypothesis suggests a down-regulation of protein production in SMA mice, in agreement with what found in literature^{249,273}.

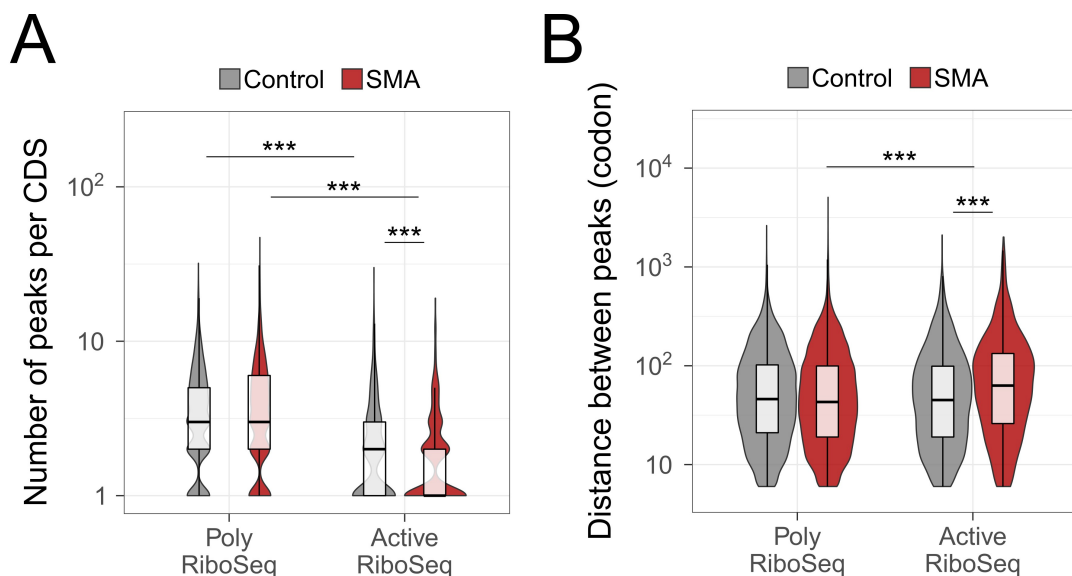


Figure 4.9. **Peaks analysis.** Violin plots showing (A) the distribution of the number of peaks per coding sequence and (B) the distribution of the distance between consecutive peaks along the coding sequence for Poly-RiboSeq and Active-RiboSeq. The box plots associated to the distributions are also reported, along with the statistical significances from the Wilcoxon-Mann-Whitney test (***) p-value < 0.001).

To better investigate whether the observed decrease of the number of ribosomes along the mRNAs in SMA is connected to the involvement of SMN in translation, I

exploited riboScan. I looked for transcript-specific and statistically significant enriched regions (hot-spots and peaks) along the sequences and for potential differences in both their position between control and SMA-affected mice. I analysed Poly-RiboSeq and Active-RiboSeq data in parallel, possibly highlighting a characteristic behaviour of active ribosomes. After the detection of hot-spots and their aggregation into peaks, I computed the number of peaks per CDS and the distance between each pair of consecutive peaks detected on the CDS (Figure 4.9).

The results showed a statistically significant decrease in the number of peaks and an increased distance in the diseased mice with respect to the control in Active-RiboSeq. These findings suggest a lower number of actively translated ribosomes along the coding sequence and longer portions of naked mRNA. As in the previous analysis, Poly RiboSeq did not show any change between SMA and control samples. Overall, these results are in agreement with ribosome drop-off that might cause a reduction of actively translating ribosomes and, consequently, an increment in the spacing between consecutive ribosomes that can be captured better with RiboLace™ than with conventional ribosome profiling.

4.2.3 Mislocalization of ribosomes along the 3' UTR in SMA?

To obtain additional insights into the possibility that SMN loss could be responsible for translational defects in SMA-affected mice, I examined more in details the outcomes of riboScan. In particular, I looked for significant enriched regions along the transcripts by computing the number of hot-spots separately for the three transcript regions (5' UTR, CDS and 3' UTR), as shown in Figure 4.10A.

Along the 5' UTR and the CDS a difference in the total number of hot-spots in Poly-RiboSeq and is 9-10 times higher than for Active-RiboSeq. This strong difference can be imputed to the lower amount of mapped reads for Active-RiboSeq with respect to Poly-RiboSeq or to the presence of a high number of not-translating ribosomes in this sample. In other words, this result is compatible with few ribosomes being under active translation in brain. Figure 4.10A also shows a similar number of detected hot-spots between control and SMA samples for both Poly-RiboSeq and Active-RiboSeq in the 5' UTR and the CDS but not in the 3' UTR. Interestingly, in this last region the diseased samples have twice as many hot-spots as the healthy ones (Figure 4.10B).

To unravel potential differences in the number of hot-spot aggregates along the transcripts, I combined hot-spots into single peaks. I confirmed a large disparity in the

number of enriched regions extracted for the two techniques (Figure 4.10C) and a huge difference in the number of identified peaks along the 3' UTR of the SMA samples (Figure 4.10D).

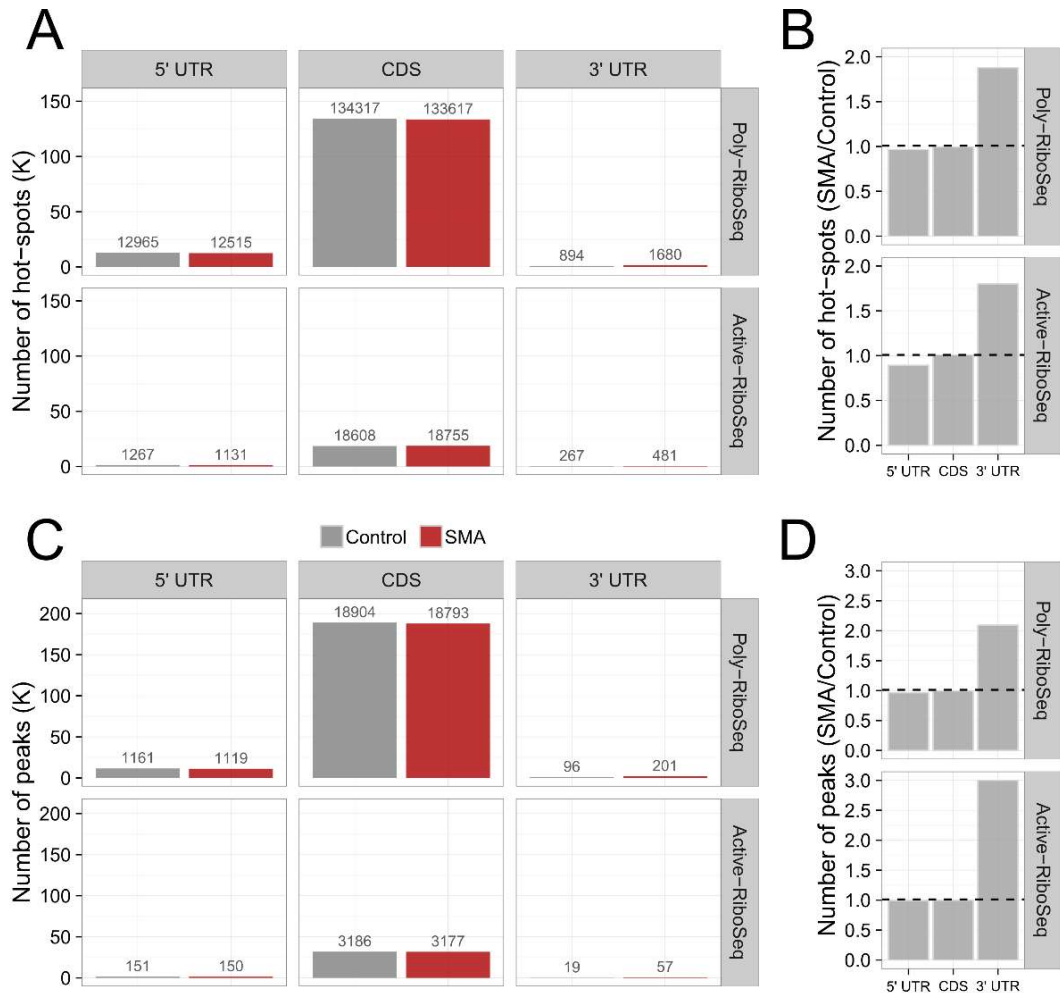


Figure 4.10. **Number of hot-spots and peaks along the ribosome occupancy profiles.** (A) Bar plots of the number of hot-spots divided by their localization along the transcripts (5' UTR, CDS, 3' UTR). The number of hot-spots is reported at the top of the bars. (B) Ratio between the number of hot-spots detected in the SMA samples and in the control ones for the three transcript regions. (C) Bar plots of the number of peaks (combined hot-spots) divided by their localization along the transcripts (5' UTR, CDS, 3' UTR). The number of peaks is reported at the top of the bars. (D) Ratio between the number of peaks detected in the SMA samples and in the control ones for the three transcript regions.

To deepen the analysis of the enriched regions I assessed the number of peaks detected in the two conditions overlap (common peaks), and those that are uniquely associated to either the healthy or the diseased samples (control and SMA specific peaks, respectively). To be as restrictive as possible in the isolation of specific peaks, I

defined two overlapping peaks as two regions that share at least one nucleotide, independently from their lengths. The percentage of common and uniquely identified peaks in the three transcript regions (5' UTR, CDS and 3' UTR) is shown in Figure 4.11. The results show that peaks uniquely detected along the 5' UTR and CDS in the SMA and control sample are relatively less abundant than the common ones for Poly-RiboSeq than for Active-RiboSeq. However, small differences in the number of specific peaks can be observed between the two conditions, while on the 3' UTR the peaks uniquely detected in the SMA samples are almost six times the number of control specific peaks. This observation suggests the possibility of ribosome readthrough of the stop codon^{89,280} or the presence of sliding ribosomes along the 3' UTR still associated to the tRNAs due to defects in the termination phase of translation, at least for a limited set of transcripts.

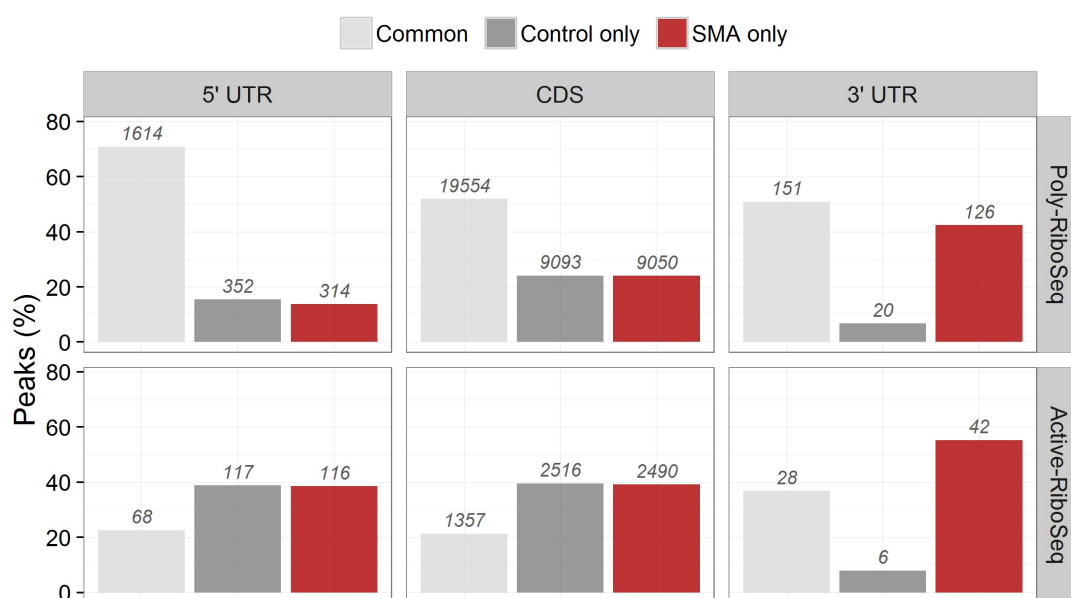


Figure 4.11. **Analysis of significant ribosome peaks.** Percentage of peaks in common between control and early symptomatic samples (Common); uniquely detected in the control (Control only) or in the early symptomatic sample (SMA only). The total number of peaks is reported at the top of the bars.

Another possible explanation for the observed signal along the 3' UTR is that these reads are the product of RNA regions protected by RNA binding proteins. To exclude the presence of protected fragments associated to RNA binding protein along the 3' UTR and captured by the RiboSeq assays, I explored if the median length of the reads along the 3'UTR might be consistent with RBP protected fragments. RBP-RNA interactions take place through protein domains covering a maximum of 4-8

nucleotides²⁸¹. Hence, RBP protected fragments are usually shorter than ribosome protected fragments unless they form complexes.

I investigated the length of the reads contributing to the SMA specific peaks for the two techniques. I compared the read length distribution between the whole SMA sample and the subset of reads specifically associated to the uniquely detected SMA peaks, performing the Wilcoxon-Mann-Whitney test for assessing the statistical significance. Figure 4.12 shows a decrease of the distribution toward shorter reads for Poly-RiboSeq, while for Active-RiboSeq no significant differences arise. The median length of reads in 3' UTR peaks is of 26 nucleotides, consistent with the length of ribosome protected fragments. Therefore, it is quite unlikely that this signal might be produced by protected fragments associated to RNA binding protein along the 3' UTR and captured by the RiboSeq assays.

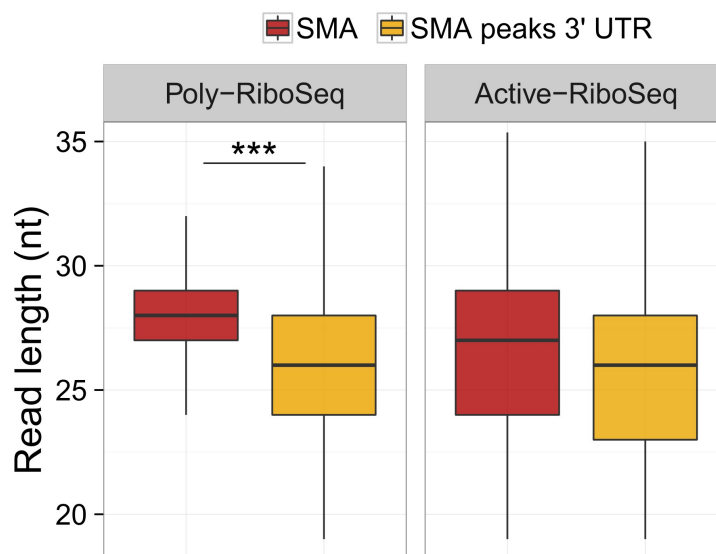


Figure 4.12. **Read length associated to SMA specific peaks.** Distribution of the length of the reads for the whole SMA dataset and for the reads that contribute to the enriched regions detected along the 3' UTRs (SMA peaks 3' UTR) for Poly-RiboSeq and Active-RiboSeq. The statistical significances from the Wilcoxon-Mann-Whitney test is shown (***) p-value < 0.001).

To further investigate the hypothesis of ribosome readthrough in the diseased samples, I performed two analyses associated to the SMA specific peaks for Poly-RiboSeq and Active-RiboSeq, separately. Since ribosome readthrough events are known to be promoted by specific sequences close to the stop codon in mammals^{26,282-284}, I investigated the presence of possible enrichment among all the sequences comprised between the last 6 nucleotides of the CDS and the first 4 of the 3' UTRs. I used as foreground the transcripts associated to the SMA specific peaks and

as background the whole set of 4324 filtered mRNAs. This analysis didn't result in any statistically significant enriched sequence motif, suggesting that ribosome readthrough is unlikely the cause of the observed effect.

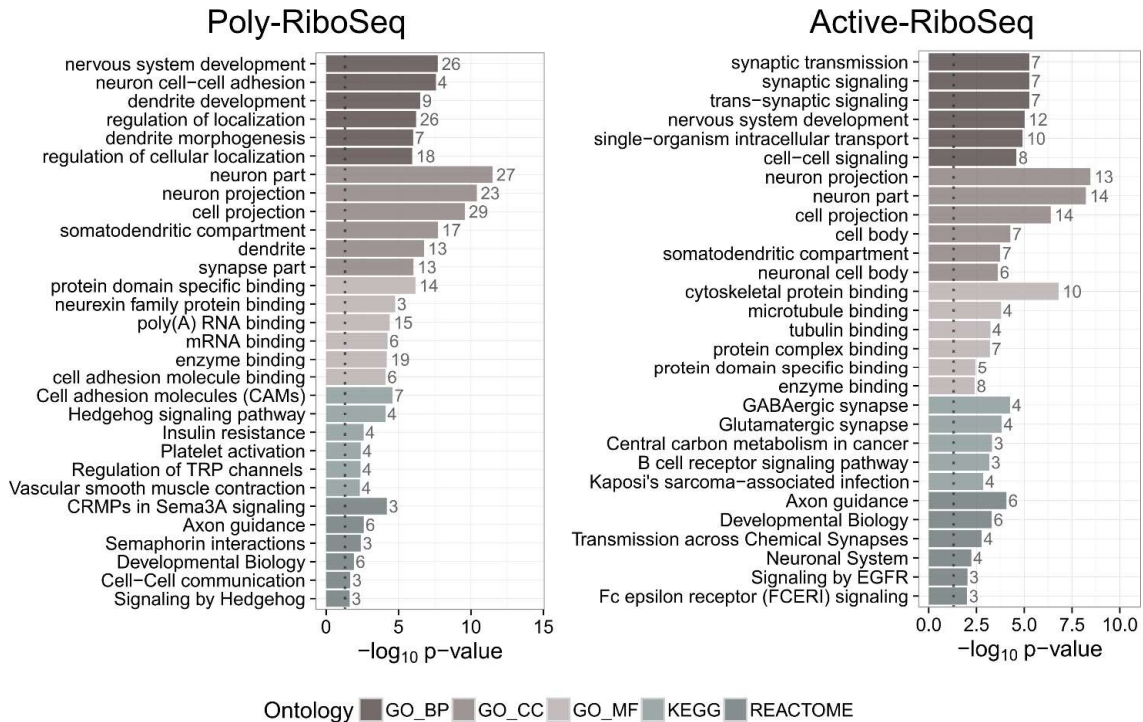


Figure 4.13. **Gene ontology analysis on transcripts associated to SMA specific peaks.** Results of Gene Ontology, KEGG and REACTOME pathway enrichment analysis on those genes with at least one peak uniquely detected in SMA and in the 3' UTR of the Poly-RiboSeq (left panel) and Active-RiboSeq (right panel). Respectively, 94 and 36 genes were used in the analysis. The number of genes associated to the corresponding term is displayed on the right of the bars. The terms are divided according to the three main categories of the Gene Ontology: biological process (GO_BP), cellular component (GO_CC) and molecular function (GO_MF).

Even if at present the unexpected increase in 3'UTR signal in SMA has no clear explanation, I decided to understand if the presence of peaks on the 3' UTR might be associated either with neuro-specific processes or with the disease. To this aim, I performed an enrichment analysis of Gene Ontology terms, KEGG and REACTOME pathways on the subpopulation of genes presenting at least one peak uniquely detected on the 3' UTR of the SMA sample of Poly-RiboSeq and Active-RiboSeq (Figure 4.13). Interestingly, a set of biological terms associated to neuro-development, cell signalling, dendrite and axon guidance mechanisms arose, coupled with many cytoskeleton-related terms for both techniques. These results definitively need much deeper investigations about the possible mechanism connected to the presence of the signal in 3'UTR. Nonetheless, the genes involved in this effect, point to diffuse and

strong connections between significant enriched regions along the 3' UTR and the development of the nervous system in early symptomatic mice.

4.3 Ribosome profiling of SMN-specialized ribosomes reveals a role for SMN in translation of the first codons

Recent studies suggest unexpected connections between the structural constituents of the translation machinery and translational controls of gene expression through the emerging hypothesis of the so-called specialized ribosomes^{68,69}. The term “specialized” refers to their unique composition in ribosomal proteins or to the definite activity carried out in cells⁶⁹, and that point to ribosome itself as a largely unexplored and direct player in the control of translation in both physiological and pathological conditions.

These findings and the fact that in the Laboratory of Translational Architectomics (IBF-CNR, Trento) preliminary data showed that SMN protein is a direct interactor of the ribosome, prompt further studies aimed at the investigation of SMN-specialized ribosomes (i.e. ribosomes that are bound by SMN), their role in tuning translation and their association to SMA, as discussed in this section.

To this aim, in the Laboratory of Translational Architectomics (IBF-CNR, Trento), a SMN-specialized ribosome profiling from healthy P5 brain was developed using a dedicated immunoprecipitation protocol (see Appendix). In parallel a control sample from immunoprecipitation of IgG for assessing the aspecific binding of ribosomes to the beads was used. The ribosome protected fragments from this sub-population of ribosomes was isolated and sequenced as detailed in Appendix. I started the analysis by identifying transcripts enriched in fragments protected by SMN-associated ribosomes with respect to the control IgG (see Appendix for details on the enrichment analysis). After the enrichment, 1095 transcripts, corresponding to 901 genes, were identified applying a double threshold on fold enrichment (>2) and statistical significance (<0.05). Annotation enrichment analysis with Gene Ontology terms, KEGG and REACTOME pathways was performed on these genes (1095 transcripts), revealing their association with neuro-related, RNA binding-and rRNA-related terms (Figure 4.14). This selected pool of mRNAs was used for the following analyses of the SMN RiboSeq data.

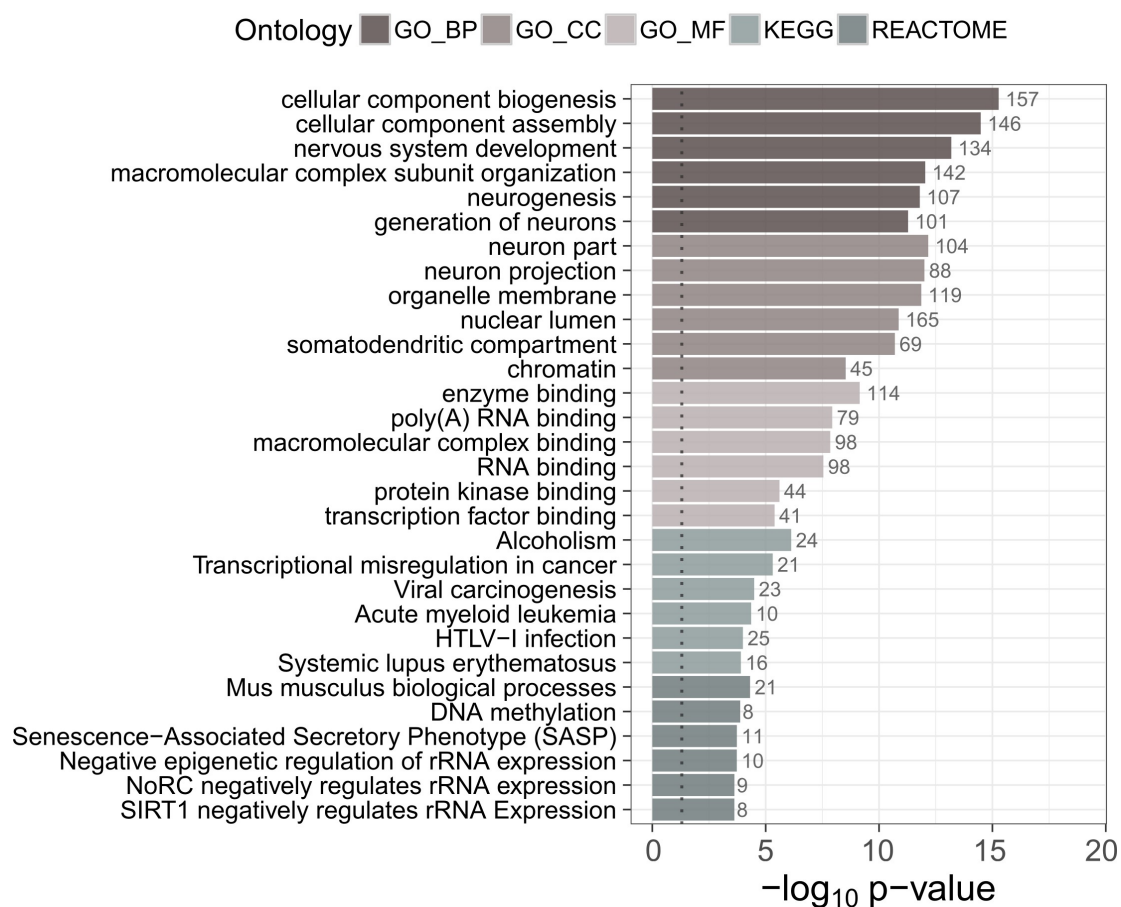


Figure 4.14. **Annotation enrichment analysis on transcripts associated with SMN interacting ribosomes.** Results of Gene Ontology, KEGG and REACTOME pathway enrichment analysis on the genes associated with SMN interacting ribosomes. 901 genes were used in the analysis. The number of genes associated to the corresponding term is displayed on the right of the bars. The terms are divided according to the three main categories of the Gene Ontology: biological process (GO_BP), cellular component (GO_CC) and molecular function (GO_MF).

Then, I generated the distribution of the read length aligning on the mRNA enriched in SMN-specialised ribosomes (Figure 4.15A). The most abundant populations always correspond to reads of 32 nucleotides, consistent with the canonical eukaryotic ribosome footprints^{97,151}. Nevertheless, a second population of reads is visible, peaking at 26 nucleotides. Interestingly, this bimodal distribution of the read length has been already observed in other ribosomal profiling data and attributed to alternative conformations of the ribosome^{275,276} that characterize different stages of ribosome translocation during protein synthesis²³. This intriguing result suggests the hypothesis that SMN-specialized ribosomes may be intimately associated to the mechanism of

translation, possibly stabilizing some specific intermediates of the ribosome during translation elongation.

To get further information on this, I investigated the accumulations of SMN-specialized ribosomes on the coding sequence and the presence of the trinucleotide periodicity along the CDS (expected for ribosomes protected fragments). I run riboWaltz on the three SMN replicas for the identification of the P-site position within the reads, using automatic detection of the optimal extremity (see *Identification of the P-site position* paragraph, method section of Chapter 3.1). The first step of RiboWaltz identified as optimal extremity for each sample the 5' end, indicating as optimal offset a stretch of 12 nucleotides. I computed the percentage of P-site falling in the three regions of the transcripts (5' UTR, CDS, and 3' UTR) for the RiboSeq SMN (Figure 4.15B). These results were compared to Poly-RiboSeq, Active-RiboSeq, used as control for the distribution in more classical ribosome profiling experiments, and PolSeq of healthy mouse brains. Figure 4.15B clearly shows that most of the SMN-specific reads map on the coding sequence, similarly to the other RiboSeq protocols and differently from PolSeq, where the signal is uniformly distributed on the full transcript. This finding suggests that SMN do immunoprecipitates ribosomes that are along the coding sequence, as expected for a ribosome profiling experiment.

Then, I explored the presence of the trinucleotide periodicity in SMN-specialized ribosomes along the coding sequences by generating the meta-profile based on the P-site of the SMN-specialized RiboSeq reads (Figure 4.15C) and by computing the percentage of P-sites according to the three reading frames for 5' UTR, CDS and 3' UTR, stratifying the reads by length (Figure 4.15D). Curiously P-site periodicity was weak for SMN-specialized ribosomes in the CDS, not much higher than what I can observe in the UTR regions. Moreover, P-sites showed two sharp peaks at the beginning of the coding sequence (Figure 4.15C), suggesting that SMN is mainly bound to ribosomes located on the AUG and around the 5th codon. As previously discussed, an analogous read accumulation have been observed in several studies^{263,278,279} and extensively investigated by Han and collaborators²⁷⁷. Prompted by this observation, I computed the ratio between the average number of P-site falling in the first 5th codons and the average number of P-sites on the rest of the coding sequences (Figure 4.15E), observing a strong accumulation of reads at the beginning of the CDSs with respect to the remaining region of the sequences.

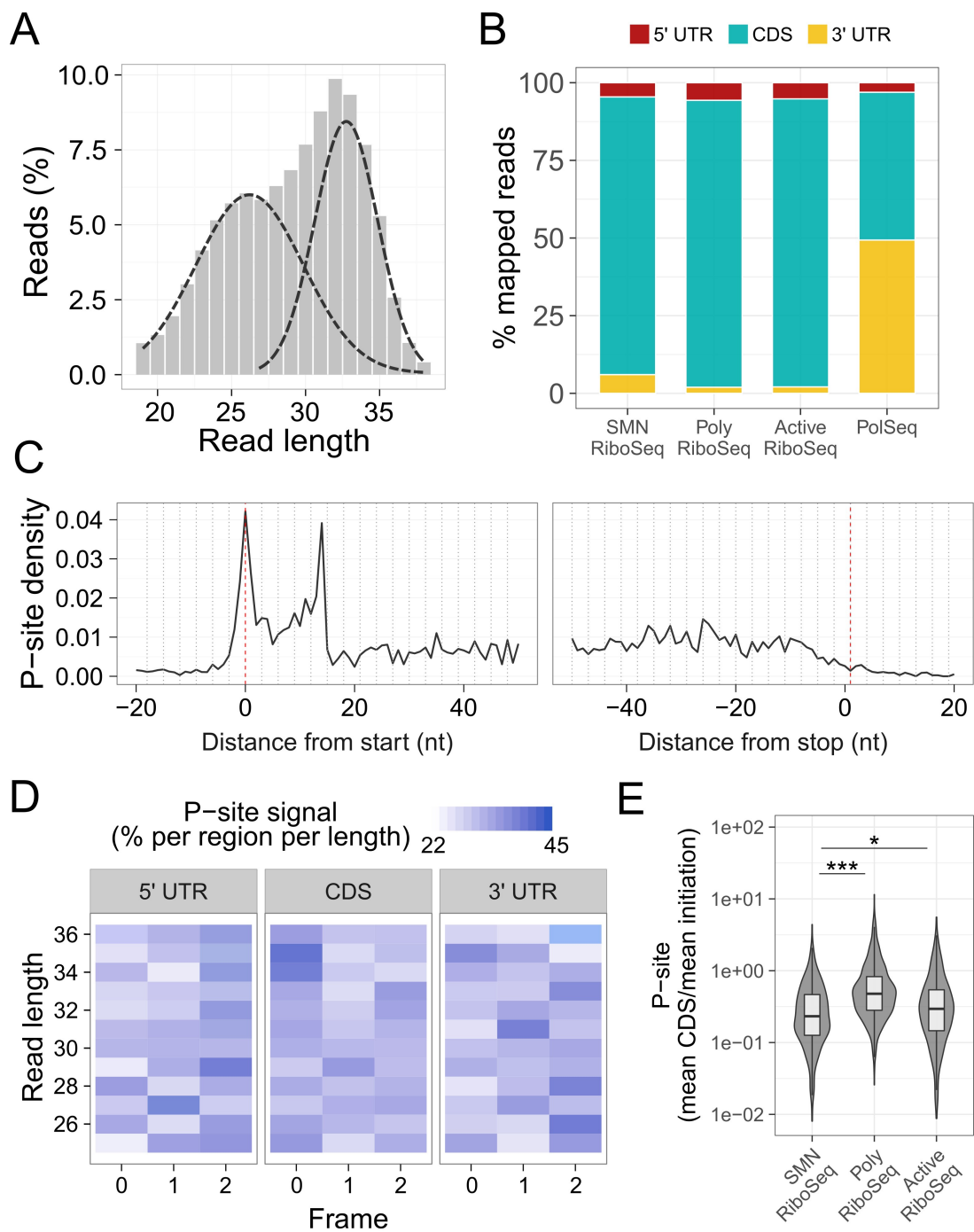


Figure 4.15. **Analysis of SMN-enriched transcripts.** (A) Distribution of the length of the reads for Poly-RiboSeq and Active-RiboSeq in SMA affected and control mice. The distribution was fitted with two Gaussian curves, represented as dashed lines. (B) Enrichment along the coding sequence of SMN RiboSeq reads. The bar plots displays the percentages of reads aligning on three mRNA regions (5' UTR, coding sequence and 3' UTR) for RiboSeq assays following SMN IP, Poly- and Active-RiboSeq of healthy mouse brains and PolSeq as control. (C) Meta-profiles based on the P-sites position of the reads along the transcripts for the SMN RiboSeq assays. (D) Heatmap of the percentage of P-sites according to the three reading frames for 5' UTR, CDS

and 3' UTR of mRNAs, stratifying the reads by length for SMN RiboSeq. (E) Violin plot showing the distributions of the ratios between the average number of P-sites on the first 5 codons and the average number of P-sites on the remaining coding sequence for SMN RiboSeq, Poly- and Active-RiboSeq of healthy mouse brains. The box plots associated to the distribution are also reported. The statistical significances from the Wilcoxon-Mann-Whitney test is shown (* p-value < 0.05, *** p-value < 0.001).

To understand how the two populations of reads observed in Figure 4.15A specifically contribute to the accumulation of reads at the beginning of the coding sequence, I generated the meta-profiles employing long (32-35 nts) and short (23-26 nts) reads (Figure 4.16).

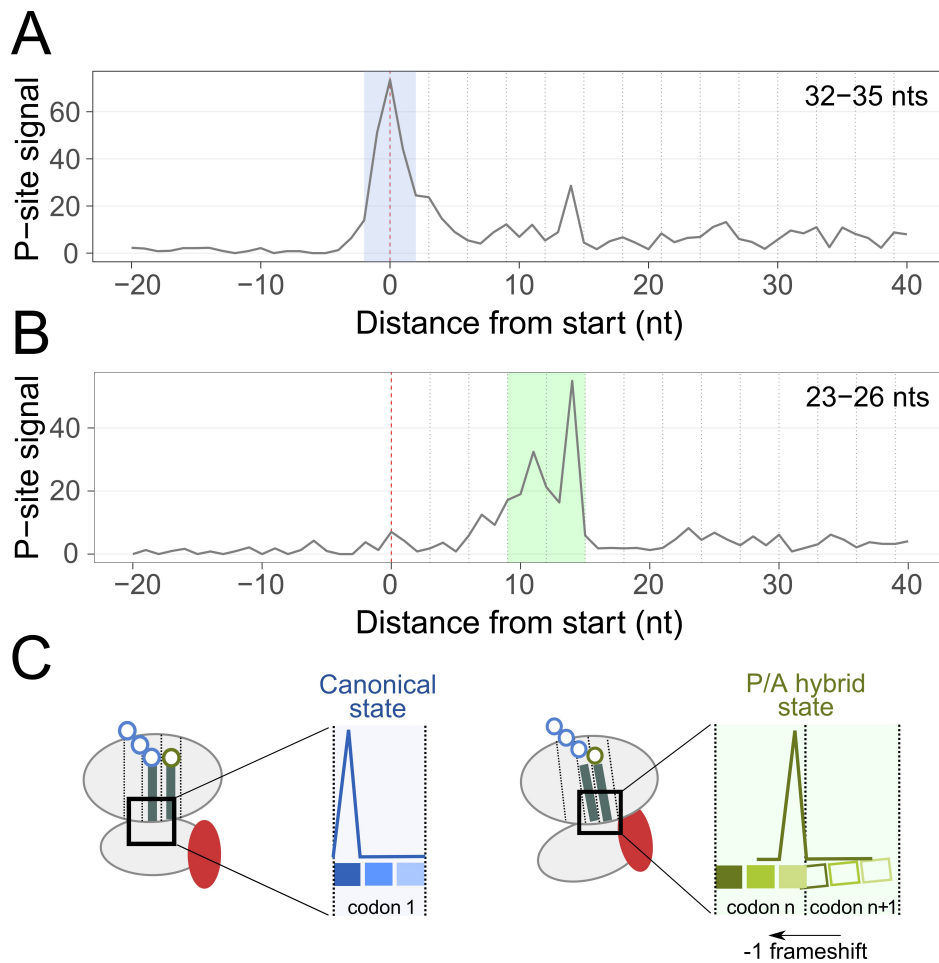


Figure 4.16. **Analysis of long and short reads.** (A) Meta profile generated by long reads (32-35 nucleotides). The blue shadow highlight the highest peak of the profile, on the start codon. (B) Meta profile generated by short reads (23-26 nucleotides). The green shadow highlight the highest peak of the profile, around the 5th codon. (C). Schematic representation of a ribosome in the correct frame on the start codon (left panel) and of a ribosome in a P/A hybrid state around the 5th codon (right panel).

Strikingly, long reads give rise to a neat in-frame peak on the start codon and the second codon, but no accumulation of reads on downstream triplets emerges to be in-frame (Figure 4.16A). On the contrary, short reads contribute to a neat increase of the signal around the 5th codon and not on the translation initiation site (Figure 4.16B). Moreover, a -1 frameshift displayed by the highest peaks of the metaprofile points to a peculiar conformation of the ribosomes translocating from one codon to the next one, e.g. an A/P hybrid state caused by the proximity of the tRNA in the A- and P-site of the ribosomes^{23,285} (Figure 4.16C).

All together, these findings are strongly suggesting the massive presence of SMN-specialized ribosomes at the beginning of the coding sequence. A possible explanation is a role of SMN in controlling the translation initiation phase: binding the ribosomal small subunit (results from the Laboratory of Translational Architectomics, IBF-CNR, Trento. Data not shown) SMN may act directly on the movement of the ribosomes on the first 5 codons of the coding sequence (Figure 4.17). Moreover, these findings coupled with the previous observations of significant ribosome enrichments along the 3' UTR SMA indicates a translational regulatory role of SMN at both the initiation and termination level, pointing to SMN as a potential determinant in ribosome recycling^{12,24,25,27,286} in physiological conditions.

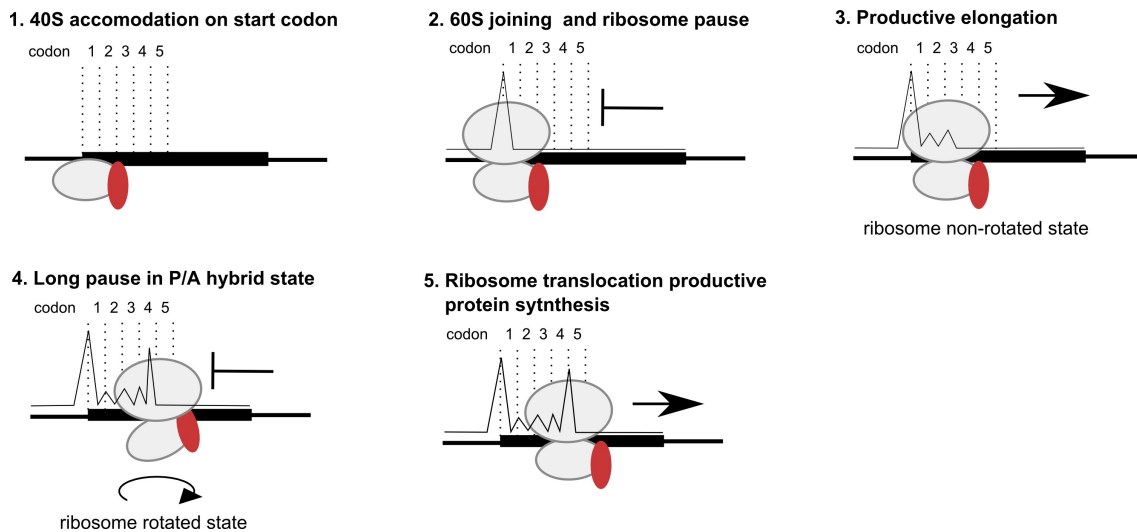


Figure 4.17. **Role of SMN in translation initiation.** Schematic representation of a possible role of SMN (in red) as regulator of translation initiation. After the formation of the 80S on the start codon, SMN controls the translation of the first 5 codons, as well as ribosome conformational changes in favour of a P/A hybrid state around the 5th codon, inducing a temporary ribosome stalling and possibly stabilizing the ribosomes before proceeding with the elongation phase.

4.4 Loss of SMN-specialized ribosomes impacts on active translation in SMA

To integrate the results obtained with ribosome profiling of SMN-specialized ribosomes with translation variations occurring in SMA, I combined SMN-RiboSeq data with Poly-RiboSeq and Active-RiboSeq on healthy and SMA-affected mouse brains. I first generated the meta-profiles of the transcripts enriched in SMN-specialized ribosomes employing Poly-RiboSeq and Active-RiboSeq data of both conditions (Figure 4.18).

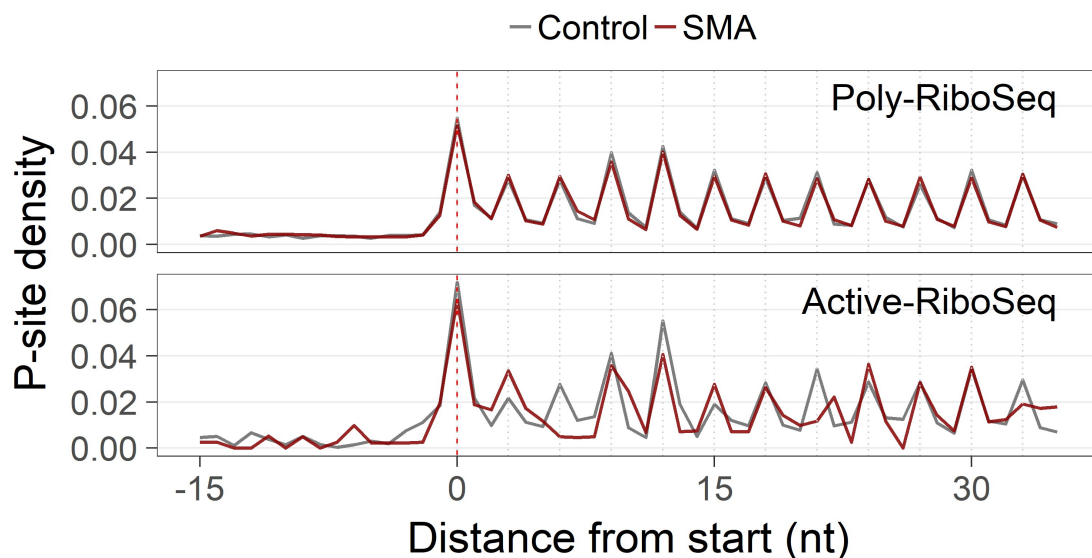


Figure 4.18. **Meta-profiles of mRNA enriched in SMN-specialized ribosomes.** Overlay meta-profiles around the start codon based on the P-site position of the reads along transcripts for control and SMA samples from Poly-RiboSeq and Active-RiboSeq assays, considering only the mRNAs enriched in SMN-specialized ribosomes.

Interestingly, in Poly-RiboSeq the profiles for the population of transcripts associated to SMN-specific ribosomes do completely overlap in control and SMA samples. Astonishingly, in Active-RiboSeq, where active ribosome protected fragments are considered, a large difference between the control and SMA metaprofile is clearly visible. In particular, the absence in the diseased sample of signal on the third codon of the coding sequence popped out, as well as the loss of a neat periodicity on downstream triplets. These findings point to a peculiar behaviour of actively translating ribosomes at the beginning of the coding sequence caused by loss of SMN (Figure 4.19). This conclusion also support the previous results concerning a possible role of SMN in regulating translation initiation and, consequently, ribosome drop-off

that might lead to the observed decrease in ribosome number in the axon of late symptomatic mice we recently observed²⁴⁹.

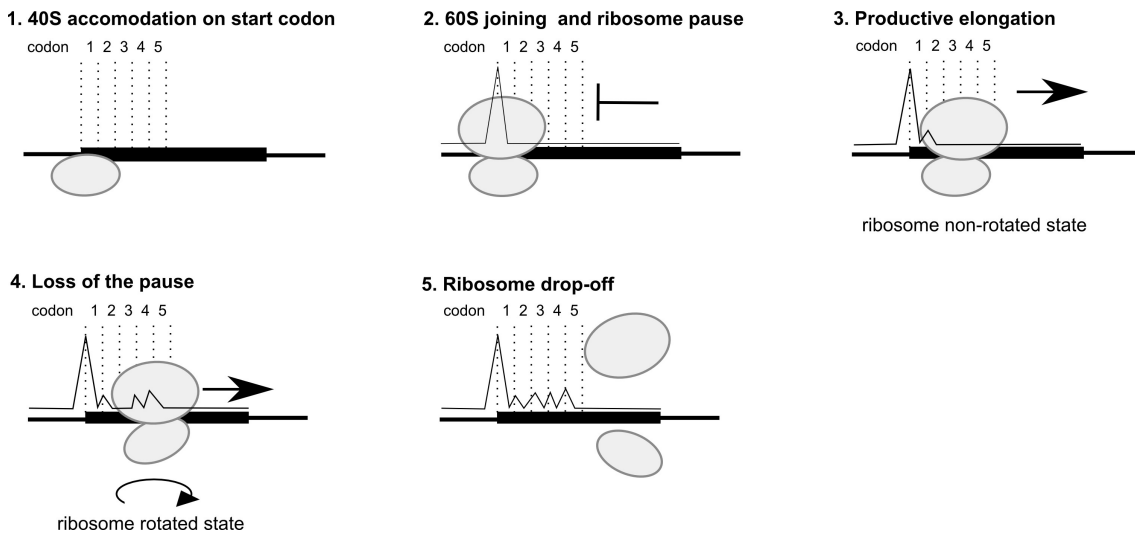


Figure 4.19. **Consequences of loss of SMN in translation initiation.** Schematic representation of translation initiation in absence of SMN. After the formation of the 80S on the start codon, the loss of SMN may induce ribosomes sliding on the first 5 codons and cause ribosomes drop-off after conformational changes at the 5th codon.

Then, to further investigate possible connections between actively translating ribosomes and SMN-specializes ribosomes, I compared SMN-RiboSeq data with Poly-RiboSeq and Active-RiboSeq on healthy and SMA-affected mouse brains. I first identified the lists of mRNAs up- and down-regulated in the diseased samples with respect to the control for both techniques. For convenience, I will refer to the considered pools of mRNAs as follows: SMN_enr (genes enriched in SMN-RiboSeq), pSMA_down (genes down-regulated in SMA from Poly-RiboSeq), pSMA_up (genes up-regulated in SMA from Poly-RiboSeq), aSMA_up (genes up-regulated in SMA from Active-RiboSeq) and pSMA_down (genes down-regulated in SMA from Active-RiboSeq). Intersection analysis of these lists is provided in (Figure 4.20A). Albeit most of the genes are exclusively present in one of the sets, some remarkable overlaps could be identified. In particular, the largest intersection includes genes shared by SMN_enr and aSMA_down lists. This means that the down-regulation of these genes in the early symptomatic SMA-affected mice, identified using Active-RiboSeq, may be caused by the absence of SMN that, in healthy conditions, is strongly associated with ribosomes translating them. Finally, I performed a term enrichment analyses (see Appendix) of the 5 groups against the EnrichR gene set libraries²⁸⁷, focusing on the sets related to motor neuron diseases (Figure 4.20B).

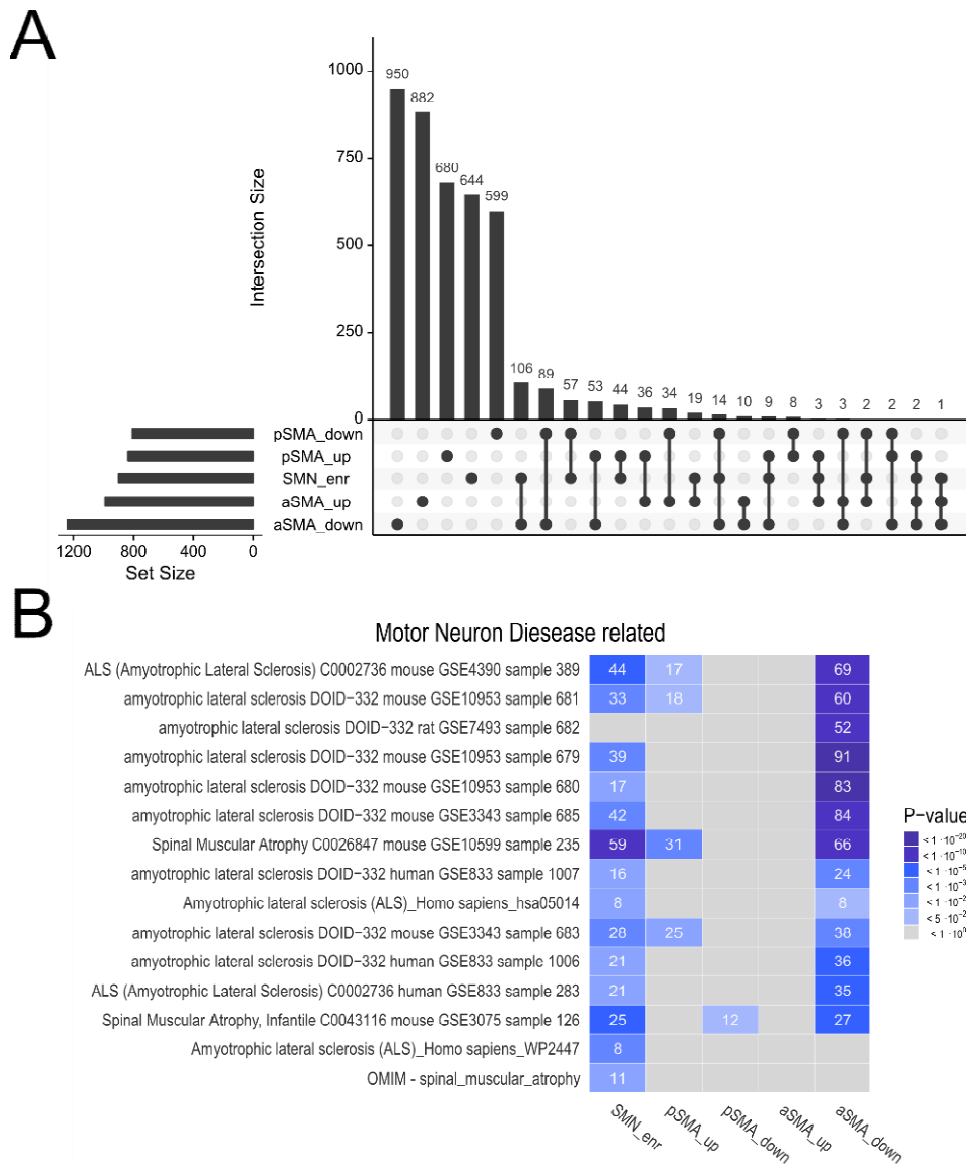


Figure 4.20. **Comparison between SMN RiboSeq and Poly and Active-RiboSeq performed on control and diseased mouse brains.** (A) The plot displays the relationship between the 5 considered sets: pSMA_down, pSMA_up, SMN_enr, aSMA_up and pSMA_down. The bottom left horizontal bars report the dimension of the corresponding set. The bottom section shows the combination matrix encoding the relationships between sets, aggregates and exclusive intersections: each column corresponds to an exclusive intersection that contains the elements of the sets listed on the left. The vertical bars reported the size of the exclusive intersection. (B) Results of the term enrichment analysis performed on the 5 groups and related to motor neuron diseased. The number in each box represent the number of gene associated to the terms listed on the left. The term of the last row comes from the OMIM database and the list of the 11 gene is reported in Table 4.1.

Strikingly, most of these sets were significantly enriched in the list of genes associated with in SMN-interacting ribosomes and down-regulated in the SMA sample from Active-RiboSeq, confirming a strong relationship between SMN and selective translation of a specific pool of transcript, which is disrupted in SMA disease. Remarkably, I could observe an enrichment for genes associated with Spinal Muscular Atrophy from the Online Mendelian Inheritance in Man database (OMIM) These 11 genes are listed in Table 4.1. Interestingly, this list includes *Smn1*, the gene encoding for SMN, confirming a potential autoregulation of this gene that was also suggested by Workman and collaborators²⁸⁸.

Gene name	Description
ARHGEF4	Rho guanine nucleotide exchange factor 4
CCND1	cyclin D1
CREB1	cAMP responsive element binding protein 1
CRHR1	corticotropin releasing hormone receptor 1
GNAI2	G protein subunit alpha i2
NCOR2	nuclear receptor corepressor 2
PRKACA	protein kinase cAMP-activated catalytic subunit alpha
RELA	RELA proto-oncogene, NF-kB subunit
SMN1	survival of motor neuron 1, telomeric
STAT3	signal transducer and activator of transcription 3
VAPB	VAMP associated protein B and C

Table 4.1. Genes associated with Spinal Muscular Atrophy from the Online Mendelian Inheritance in Man database (OMIM) enriched in SMN-specialized ribosomes.

4.5 Conclusions

Summarising, I pointed out the significant contribution of dedicated computational pipelines for the high-resolution analysis of ribosome profiling data, based on the extraction of positional information such as the identification of ribosome P-site within the reads and the detection of ribosome peaks along the transcripts. Firstly, I demonstrated the ability of Active-RiboSeq in capturing ribosome protected fragments, highlighting some emerging differences with respect to Poly-RiboSeq. Secondly, I observed a relative accumulation of mapping reads associated to actively translated ribosomes drop-off along the coding sequence in the SMA condition. Thirdly, I showed quantitative differences in the amount and in the position of ribosome peaks detected along the 3' UTR between control and SMA-affected mouse brains, confirmed by both Poly- and Active-RiboSeq, demonstrating a strong connection between transcripts harbouring these peaks and neuro-related functions.

Fourthly, investigating SMN RiboSeq data I identified a specific pool of mRNAs associated with SMN interacting ribosomes. This pool is enriched in neuro-related functions, RNA-related terms and genes responsible for the onset of SMA disease according to OMIM. I revealed a significant accumulation of long reads on the start codon and of out-of-frame short reads on the 5th codon of the CDS, pointing to conformational changes of ribosomes at the beginning of the coding sequence and suggesting a potential role of SMN in binding ribosomes in either translation initiation or the early phase of translation elongation. Finally, I demonstrated a remarkable relationship between SMN and actively translating ribosomes resulting in alterations of the meta-profile of transcripts enriched in SMN-specialised ribosomes in SMA sample from Active-RiboSeq. In addition, I showed a strong connection between genes associated with in SMN-interacting ribosomes and down-regulated in the SMA-affected mouse brains from Active-RiboSeq.

Overall, these findings show that the initiation, elongation and probably termination phase are altered in the diseased samples, suggesting a critical role of SMN in regulating polysome activity at multiple levels. Further experimental investigation may be driven by this conclusion, eventually resolving the clear mechanism connecting SMN and translation, and the dysregulation of this interaction in SMA.

5 Discussion

Translation is a fundamental biological process occurring in cells, carried out on mRNA molecules that can be bound by many ribosomes at a time (polyribosomes)^{16–18}. It is obviously the primary determinant in regulating protein expression in a wide range of physiological processes^{37–40}, including several neurological disorders^{47–52} and tumours^{41–46}. Recent findings demonstrated the existence of a wide collection of *cis*- (nucleotide composition of the mRNAs and their secondary structures)^{57,58} and *trans*- (ncRNAs and RNA binding proteins)^{59,60,62} factors, acting as translational regulatory mechanisms, which regulate the movement and position of ribosomes along the transcripts. Nevertheless, despite recent advances in dissecting the mechanism of translation¹²³, a complete characterization of polysome organization (for example the number and position of ribosomes along the transcript) and the functional controls directed in shaping cellular phenotypes is still lacking. For these reasons, I propose RiboWaves, an integrated bioinformatics suite that mixes experimental data (e.g. Atomic Force Microscopy images, Next-Generation Sequencing data) and computational assays (deterministic and stochastic modeling, pipelines for high-throughput data analysis) for a comprehensive understanding of translation regulation and polysome organizational rules governing the number of ribosome per polysome and ribosome localization along the mRNAs.

Mathematical model of translations

The first contribution of my work consisted in the development of a deterministic and a stochastic modeling module, *riboAbacus* and *riboSim*, to forecasts the number and the position of ribosome within polysomes, respectively.

The number of ribosomes per transcript, together with ribosome localization along the mRNA, is a fundamental aspect of a polysome at the steady state. I demonstrated that assessing the number of ribosomes within polysomes can clarify the impact of some regulatory elements in controlling translation. Nevertheless, at present the challenge in experimentally measuring the precise number of ribosome per transcript hindered the use of this parameter for translational studies as well as in mathematical and computational models. To overcome this problem, I took advantage of the number of ribosomes per transcript calculated from polysome images acquired by Atomic Force Microscopy⁹⁵. In principle, polysome profiling followed by microarray¹³⁷ and more recent high-throughput techniques such as ribosome profiling¹²³ might be employed to

experimentally deduce the projected number of ribosome per transcript. However, comparing the distribution of the number of ribosomes per transcript obtained by AFM in yeast with the data provided by Arava and collaborators¹³⁷ produced by polysome profiling and microarray analyses, revealed the higher accuracy of AFM images in computing the number of ribosome per polysome. On the other hand, ribosome profiling, based on deep sequencing of ribosome protected mRNA fragments^{123,150}, has been extensively used to measure ribosome density along mRNAs, typically starting from the total amount of reads mapping on the transcripts¹⁶⁹. Yet, ribosome density is not the precise calculation of the number of ribosomes per transcript. In fact, RiboSeq data are projection of reads from thousands of single mRNA molecules in a transcriptome and are typically very noisy. Therefore, a definitive approach for precisely localising single ribosomes along mRNA is missing and the calculation of the number of ribosome per polysome starting from RiboSeq data is still an open challenge full of pitfalls. Nevertheless, ribosome occupancy profiles provides meaningful positional information, often employed for obtaining precise information about translation at single nucleotide resolution^{141,142}.

Thus, I took advantage of AFM images of polysome and RiboSeq data as experimental benchmark for riboAbacus and riboSim, respectively. To have a comprehensive view of polysome organization, the number of ribosomes per transcript predicted by riboAbacus and the positional information supplied by riboSim were analysed in parallel. riboAbacus and riboSim were run using a progressive approach: starting from the simplest model I introduced at each step a new feature, i.e. the codon usage bias and the ramp.

The first significant finding emerged after the introduction in both models the codon usage bias^{79,80,114,118,125,127,289}. rA_Figure 4A (left panel)² shows that in human riboAbacus significantly increases the similarity between experimental and predicted profiles but still overestimates the number of ribosomes per transcript. This observation is supported by riboSim that in two mammals (mouse and human) does not display any improvement in the prediction of ribosome localization along mRNAs (Figure 2.6B), indicating that the codon usage bias alone is not sufficient to account for ribosome dynamics for high eukaryotes. In fact, the simulation performed in yeast by riboSim showed a significant increase of the correlation distribution towards positive values, indicating that the codon usage bias alone is able to enhance ribosome

² "rA_" stands for riboAbacus. If the figure number is preceded by this suffix, please refers to corresponding panel in "RiboAbacus: a model trained on polyribosome images predicts ribosome density and translational efficiency from mammalian transcriptomes" by Lauria et al., reported in section 2.1.

positional predictions in simple organism, partially confirming its connection with ribosome translocation^{253,290}, pauses^{97,98} and drop-offs^{99,100}.

I then introduced in both models an additional level of complexity provided by the 5' ramp i.e. a region at the beginning of the coding sequence showing high ribosome density and reduced elongation speed with respect to the remaining coding sequence^{81,82,121,253}. Its length ranges from 5 to 50 nucleotides^{82,127}. Several biological explanations have been proposed for the existence of the ramp, such as local codon usage bias^{113,122,127}, mRNAs secondary and tertiary structures^{116,127}, rapid initiation rates¹²⁹ and the concurrence of many of these causes^{81,115,128}. Gerashchenko and Gladyshev recently proposed that the ramp may be an artefact caused by the use of cycloheximide during the preparation of the samples²⁹¹, although it has been shown that an increased density ribosomes at the beginning of the coding sequence can be equally observed in datasets produced without cycloheximide treatment^{170,292}. Even though the precise nature of the ramp is still a matter of debate, I modelled the ramp effect by introducing two parameters that characterise this phenomenon: ramp length and ribosome slowdown rate. I optimised both parameters in human Hek-293 translome, obtaining the best fit with experimental data setting 50 codons of ramp length and 70% of ribosome slowdown rate (rA_Figure 3B). The ramp length is in agreement with data available in literature^{82,115,128,138} and exceeds the portion of mRNA that, according to recent literature, may be affected by cycloheximide-related biases: 8 codons²⁹³. My results also provide for the first time an estimation of the ribosome slowdown rate.

In addition, the cross-validation used for optimising the parameters and the two validations performed in MCF-7 and rabbit reticulocytes (rA_Figure 4, 5 and 6) highlights the importance of the ramp in determining the number of ribosomes per transcript in human. riboSim simulations after the introduction of the optimised ramp parameters proved that in complex biological systems its performance in predicting ribosome localization are not improved by a slowdown mechanism (Figure 2.7B). This result does not match with the findings of riboAbacus about the fundamental role of the ramp in providing improved predictions of the number of ribosomes per transcript in both mouse and human. Nevertheless, the inclusion of ramp parameters produces a significant improvement in riboSim predictions in yeast. A possible explanation of this difference between the two models in mammals could be that, while the overall number of ribosomes per transcript can be adequately estimated by few parameters, a larger set of features, still to be discovered, is required to properly model ribosome localization along mRNAs in more complex biological systems than yeast. In fact, my results clearly point to additional translational regulatory elements acting in higher

eukaryotes, suggesting different and more sophisticated strategies of translation regulation depending on the complexity of the species. This hypothesis is supported by the work by Harrison and collaborators²⁹⁴, which showed a widespread lack of “RNA interference” (namely RNA-driven controls of cellular processes) in *Saccharomyces cerevisiae*.

As a consequence, stochastic simulation of translation may require to model the presence of other *cis* and *trans* factors acting as translational regulators. For example, mRNA secondary structures³⁴, RBP³⁷ and ncRNAs binding ribosomes⁶² may have a crucial role in tuning ribosome movements and define their precise localization along the mRNAs.

Following the promising results obtained with riboAbacus, I tried to elucidate potential connections between the number of ribosomes per transcript and the total protein level in cells³. In previous years a general poor correlation between experimental measures of transcriptome (mRNA levels) and proteome (protein abundance) has been reported in several works in mammalian cells^{54,56}. I showed that taking into account the number of ribosomes per transcript, as RiboAbacus does, significantly increased the experimental correlation between transcriptome and proteome in 3 different datasets, especially when the ramp effect is included in the model. These findings demonstrate that up to 10% of protein levels can be explained by the number of ribosomes per transcript and support the crucial role of the slowdown mechanism at the beginning of the coding sequence in regulating translation and the final protein abundance.

In summary, I showed that the number of ribosomes per transcript and their localization are strongly-related elements that characterise polysomes. These parameters should be investigated in parallel for a comprehensive description of the translational machinery. The good level of predictions of the number of ribosomes per transcript already achieved by riboAbacus, coupled with the introduction in riboSim of local slowdowns, which represent an additional feature, makes the combination of the two models an optimal tool for the analysis of translation.

Undoubtedly, the features and organisms examined in this work represent only a starting point for a deeper and comprehensive understanding of translational regulatory mechanisms acting both in *cis* and in *trans*. Thus, features such as mRNA

³ For a more dilated discussion about the use of riboAbacus in predicting protein abundance please refer to "*RiboAbacus: a model trained on polyribosome images predicts ribosome density and translational efficiency from mammalian transcriptomes*" by Lauria et al., reported in section 2.1.

secondary structures RBP interactions and should be investigated in a broader range of species. I will integrate riboAbacus and riboSim adding the two abovementioned features taking advantage of SHAPE and CLIP data for mRNA secondary structures and RNA binding protein, respectively. Then, to unearth potential differences in ribosome number and localisation throughout evolution, I will estimate the amount and the position of ribosomes along mRNAs for an extended collection of organisms. Finally, I will compare the outcomes of the models to experimental data from different conditions (e.g. physiological, pathological or stress conditions) to find a direct connection between the considered features and possible alterations in ribosome arrangement within polysomes.

From a computational point of view, to provide extended simulations of translation and lighten the current assumptions, riboAbacus and riboSim can be improved including the initiation phase (e.g. taking into account the scanning of the 5' UTR by the ribosomal small subunit 40S and the termination steps). Finally, I will integrate riboAbacus and riboSim to combine the advantages of the two models e.g. the speed of the deterministic approach and the possibility to easily incorporate multiple features of the stochastic one.

Concluding, my results pinpoint the importance of a double approach based on deterministic and stochastic models for better understanding the role of translational regulatory elements in tuning polysome organization throughout evolution.

Analysis of ribosome profiling data, and application to unravel at single nucleotide resolution the mechanism leading to translational defects in spinal muscular atrophy

I previously discussed the widespread diffusion of RiboSeq for the study of translation with unprecedented resolution^{123,150} and for the extraction of detailed information about position and fluxes of ribosomes along the mRNAs¹⁴⁰⁻¹⁴² at transcript-level^{126,163,166}. Much of this information relies on the ability to determine the exact localization of the P-site within ribosome protected fragments (reads). The P-site offset is of crucial importance for a wide range of RiboSeq analyses such as verifying the trinucleotide periodicity of the ribosome along the coding sequence^{123,180}, derive accurately estimations of codon usage bias and translation pauses^{170,184} and reveal novel translated regions in known protein coding transcripts or ncRNAs¹⁶⁵⁻¹⁶⁷. Despite the many efforts aimed at dealing with ribosome profiling data, some aspects such as statistical procedures for the extraction of meaningful positional information still need to be computationally addressed. In particular, significant read accumulations along the mRNAs may be related to ribosome slowdown^{82,253} and ribosome stalling^{97,254}, two

scenarios connected to many pathologies such as neurodegenerative diseases^{47,255}, diabetes and multi-systemic failure²⁵⁶.

Recently, many biases introduced by the alignment and the preprocessing of ribosome profiling fragments have been discussed^{190,191}. For example, ambiguous reads mapping to mRNA isoforms, missing normalizations and alignment of selected subsets of reads¹⁹³⁻¹⁹⁵ may lead to very noisy and misleading occupancy profiles, making it difficult to identify regions truly associated to ribosome pauses and slowdowns. Albeit a few procedures were proposed to improve data analysis^{61,170,196}, a conclusive approach for the extraction of meaningful biological information is still missing.

Overall, these considerations prompt the development of riboWaltz and riboScan, two computational tools for accurate analyses of ribosome profiling data. More in detail, riboWaltz is an R package aimed at identifying the optimal P-site position within the reads, while riboScan is a pipeline dedicated to the extraction of those regions of the mRNAs with a statistically significant enrichment of ribosome protected fragments (namely ribosome hot-spots). In this section I will discuss the results obtained applying the two pipelines in a case study, which is ribosome profiling assays performed on healthy and SMA-affected mouse brains as well as SMN-specialized ribosomes.

Spinal muscular atrophy (SMA) is a neuromuscular disease caused by genetic alterations of the Survival of Motor Neuron gene (*Smn*) that induce the production of low level of SMN protein²⁴³. Recent findings demonstrated the relationship between SMN and the translational machinery²⁷² and its association to polysomes *in vivo*²⁴⁵ and *in vitro*²⁴⁶. Moreover, the Laboratory of Translational Architectomics (IBF-CNR, Trento) has demonstrated that SMN is tightly associated to ribosomes, highlighting a still unknown role of this protein in the cytoplasm and translation. Nevertheless, a clear mechanism connecting SMN and translation has not yet been established.

To identify possible mislocalization of ribosomes enrichments along transcripts in early-symptomatic SMA brains, I took advantage of two ribosome profiling techniques, one collecting ribosome footprints from ribosomes obtained from polysomal fractions (Poly-RiboSeq)¹⁵⁹, the other selecting actively translating ribosomes in polysomal fractions (Active-RiboSeq) by using a new technology called RiboLaceTM, developed by IMMAGINA Biotechnology. First of all I applied riboWaltz to all the datasets to assess that both Poly-RiboSeq and Active-RiboSeq in the two conditions are capturing ribosome protected fragments on the coding sequence. For the first time, the ability of Active-RiboSeq in extracting *bona-fide* ribosome protected fragments was demonstrated by verifying the enrichment of ribosome P-sites along the CDS with

respect to UTRs (Figure 4.5), and the presence of a clear trinucleotide periodicity^{123,263} confined in the CDS of transcripts (Figure 4.6). The meta-profiles in Figure 4.7 revealed an increased signal on the 5th codon, especially in the Active-RiboSeq samples. A signal enrichment in the same position have been already observed in several papers^{263,278,279} and accurately investigated by Han and collaborators²⁷⁷, that attributed the early ribosome pause to the geometry of their exit tunnel for productive translation. Interestingly, they claim that a pause in this position is necessary to reduce the risk of further ribosome drop-off, resulting in incomplete translation events. In light of this hypothesis, the presence of a peak on the 5th codon can be considered a positive indicator of successful translation.

I also explored the potential connection between this result and the nucleotide composition of mRNAs performing a sequence enrichment analysis around the 5th codon for the Active-RiboSeq assay. In accordance to the results obtained by Han and collaborators²⁷⁷, no statistically significant enriched sequence emerged, leading to the conclusion that the nucleotide composition of the mRNAs do not account for a slowdown of the ribosomes at the beginning of the coding sequence in early symptomatic mice. This result suggest that other mechanisms or molecules might be responsible for this highly specific pause.

Following the approaches used by Han and collaborators²⁷⁷ for the computational analysis of this phenomena, I computed the ratio between the signal on the first 5 codons and the remaining part of the coding sequence for all the selected transcript. Figure 4.8 shows a significant increment of the ratio distribution in Active-RiboSeq with respect to Poly-RiboSeq for both the conditions and, interestingly, an increase in the SMA sample with respect to the control only for actively translating ribosomes. The difference between the two techniques may be associated to the presence of non-translating stalled ribosomes^{97,254,295} along the CDS, that are particularly found in neuronal tissues, that are captured by Poly-RiboSeq protocols but not by the RiboLaceTM technology. The increased ratio of active ribosomes detected in the SMA sample may also be attributes to lower signal on the CDS caused by ribosome drop-off, an event that has been widely investigated from both experimental^{99,100,296} and computational point of views^{86,87,230,297}. Another explanation may be the absence of local but functional ribosome slowdowns induces by SMN, similarly to what have been observed for FMRP in the context of the Fragile X Syndrome⁴⁷.

To better investigate whether a potential mislocalization of ribosomes along the mRNAs at this stage of the disease is connected to the loss of the physiological role exerted by SMN in translation I employed riboScan. I detected statistically significant

enriched regions in ribosomes (both hot-spots and peaks) and searched for any difference between control and SMA affected mice.

Interestingly, the analysis of the peaks detected along the coding sequence for Active-RiboSeq (Figure 4.9) support the hypothesis of ribosome drop-off caused by loss of SMN. In fact, I demonstrated the presence of a reduced number of peaks and an increase in the distance between consecutive peaks in the SMA sample, in agreement with ribosome drop-off during the elongation phase.

In addition, I revealed strong connections between significant enriched regions along the 3' UTR and the development of the nervous system in early symptomatic mice (Figure 4.10 and Figure 4.11). In absence of further experimental validation, this result should be taken with caution, yet it leads to some very interesting causative hypothesis: readthrough of the stop codon^{89,280}, non-translating ribosomes sliding on the 3' UTR and even the presence of reads associated to clusters of RNA binding proteins. The first two hypotheses may lead to the production of non-functional proteins⁸⁹ and short peptide produced by post-termination ribosomes²⁹⁸ after translation reinitiation in the 3' UTR^{299,300}, respectively. Both possibilities point to the observed ribosome enrichment along the 3' UTR as a crucial determinant in the development of SMA, potentially related to defects in the termination phase of translation. To test this hypothesis in the Laboratory of Translational Architectomics (IBF-CNR, Trento) experimental validations using cell lines depleted by SMN³⁰¹ and stop codon readthrough assays³⁰² have been planned.

Even though the unexpected increase in 3'UTR signal in SMA has no clear explanation, I performed an enrichment analysis with Gene Ontology terms, KEGG and REACTOME pathways on the subpopulation of genes presenting enriched regions on 3' UTR of the SMA sample (Figure 4.13). Strikingly, the results show a sizeable set of biological terms associated to neuro-development, cell signalling, dendrite and axon guidance mechanisms. Remarkably, many terms supports previous observations of SMN-related defects in the transport of mRNA and RBPs³⁰³⁻³⁰⁵, components of the translational machinery²⁴⁴ and β -actin mRNA in the growth cone of SMA-affected motor neurons³⁰⁶.

An emerging hypothesis connecting the observed ribosome enrichments along the 3' UTR and ribosome drop-off in SMA-affected mice is the lack of SMN-driven controls of translation. Lower level of SMN may induce a dysregulation of both translation elongation, termination and possibly ribosome recycling, in line with recent studies suggesting the existence of specialized ribosomes^{68,69} that can be post-translationally modified⁵⁷.

To deeper examine the role of SMN in translation in healthy mouse brains, I analysed data obtained by SMN-specific RiboSeq performed after sub-fractionation of ribosomes followed by immunoprecipitation of SMN-specialized ribosomes. A pool of mRNAs enriched in SMN-specialized ribosomes with respect to IgG-aspecifically associated transcripts was identified and its connection with neuro-related processes demonstrated (Figure 4.14). For mRNAs enriched in SMN-specialized ribosomes, I showed the presence of two distinct populations of reads (Figure 4.15A): long (32-35 nucleotides) and short (23-26 nucleotides), that has been already associated to two different ribosome conformations^{23,171,285} characterizing different stages of translation. Moreover, an accumulation of mapping reads along the coding sequence emerges (Figure 4.15B) although it is not associated to a trinucleotide periodicity (Figure 4.15C and D). This observation, coupled with the high ratio between the P-sites on the first 5 codons and the P-site on remaining part of the coding sequence (median >5) displayed in Figure 4.15E, points to a major role of SMN in modulating polysomes activity at the very beginning of the translation of a transcript by directly controlling ribosomes at the translation initiation phase and during the elongation stage of the very first codons.

Further confirmations of this hypothesis arise from the analysis of the two populations of reads (long and short) associated to SMN-specialized ribosomes (Figure 4.16A and B). By a meta-gene analysis, I demonstrated that the longer reads mainly align on the translation start site, while the shorter reads show accumulation around the 5th codon of the coding sequence with a -1 frameshift. The different but exact localization of long and short reads clearly points to multiple conformations of SMN-interacting ribosomes along the CDS. In particular, the -1 frameshift associated to short reads around the 5th codon suggests a rotated state of ribosomes due to the proximity of the tRNAs accommodated in the A and P-site (Figure 4.16C), as already discussed by Matsuo and colleagues²³ and Sulima and collaborators²⁸⁵. From these results, an hypothesis of the mechanistic role of SMN in controlling translation initiation has been outlined (see Figure 4.17): SMN controls the translation of the first 5 codons and the stabilization of the ribosome conformation in a P/A hybrid state on the 5th codon, inducing a temporary ribosome stalling before proceeding with the productive elongation of the rest of the transcript. This hypothesis is also supported by the meta-gene analysis of the transcripts enriched in SMN-specialised ribosomes (Figure 4.18). In fact, meta-profiles of Active-RiboSeq data show a lack of reads on the 3rd codon of the coding sequence in absence of SMN with respect to the control, and the loss of the trinucleotide periodicity on the CDS. These findings confirm an altered dynamics of actively translating ribosomes at the beginning of the coding sequence caused by SMA i.e. when SMN expression is lost (Figure 4.19).

Remarkably, the possible contribution of SMN in regulating translation initiation coupled with the presence of significant enriched regions on the 3' UTR also point to an explored role of SMN as a major determinant in ribosome recycling^{12,24,25,27,286} in physiological conditions.

Finally, by intersecting the SMN-specific ribosome profiling data with Poly-RiboSeq and Active-RiboSeq on healthy and SMA-affected mouse brains, an intriguing link between SMN enriched genes and down-regulated genes in Active-RiboSeq SMA data emerged (Figure 4.20A). The results of term enrichment analyses showed that, focusing on motor neuron diseases-related terms (Figure 4.20B), most of them are connected to genes enriched in SMN and down-regulated in the SMA sample in Active-RiboSeq but not in Poly-RiboSeq. This confirms a strong relationship between the expression of SMN and positive translation of a specific pool of transcript, which is disrupted in SMA disease possibly due to SMN loss. Remarkably, the SMN enriched genes are also associated to term specifically related to the spinal muscular atrophy from the OMIM database. This set includes *smn1* (see Table 4.1), i.e. the gene that encodes for SMN, indicating a possible auto-regulatory feedback control of SMN expression, already discussed by Workman and collaborators²⁸⁸.

In summary, the present study shows how the most classical computational assays for quantitative examination of ribosome profiling data (e.g. enrichment analysis) can be optimally integrated with original and dedicated pipelines focused on the positional aspect of the analyses. Applying this tools to the investigation of RiboSeq assays performed on healthy and SMA-affected mouse at the early-symptomatic stage of the disease, I discovered a possible mechanism of action of SMN that is in nice accordance with previous observation about possible SMN-related dysregulations of local protein synthesis in neurons³⁰⁷⁻³⁰⁹. Indeed, lack of controls at very first initial phases of translation may be also the cause of the hypothesised ribosome drop-off and stop codon readthrough observed in SMA-affected mouse brains, reinforcing the evidences indicating SMN as a possible regulator of ribosome and polysomes activity.

Despite the many hints provided by riboWaltz and riboScan to understand translational defects in SMA, the biological factors explaining the abovementioned ribosome stalling, slowdowns and readthrough can be identified only through additional investigations. To this aim, I would like to further examine the observed accumulation of SMN-specialised ribosome around the 5th codon of the coding sequence: I will look for recurring motifs in the nucleotide sequences immediately downstream the 5th codon of the CDS associated to mRNA secondary structures that may induce ribosome slowdown.

Additional sequence enrichment analyses are also required to explore the hypothesis of SMN-related defects in the transport of mRNAs. I will look for binding motifs within the 3' UTR of the transcripts enriched in SMN-specialised ribosomes known to be associated to RBPs involved in axonal mRNA transport.

A critical point of this work consists in the hot-spots and peaks detected along the 3' UTR that must be further investigate to either confirm or disprove the presence of ribosomes downstream the coding sequence. For this reason, I will look for the trinucleotide periodicity on the 3' UTR employing only the reads contributing to the SMA specific peaks detected in this region. This can be approached by computing the percentage of P-sites corresponding to the three reading frames, generating the meta-profile of the transcripts showing at least one peak on the 3' UTR uniquely detected in the SMA sample. Moreover, to investigate the role of the 3' UTR in the development of SMA and possible defects in the termination phase of translation experimental validations using cell lines depleted by SMN and stop codon readthrough assays have been planned in my Lab.

Finally, to enhance the previous analyses and detail possible mechanisms of the disease I will progressively reduce the quantity of examined data. First, I will identify specific subsets of transcripts, e.g. showing significant differences in the number or localization of reads, hot-spots and peaks along the mRNAs in SMA-affected mice with respect to the healthy ones. Second, I will further narrow the investigation by single-transcript analyses.

Concluding, although experimental validations of the obtained results are required, the present work provides a new integrated scenario for better understanding translation and a first step for paving the way for understanding fine alteration of translation in diseases.

Appendix

Poly-RiboSeq and SMN-specific RiboSeq

Cytoplasmic lysates were prepared pulverizing frozen mouse brains in a pestle and mortar cooled by liquid nitrogen and dissolving the powder in lysis buffer (10 mM NaCl, 10 mM MgCl₂, 10 mM Tris-HCl pH 7.5, 1% Triton-X100, 1% NaDeoxycholate, 0.6 U/μL RNase inhibitor, 1 mM dithiothreitol, 200 μg/ml cycloheximide, 0.005 U/ μL DNase I). Lysates were clarified by two following centrifugations.

For Poly-RiboSeq the brain lysates were then loaded in a 15-50% linear sucrose gradient and polysomes separated by ultracentrifugation. Polysomal fractions were pooled and digested with Rnase I for 2h at 4°C. Ribosome Protected Fragments (RPFs) were extracted with acid-phenol:chloroform:isoamylalcohol and isopropanol precipitation.

For SMN-specific RiboSeq the concentration of NaCl of the brain lysates were increased to 150 mM and the lysates were digested with RNase I for 45 min at RT by gentle agitation. The digested lysates were then centrifuged for 67 min at 100000 rpm at 4°C in an ultracentrifuge. The pellets were resuspended in resuspension buffer (150 mM NaCl, 10 mM MgCl₂, 10 mM Tris-HCl pH 7.5, 1% Triton-X100, 1% NaDeoxycholate, 0.6 U/μL RNase inhibitor, 1 mM dithiothreitol, 200 μg/ml cycloheximide, 0.005 U/ μL DNase I) and incubated with 2 micrograms of SMN antibody (BD Bioscience, cod. 610646) for 1.40h at 4°C on rotating wheel. Protein G Dynabeads (ThermoFisher Scientific) were then added and the solutions incubated for 1h at 4°C on a rotating wheel. The beads were then wash twice on magnet to remove unbound proteins/ribosomes and the SMN-specific RPFs were extracted by Trizol.

The extracted RPFs (both from Poly-RiboSeq and SMN-specific RiboSeq) were size selected (25-35 nt) by UREA Page, treated to remove 3' phosphate, ligated to the 3' adaptor and purified again by UREA page. Ligated RPFs were retrotranscribed and the cDNA circularized. Libraries were finally amplified by PCR using Illumina PCR Primer Index. Libraries were sequenced on an Illumina HiSeq2500.

Ribosome profiling analysis

Ribosome profiling samples from healthy and SMA-affected mouse brains were sequenced with the Illumina HiSeq 2000 platform. Raw reads were processed by removing 5' adapters CTGTAGGCACCATCAAT, discarding reads shorter than 20

nucleotides and trimming the first nucleotide (using Trimmomatic v0.36). The reads mapping on the collection of mouse rRNAs and tRNAs (downloaded from the SILVA rRNA and the Genomic tRNA databases respectively) were removed. The remaining reads were aligned to the mouse transcriptome using the Gencode M6 transcript annotation, based on ENSEMBL version 81 and on GRCm38 genome reference. All reads aligning to the very same region were collapsed to avoid potential PCR duplicates, and only strand-specific reads were kept. All the alignments were performed with Bowtie2 (v2.2.6) employing the default settings.

Differential translation and annotation enrichment analyses

The analysis of differential expression between healthy and diseased mice and techniques (Poly- and Active-RiboSeq) were performed by differential expression tests provided by the EdgeR package³¹⁰. Annotation enrichment analyses were performed using the clusterProfiler³¹¹ package and the EnrichR tool³¹². Table A1 displays the number of reads left after each step of the alignment.

Conditon	Technique	Replica	Number of reads (x10 ⁶)				
			Initial amount	Clipping / trimming	rRNA align.	tRNA align.	mRNA align.
Healthy	Poly RiboSeq	1	58.23	56.30	5.55	5.53	2.63
		2	125.94	113.64	14.07	14.03	10.62
	Actve RiboSeq	1	47.96	32.91	5.24	4.66	2.09
		2	83.88	75.94	29.27	5.32	0.86
	SMN specific RiboSeq	1	49.29	40.65	7.41	7.29	2.57
		2	33.47	29.77	4.01	3.96	1.26
SMA early symptomatic	Poly RiboSeq	1	80.65	77.64	7.76	7.73	5.30
		2	111.92	105.72	11.09	11.06	7.82
	Actve RiboSeq	1	87.59	69.87	28.72	5.45	0.37
		2	90.19	75.51	33.81	5.95	0.38

Table A1. Number of reads from RiboSeq assays at each step of the alignment process.

Bibliography

1. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–3 (1970).
2. Matera, A. G., Terns, R. M. & Terns, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. cell Biol.* **8**, 209–220 (2007).
3. Deniz, E. & Erman, B. Long noncoding RNA (lincRNA), a new paradigm in gene expression control. *Funct. Integr. Genomics* 1–9 (2016).
4. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
5. Gradi, A. *et al.* A novel functional human eukaryotic translation initiation factor 4G. *Mol. Cell. Biol.* **18**, 334–42 (1998).
6. Matsuo, H. *et al.* Structure of translation factor eIF4E bound to m7GDP and interaction with 4E-binding protein. *Nat. Struct. Biol.* **4**, 717–724 (1997).
7. Wells, S. E., Hillner, P. E., Vale, R. D. & Sachs, A. B. Circularization of mRNA by Eukaryotic Translation Initiation Factors. *Mol. Cell* **2**, 135–140 (1998).
8. Tarun, S. Z. & Sachs, A. B. Binding of eukaryotic translation initiation factor 4E (eIF4E) to eIF4G represses translation of uncapped mRNA. *Mol. Cell. Biol.* **17**, 6876–86 (1997).
9. López-Lastra, M., Rivas, A. & Barría, M. I. Protein synthesis in eukaryotes: The growing biological relevance of cap-independent translation initiation. *Biological Research* **38**, 121–146 (2005).
10. Sonenberg, N., Trachsel, H., Hecht, S. & Shatkin, a J. Differential stimulation of capped mRNA translation in vitro by cap binding protein. *Nature* **285**, 331–333 (1980).
11. Craig, A. W., Haghghat, A., Yu, A. T. & Sonenberg, N. Interaction of polyadenylate-binding protein with the eIF4G homologue PAIP enhances translation. *Nature* **392**, 520–3 (1998).
12. Becker, T. *et al.* Structural basis of highly conserved ribosome recycling in eukaryotes and archaea. *Nature* **482**, 501–6 (2012).
13. Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192**, 1227–1232 (1961).
14. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**, 229–64 (1992).
15. Sengupta, S. & Higgs, P. G. Pathways of Genetic Code Evolution in Ancient and Modern Organisms. *Journal of Molecular Evolution* **80**, 229–243 (2015).
16. Warner, J. R., Rich, A. & Hall, C. E. Electron microscope studies of ribosomal clusters synthesizing hemoglobin. *Science (80-)*. **138**, 1399–1403 (1962).
17. Palade, G. E. A small particulate component of the cytoplasm. *J. Biophys. Biochem. Cytol.* **1**, 59–68 (1955).
18. Wettstein, F. O., Staehelin, T. & Noll, H. Ribosomal aggregate engaged in protein

- synthesis: characterization of the ergosome. *Nature* **197**, 430–435 (1963).
19. Hinnebusch, A. G. The Scanning Mechanism of Eukaryotic Translation Initiation. *Annu. Rev. Biochem.* **83**, 779–812 (2014).
 20. Jackson, R. J., Hellen, C. U. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **324**, 113–127 (2010).
 21. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11**, 113–127 (2010).
 22. Rodnina, M. V, Fischer, N., Maracci, C. & Stark, H. Ribosome dynamics during decoding. *Phil. Trans. R. Soc. B* **372**, 20160182 (2017).
 23. Matsuo, Y. *et al.* Ubiquitination of stalled ribosome triggers ribosome-associated quality control. *Nat. Commun.* **8**, 159 (2017).
 24. Dever, T. E. & Green, R. The elongation, termination, and recycling phases of translation in eukaryotes. *Cold Spring Harb. Perspect. Biol.* **4**, 1–16 (2012).
 25. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. Termination and post-termination events in eukaryotic translation. *Advances in Protein Chemistry and Structural Biology* **86**, 45–93 (2012).
 26. Beznosková, P., Wagner, S., Jansen, M. E., Von Der Haar, T. & Valášek, L. S. Translation initiation factor eIF3 promotes programmed stop codon readthrough. *Nucleic Acids Res.* **43**, 5099–5111 (2015).
 27. Nürenberg, E. & Tampé, R. Tying up loose ends: ribosome recycling in eukaryotes and archaea. *Trends Biochem Sci* **38**, 64–74 (2013).
 28. Mathews, M. B. & Hershey, J. W. B. The translation factor eIF5A and human cancer. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1849**, 836–844 (2015).
 29. Rossi, D., Kuroshu, R., Zanelli, C. F. & Valentini, S. R. eIF5A and EF-P: Two unique translation factors are now traveling the same road. *Wiley Interdisciplinary Reviews: RNA* **5**, 209–222 (2014).
 30. Gutierrez, E. *et al.* eif5A promotes translation of polyproline motifs. *Mol. Cell* **51**, 35–45 (2013).
 31. Ryazanov, A. G., Shestakova, E. A. & Natapov, P. G. Phosphorylation of elongation factor 2 by EF-2 kinase affects rate of translation. *Nature* **334**, 170–3 (1988).
 32. Dever, T. E. *et al.* Phosphorylation of initiation factor 2 α by protein kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. *Cell* **68**, 585–596 (1992).
 33. Baird, T. D. & Wek, R. C. Eukaryotic initiation factor 2 phosphorylation and translational control in metabolism. *Adv. Nutr.* **3**, 307–21 (2012).
 34. Loreni, F., Mancino, M. & Biffo, S. Translation factors and ribosomal proteins control tumor onset and progression: how? *Oncogene* **33**, 2145–56 (2014).
 35. Russell, J. B. & Cook, G. M. Energetics of bacterial growth: balance of anabolic and catabolic reactions. *Microbiol. Rev.* **59**, 48–62 (1995).
 36. Buttgereit, F; Brand, M. D. A hierarchy of ATP-consuming processes in mammalian cells.

- Biochem. J.* **312** (Pt 1, 163–167 (1995).
37. Kondrashov, N. *et al.* Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. *Cell* **145**, 383–397 (2011).
 38. Piccirillo, C. A., Bjur, E., Topisirovic, I., Sonenberg, N. & Larsson, O. Translational control of immune responses: from transcripts to translatoemes. *Nat. Immunol.* **15**, 503–511 (2014).
 39. Kandel, E. R., Dudai, Y. & Mayford, M. R. The molecular and systems biology of memory. *Cell* **157**, 163–186 (2014).
 40. Fioriti, L. *et al.* The Persistence of Hippocampal-Based Memory Requires Protein Synthesis Mediated by the Prion-like Protein CPEB3. *Neuron* **86**, 1433–1448 (2015).
 41. Lasko, P. mRNA localization and translational control in Drosophila oogenesis. *Cold Spring Harbor Perspectives in Biology* **4**, (2012).
 42. Sonenberg, N. & Hinnebusch, A. G. New Modes of Translational Control in Development, Behavior, and Disease. *Mol. Cell* **28**, 721–729 (2007).
 43. Rosenwald, I. B. Upregulated expression of the genes encoding translation initiation factors eIF-4E and eIF-2alpha in transformed cells. *Cancer Lett.* **102**, 113–123 (1996).
 44. Gingras, A. C., Kennedy, S. G., O’Leary, M. A., Sonenberg, N. & Hay, N. 4E-BP1, a repressor of mRNA translation, is phosphorylated and inactivated by the Akt(PKB) signaling pathway. *Genes Dev.* **12**, 502–513 (1998).
 45. Calkhoven, C. F., Müller, C. & Leutz, A. Translational control of gene expression and disease. *Trends in Molecular Medicine* **8**, 577–583 (2002).
 46. Hershey, J. W. B., Sonenberg, N. & Mathews, M. B. Principles of translational control: An overview. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).
 47. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247–261 (2011).
 48. Ling, S. C., Polymenidou, M. & Cleveland, D. W. Converging mechanisms in als and FTD: Disrupted RNA and protein homeostasis. *Neuron* **79**, 416–438 (2013).
 49. Scheper, G. C., van der Knaap, M. S. & Proud, C. G. Translation matters: protein synthesis defects in inherited disease. *Nat. Rev. Genet.* **8**, 711–723 (2007).
 50. Ding, Q., Markesbery, W. R., Chen, Q., Li, F. & Keller, J. N. Ribosome Dysfunction Is an Early Event in Alzheimer’s Disease. *J. Neurosci.* **25**, 9171–9175 (2005).
 51. Vilotti, S. *et al.* Parkinson’s disease DJ-1 I166p alters rRNA biogenesis by exclusion of TTRAP from the nucleolus and sequestration into cytoplasmic aggregates via TRAF6. *PLoS One* **7**, (2012).
 52. Lee, J. *et al.* Dysregulation of upstream binding factor-1 acetylation at K352 is linked to impaired ribosomal DNA transcription in Huntington’s disease. *Cell Death Differ.* **18**, 1726–35 (2011).
 53. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).

54. Tebaldi, T. *et al.* Widespread uncoupling between transcriptome and translome variations after a stimulus in mammalian cells. *BMC Genomics* **13**, 220 (2012).
55. Kuersten, S., Radek, A., Vogel, C. & Penalva, L. O. F. Translation regulation gets its 'omics' moment. *Wiley Interdisciplinary Reviews: RNA* **4**, 617–630 (2013).
56. Vogel, C. *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *TL - 6. Mol. Syst. Biol.* **6 VN-re**, 400 (2010).
57. Xue, S. & Barna, M. Cis-regulatory RNA elements that regulate specialized ribosome activity. *RNA Biol.* **12**, 1083–7 (2015).
58. Gebauer, F., Preiss, T. & Hentze, M. W. From cis-regulatory elements to complex RNPs and back. *Cold Spring Harb. Perspect. Biol.* **4**, 1–14 (2012).
59. Kraushar, M. L. *et al.* Temporally defined neocortical translation and polysome assembly are determined by the RNA-binding protein Hu antigen R. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E3815–24 (2014).
60. Friend, K. *et al.* A conserved PUF–Ago–eEF1A complex attenuates translation elongation. *Nat. Struct. Mol. Biol.* **19**, 176–183 (2012).
61. O'Connor, P. B. F., Andreev, D. E. & Baranov, P. V. Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat. Commun.* **7**, 12915 (2016).
62. Pircher, A., Bakowska-Zywicka, K., Schneider, L., Zywicki, M. & Polacek, N. An mRNA-Derived Noncoding RNA Targets and Regulates the Ribosome. *Mol. Cell* **54**, 147–155 (2014).
63. Ben-Shem, A. *et al.* SOM: The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**, 1524–9 (2011).
64. Anger, A. M. *et al.* Structures of the human and Drosophila 80S ribosome. *Nature* **497**, 80–85 (2013).
65. Schluenzen, F. *et al.* Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Å Resolution. *Cell* **102**, 615–623 (2000).
66. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution. *Science (80-.)*. **289**, 905–920 (2000).
67. Melnikov, S. *et al.* One core, two shells: bacterial and eukaryotic ribosomes. *Nat. Struct. Mol. Biol.* **19**, 560–567 (2012).
68. Mauro, V. P. & Edelman, G. M. The ribosome filter redux. *Cell Cycle* **6**, 2246–2251 (2007).
69. Xue, S. & Barna, M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat. Rev. Mol. Cell Biol.* **13**, 355–369 (2012).
70. Simsek, D. *et al.* The Mammalian Ribo-interactome Reveals Ribosome Functional Diversity and Heterogeneity. *Cell* **169**, 1051–1065.e18 (2017).
71. Shi, Z. *et al.* Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of

- mRNAs Genome-wide. *Molecular Cell* (2016). doi:10.1016/j.molcel.2017.05.021
72. Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.* **84**, 165–198 (2015).
 73. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 178–189 (2015).
 74. Lee, D. Y., Hayes, J. J., Pruss, D. & Wolffe, A. P. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73–84 (1993).
 75. Kwon, J., Morshead, K. B., Guyon, J. R., Kingston, R. E. & Oettinger, M. a. Histone acetylation and hSWI/SNF remodeling act in concert to stimulate V(D)J cleavage of nucleosomal DNA. *Mol. Cell* **6**, 1037–48 (2000).
 76. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5' - untranslated regions of eukaryotic mRNAs. *Science (80-.)*. **352**, 1413–1416 (2016).
 77. Gray, N. K. & Hentze, M. W. Regulation of protein synthesis by mRNA structure. *Mol. Biol. Rep.* **19**, 195–200 (1994).
 78. Sonenberg, N. & Hinnebusch, A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* **136**, 731–745 (2009).
 79. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.* **9**, 213 (1981).
 80. Lynn, D. J., Singer, G. A. C. & Hickey, D. A. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* **30**, 4272–7 (2002).
 81. Dana, A. & Tuller, T. Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells. *PLoS Comput. Biol.* **8**, (2012).
 82. Li, Q. & Qu, H. Q. Human Coding Synonymous Single Nucleotide Polymorphisms at Ramp Regions of mRNA Translation. *PLoS One* **8**, (2013).
 83. Xue, S. *et al.* RNA regulons in Hox 5 [prime] UTRs confer ribosome specificity to gene regulation. *Nature* **517**, 33–38 (2015).
 84. Noller, H. F. Evolution of protein synthesis from an RNA world. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).
 85. des Georges, A. *et al.* Structure of the mammalian ribosomal pre-termination complex associated with eRF1 \cdot eRF3 \cdot GDPNP. *Nucleic Acids Res* **42**, 3409–3418 (2014).
 86. Sin, C., Chiarugi, D. & Valleriani, A. Quantitative assessment of ribosome drop-off in E. coli. *Nucleic Acids Res.* **44**, 2528–2537 (2016).
 87. Valleriani, a., Ignatova, Z., Nagar, a. & Lipowsky, R. Turnover of messenger RNA: Polysome statistics beyond the steady state. *EPL (Europhysics Lett.* **89**, 58003 (2010).
 88. Jungreis, I. *et al.* Evolutionary dynamics of abundant stop codon readthrough. *Mol. Biol. Evol.* **33**, 3108–3132 (2016).
 89. Dabrowski, M., Bukowy-Bieryllo, Z. & Zietkiewicz, E. Translational readthrough potential

- of natural termination codons in eucaryotes--The impact of RNA sequence. *RNA Biol.* **12**, 950–958 (2015).
90. Ortiz, J. O. *et al.* Structure of hibernating ribosomes studied by cryoelectron tomography in vitro and in situ. *J. Cell Biol.* **190**, 613–621 (2010).
 91. Afonina, Z. A., Myasnikov, A. G., Shirokov, V. A., Klaholz, B. P. & Spirin, A. S. Formation of circular polyribosomes on eukaryotic mRNA without cap-structure and poly(A)-tail: A cryo electron tomography study. *Nucleic Acids Res.* **42**, 9461–9469 (2014).
 92. Myasnikov, A. G. *et al.* The molecular structure of the left-handed supra-molecular helix of eukaryotic polyribosomes. *Nat. Commun.* **5**, 5294 (2014).
 93. Brandt, F. *et al.* The native 3D organization of bacterial polysomes. *Cell* **136**, 261–271 (2009).
 94. Brandt, F., Carlson, L. A., Hartl, F. U., Baumeister, W. & Grünewald, K. The Three-Dimensional Organization of Polyribosomes in Intact Human Cells. *Mol. Cell* **39**, 560–569 (2010).
 95. Viero, G. *et al.* Three distinct ribosome assemblies modulated by translation are the building blocks of polysomes. *J. Cell Biol.* **208**, 581–596 (2015).
 96. Lunelli, L. *et al.* Peering at brain polysomes with Atomic Force Microscopy. *J Vis Exp.* **Accepted**, 1–8 (2015).
 97. Wolin, S. L. & Walter, P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J.* **7**, 3559–3569 (1988).
 98. Li, G.-W., Oh, E. & Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**, 538–541 (2012).
 99. Herr, A. J., Wills, N. M., Nelson, C. C., Gesteland, R. F. & Atkins, J. F. Drop-off during ribosome hopping. *J. Mol. Biol.* **311**, 445–452 (2001).
 100. Cruz-Vera, L. R., Magos-Castro, M. A., Zamora-Romo, E. & Guarneros, G. Ribosome stalling and peptidyl-tRNA drop-off during translational delay at AGA codons. *Nucleic Acids Res.* **32**, 4462–4468 (2004).
 101. Sharp, P. M. & Li, W. H. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res.* **14**, 7737–7749 (1986).
 102. Ferguson, J., Ho, J. Y., Peterson, T. A. & Reed, S. I. Nucleotide sequence of the yeast cell division cycle start genes CDC28, CDC36, CDC37, and CDC39, and a structural analysis of the predicted products. *Nucleic Acids Res.* **14**, 6681–6697 (1986).
 103. Behura, S. K. & Severson, D. W. Codon usage bias: Causative factors, quantification methods and genome-wide patterns: With emphasis on insect genomes. *Biol. Rev.* **88**, 49–61 (2013).
 104. Stadler, M. & Fire, A. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA* **17**, 2063–2073 (2011).
 105. Novoa, E. M. & Ribas de Pouplana, L. Speeding with control: Codon usage, tRNAs, and ribosomes. *Trends in Genetics* **28**, 574–581 (2012).
 106. Ermolaeva, M. D. Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* **3**, 91–97

- (2001).
107. Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160**, 1111–1124 (2015).
 108. Drummond, D. A. & Wilke, C. O. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* **134**, 341–352 (2008).
 109. Dong, H., Nilsson, L. & Kurland, C. G. Co-variation of tRNA Abundance and Codon Usage in *Escherichia coli* at Different Growth Rates. *J. Mol. Biol.* **260**, 649–663 (1996).
 110. Moriyama, E. N. & Powell, J. R. Codon usage bias and tRNA abundance in *Drosophila*. *J. Mol. Evol.* **45**, 514–523 (1997).
 111. Berg, O. G. & Kurland, C. . Growth rate-optimised tRNA abundance and codon usage. *J. Mol. Biol.* **270**, 544–550 (1997).
 112. Dittmar, K. A., Goodenbour, J. M. & Pan, T. Tissue-specific differences in human transfer RNA expression. *PLoS Genet.* **2**, 2107–2115 (2006).
 113. Park, J.-H. *et al.* Preferential use of minor codons in the translation initiation region of human genes. *Hum. Genet.* **136**, 67–74 (2017).
 114. Komar, A. A., Lesnik, T. & Reiss, C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.* **462**, 387–391 (1999).
 115. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).
 116. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–8 (2009).
 117. Nakahigashi, K. *et al.* Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. *BMC Genomics* **15**, 1115 (2014).
 118. dos Reis, M., Savva, R. & Wernisch, L. Solving the riddle of codon usage preferences: A test for translational selection. *Nucleic Acids Res.* **32**, 5036–5044 (2004).
 119. Sharp, P. M. & Li, W. The codon Adaptation Index - A measure of directional synonymous codon usage bias, and its possible applications. *Nucleic Acids Res.* **15**, 1281–1295 (1987).
 120. Tuller, T., Kupiec, M. & Ruppin, E. Determinants of protein abundance and translation efficiency in *S. cerevisiae*. *PLoS Comput Biol* **3**, e248 (2007).
 121. Zupanic, A. *et al.* Detecting translational regulation by change point analysis of ribosome profiling data sets. *RNA* **2014**, 1507–1518 (2014).
 122. Zhang, D., Chen, D., Cao, L., Li, G. & Cheng, H. The effect of codon mismatch on the protein translation system. *PLoS One* **11**, (2016).
 123. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation Using Ribosome Profiling. *Science (80-)*. **324**, 218–23 (2009).
 124. Tuller, T. & Zur, H. Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res.* **43**, 13–28 (2015).

125. Chu, D. *et al.* Translation elongation can control translation initiation on eukaryotic mRNAs. *EMBO J.* **33**, 21–34 (2014).
126. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).
127. Bentele, K., Saffert, P., Rauscher, R., Ignatova, Z. & Blüthgen, N. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* **9**, 675 (2013).
128. Tuller, T. *et al.* Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.* **12**, R110 (2011).
129. Shah, P., Ding, Y., Niemczyk, M., Kudla, G. & Plotkin, J. B. Rate-limiting steps in yeast protein translation. *Cell* **153**, 1589–601 (2013).
130. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–32 (2012).
131. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270 (2014).
132. McShane, E. *et al.* Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell* **167**, 803–815.e21 (2016).
133. Kapeli, K. & Yeo, G. W. Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: Implications for neurological diseases. *Frontiers in Neuroscience* (2012). doi:10.3389/fnins.2012.00144
134. King, H. A. & Gerber, A. P. Translatome profiling: Methods for genome-scale analysis of mRNA translation. *Briefings in Functional Genomics* **15**, 22–31 (2016).
135. Mašek, T., Valášek, L. & Pospíšek, M. *Polysome analysis and RNA purification from sucrose gradients. Methods in Molecular Biology (Clifton, N.J.)* **703**, (2011).
136. Spangenberg, L. *et al.* Polysome profiling shows extensive posttranscriptional regulation during human adipocyte stem cell differentiation into adipocytes. *Stem Cell Res.* **11**, 902–912 (2013).
137. Arava, Y. *et al.* Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci* **100**, 3889–3894 (2003).
138. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science (80-.)*. **324**, 218–223 (2009).
139. Heiman, M. *et al.* A Translational Profiling Approach for the Molecular Characterization of CNS Cell Types. *Cell* **135**, 738–748 (2008).
140. Martens, A. T., Taylor, J. & Hilser, V. J. Ribosome A and P sites revealed by length analysis of ribosome profiling data. *Nucleic Acids Res.* **43**, 3680–3687 (2015).
141. Legendre, R., Baudin-Baillieu, A., Hatin, I. & Namy, O. RiboTools: A Galaxy toolbox for qualitative ribosome profiling analysis. *Bioinformatics* **31**, 2586–2588 (2015).
142. Andreev, D. E. *et al.* Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res.* gkw1190 (2016).

143. Qin, X., Ahn, S., Speed, T. P. & Rubin, G. M. Global analyses of mRNA translational control during early Drosophila embryogenesis. *Genome Biol* **8**, R63 (2007).
144. Picard, F. *et al.* Bacterial translational regulations: high diversity between all mRNAs and major role in gene expression. *BMC Genomics* **13**, 528 (2012).
145. Janas, M. M. *et al.* Reduced Expression of Ribosomal Proteins Relieves MicroRNA-Mediated Repression. *Mol. Cell* **46**, 171–186 (2012).
146. Molotski, N. & Soen, Y. Differential association of microRNAs with polysomes reflects distinct strengths of interactions with their mRNA targets. *RNA* **18**, 1612–23 (2012).
147. Park, J. E., Yi, H., Kim, Y., Chang, H. & Kim, V. N. Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. *Mol. Cell* **62**, 462–471 (2016).
148. Nussbacher, J. K., Batra, R., Lagier-Tourenne, C. & Yeo, G. W. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends in Neurosciences* **38**, 226–236 (2015).
149. Reschke, M. *et al.* Characterization and analysis of the composition and dynamics of the mammalian riboproteome. *Cell Rep.* **4**, 1276–1287 (2013).
150. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012).
151. Steitz, J. A. Polypeptide Chain Initiation: Nucleotide Sequences of the Three Ribosomal Binding Sites in Bacteriophage R17 RNA. *Nature* **224**, 957–964 (1969).
152. David, A. *et al.* Nuclear translation visualized by ribosome-bound nascent chain puromycylation. *J. Cell Biol.* **197**, 45–57 (2012).
153. RICH, A., PENMAN, S., BECKER, Y., DARNELL, J. & HALL, C. POLYRIBOSOMES: SIZE IN NORMAL AND POLIO- INFECTED HELA CELLS. *Science* **142**, 1658–1663 (1963).
154. Warner, J. R., Knopf, P. M. & Rich, A. A multiple ribosomal structure in protein synthesis. *Proc. Natl. Acad. Sci.* **49**, 122–129 (1963).
155. Penman, S., Scherrer, K., Becker, Y. & Darnell, J. E. Polyribosomes in normal and poliovirus-infected HeLa cells and their relationship to messenger-RNA. *Proc. Natl. Acad. Sci.* **49**, 654–662 (1963).
156. Munro, A. J., Jackson, R. J. & Korner, A. Studies on the nature of polysomes. *Biochem. J.* **92**, 289 (1964).
157. Yan, X., Hoek, T. A., Vale, R. D. & Tanenbaum, M. E. Dynamics of Translation of Single mRNA Molecules in Vivo. *Cell* **165**, 976–989 (2016).
158. Heyer, E. E. & Moore, M. J. Redefining the Translational Status of 80S Monosomes. *Cell* **164**, 757–769 (2016).
159. Aspden, J. L. *et al.* Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *Elife* **3**, e03528 (2014).
160. El Fatimy, R. *et al.* Tracking the Fragile X Mental Retardation Protein in a Highly Ordered Neuronal RiboNucleoParticles Population: A Link between Stalled Polyribosomes and RNA Granules. *PLoS Genet.* **12**, (2016).

161. Graber, T. E. *et al.* Reactivation of stalled polyribosomes in synaptic plasticity. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16205–10 (2013).
162. Clamer, M. *et al.* Active ribosome profiling with RiboLace. *bioRxiv*
163. Olshen, A. B. *et al.* Assessing gene-level translational control from ribosome profiling. *Bioinformatics* **29**, 2995–3002 (2013).
164. Arava, Y., Boas, F. E., Brown, P. O. & Herschlag, D. Dissecting eukaryotic translation and its control by ribosome density mapping. *Nucleic Acids Res* **33**, 2421–2432 (2005).
165. Hsu, P. Y. *et al.* Super-Resolution Ribosome Profiling Reveals Novel Translation Events in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **113**, In Revision (2016).
166. Kochetov, A. V *et al.* AltORFev facilitates the prediction of alternative open reading frames in eukaryotic mRNAs. *Bioinformatics* btw736 (2016).
167. Raj, A. *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5**, (2016).
168. Kapeli, K. & Yeo, G. Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases. *Front. Neurosci.* **6**, 144 (2012).
169. Michel, A. M., Andreev, D. E. & Baranov, P. V. Computational approach for calculating the probability of eukaryotic translation initiation from ribo-seq data that takes into account leaky scanning. *BMC Bioinformatics* **15**, 380 (2014).
170. Weinberg, D. E. *et al.* Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep.* **14**, 1787–1799 (2016).
171. Lareau, L. F., Hite, D. H., Hogan, G. J. & Brown, P. O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife* **2014**, (2014).
172. Xiao, Z., Zou, Q., Liu, Y. & Yang, X. Genome-wide assessment of differential translations with ribosome profiling data. *Nat. Commun.* **7**, 11194 (2016).
173. Zhong, Y. *et al.* RiboDiff: Detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* **33**, 139–141 (2017).
174. O'Connor, P. B. F., Andreev, D. E. & Baranov, P. V. Comparative survey of the relative impact of mRNA features on local ribosome profiling read density. *Nat. Commun.* **7**, 12915 (2016).
175. Crappé, J. *et al.* PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* **43**, e29--e29 (2015).
176. Chun, S. Y., Rodriguez, C. M., Todd, P. K. & Mills, R. E. SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data. *bioRxiv* 1–6 (2015). doi:http://dx.doi.org/10.1101/034777
177. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 1–9 (2015).
178. de Klerk, E. *et al.* Assessing the translational landscape of myogenic differentiation by

- ribosome profiling. *Nucleic Acids Res.* gkv281 (2015).
179. Malone, B. *et al.* Bayesian prediction of RNA translation from ribosome profiling. *Nucleic Acids Res.* **45**, 2960–2972 (2017).
 180. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010).
 181. Gritsenko, A. A., Hulsman, M., Reinders, M. J. T. & de Ridder, D. Unbiased Quantitative Models of Protein Translation Derived from Ribosome Profiling Data. *PLoS Comput. Biol.* **11**, (2015).
 182. Dana, A. & Tuller, T. Mean of the typical decoding rates: a new translation efficiency index based on the analysis of ribosome profiling data. *G3 Genes/ Genomes/ Genet.* **5**, 73–80 (2015).
 183. Dana, A. & Tuller, T. Properties and determinants of codon decoding time distributions. *BMC Genomics* **15 Suppl 6**, S13 (2014).
 184. Pop, C. *et al.* Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.* **10**, 770–770 (2014).
 185. Dunn, J. G. & Weissman, J. S. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. *BMC Genomics* **17**, 958 (2016).
 186. Popa, A. *et al.* RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Research* **5**, 1309 (2016).
 187. H Backman, T. W. & Girke, T. systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics* **17**, 388 (2016).
 188. Michel, A. M. *et al.* GWIPS-viz: Development of a ribo-seq genome browser. *Nucleic Acids Res.* **42**, (2014).
 189. Xie, S. Q. *et al.* RPFdb: A database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res.* **44**, D254–D258 (2016).
 190. Ji, Z., Song, R., Huang, H., Regev, A. & Struhl, K. Transcriptome-scale RNase-footprinting of RNA-protein complexes. *Nat. Biotechnol.* **34**, 2–6 (2016).
 191. Bartholomewus, A., Del Campo, C. & Ignatova, Z. Mapping the non-standardized biases of ribosome profiling. *Biological Chemistry* **397**, 23–35 (2016).
 192. Lecanda, A. *et al.* Dual randomization of oligonucleotides to reduce the bias in ribosome-profiling libraries. *Methods* **107**, 89–97 (2016).
 193. Bohnert, R. & Ratsch, G. rQuant. web: a tool for RNA-Seq-based transcript quantitation. *Nucleic Acids Res* **38**, W348–W351 (2010).
 194. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**, e131 (2010).
 195. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genom Biol* **12**, r22 (2010).
 196. Diamant, A. & Tuller, T. Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biol. Direct* **11**, 24 (2016).

197. Burton, D. M. *The History of Mathematics: An Introduction. Book 36*, (2010).
198. Bailey, N. The mathematical theory of infectious diseases and its applications. 2nd edition. *Math. theory Infect. Dis. its Appl. 2nd Ed.* **413**, (1975).
199. Abrams, P. A. The Evolution of Predator-Prey Interactions: Theory and Evidence. *Annu. Rev. Ecol. Syst.* **31**, 79–105 (2000).
200. Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nat. Rev. Microbiol.* (2008). doi:10.1038/nrmicro1845
201. Karle, J. Direct methods in protein crystallography. *Acta Crystallographica Section A: Foundations of Crystallography* **45**, (1989).
202. Glynn, P., Unudurthi, S. D. & Hund, T. J. Mathematical modeling of physiological systems: An essential tool for discovery. *Life Sciences* **111**, 1–5 (2014).
203. Tuszynski, J. a *et al.* Mathematical and computational modeling in biology at multiple scales. *Theor. Biol. Med. Model.* **11**, 52 (2014).
204. Winterbach, W., Mieghem, P. Van, Reinders, M., Wang, H. & Ridder, D. de. Topology of molecular interaction networks. *BMC Syst. Biol.* **7**, 1–15 (2013).
205. Gerst, I. & Levine, S. N. Kinetics of protein synthesis by polyribosomes. *J Theor Biol* **9**, 16–36 (1965).
206. Zur, H. & Tuller, T. Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution. *Nucleic Acids Res.* **44**, 9031–9049 (2016).
207. von der Haar, T. Mathematical and Computational Modelling of Ribosomal Movement and Protein Synthesis: an overview. *Comput. Struct. Biotechnol. J.* **1**, 1–7 (2012).
208. Lodish, H. F. Model for the regulation of mRNA translation applied to haemoglobin synthesis. *Nature* **251**, 385–388 (1974).
209. Zouridis, H. & Hatzimanikatis, V. Effects of codon distributions and tRNA competition on protein translation. *Biophys J* **95**, 1018–1033 (2008).
210. Heinrich, R. & Rapoport, T. A. Mathematical modelling of translation of mRNA in eucaryotes; steady states, time-dependent processes and application to reticulocyttest. *J. Theor. Biol.* **86**, 279–313 (1980).
211. Zhang, G. & Ignatova, Z. Generic algorithm to predict the speed of translational elongation: Implications for protein biogenesis. *PLoS One* **4**, (2009).
212. Zouridis, H. & Hatzimanikatis, V. A model for protein translation: polysome self-organization leads to maximum protein synthesis rates. *Biophys J* **92**, 717–730 (2007).
213. Zhang, S., Goldman, E. & Zubay, G. Clustering of low usage codons and ribosome movement. *Journal of theoretical biology* **170**, 339–54 (1994).
214. Mitarai, N., Sneppen, K. & Pedersen, S. Ribosome Collisions and Translation Efficiency: Optimization by Codon Usage and mRNA Destabilization. *J. Mol. Biol.* **382**, 236–245 (2008).
215. Dong, J. J., Schmittmann, B. & Zia, R. K. P. Towards a Model for Protein Production Rates. *J. Stat. Phys.* **128**, 21–34 (2006).

216. Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).
217. Chu, D., Barnes, D. J. & Von Der Haar, T. The role of tRNA and ribosome competition in coupling the expression of different mRNAs in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **39**, 6705–6714 (2011).
218. Chou, T. & Lakatos, G. Clustered bottlenecks in mRNA translation and protein synthesis. *Phys. Rev. Lett.* **93**, (2004).
219. Ciandrini, L., Stansfield, I. & Romano, M. C. Ribosome Traffic on mRNAs Maps to Gene Ontology: Genome-wide Quantification of Translation Initiation Rates and Polysome Size Regulation. *PLoS Comput. Biol.* **9**, (2013).
220. Raveh, A., Margaliot, M., Sontag, E. D. & Tuller, T. A model for competition for ribosomes in the cell. *J. R. Soc. Interface* **13**, 20151062 (2016).
221. Helbing, D. Traffic and related self-driven many-particle systems. *Rev. Mod. Phys.* **73**, 1067–1141 (2001).
222. Gillespie, D. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
223. Neyman, J. Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability. *Philos. Trans. R. Soc. London. Ser. A, Math. Phys. Sci.* **236**, 333 LP-380 (1937).
224. Gorgoni, B., Ciandrini, L., McFarland, M. R., Romano, M. C. & Stansfield, I. Identification of the mRNA targets of tRNA-specific regulation using genome-wide simulation of translation. *Nucleic Acids Res.* **44**, 9231–9244 (2016).
225. Rogers, D. W., Boettcher, M. A., Traulsen, A. & Greig, D. Ribosome reinitiation can explain length-dependent translation of messenger RNA. *PLOS Comput. Biol.* **13**, e1005592 (2017).
226. Mao, Y., Liu, H., Liu, Y. & Tao, S. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **42**, 4813–4822 (2014).
227. Cook, L. J., Zia, R. K. P. & Schmittmann, B. Competition between multiple totally asymmetric simple exclusion processes for a finite pool of resources. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **80**, (2009).
228. Duc, K. D., Saleem, Z. H. & Song, Y. S. Theoretical quantification of interference in the TASEP: Application to mRNA translation shows near-optimality of termination rates. *bioRxiv* 147017 (2017).
229. Gilchrist, M. A. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol. Biol. Evol.* **24**, 2362–2372 (2007).
230. Gilchrist, M. A. & Wagner, A. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. *J. Theor. Biol.* **239**, 417–434 (2006).
231. Kubatko, L., Shah, P., Herbei, R. & Gilchrist, M. A. A codon model of nucleotide substitution with selection on synonymous codon usage. *Mol. Phylogenet. Evol.* **94**,

- 290–297 (2016).
232. Innocentini, G. C. P., Forger, M., Radulescu, O. & Antoneli, F. Protein Synthesis Driven by Dynamical Stochastic Transcription. *Bull. Math. Biol.* **78**, 110–131 (2016).
 233. Zhang, G. *et al.* Global and local depletion of ternary complex limits translational elongation. *Nucleic Acids Res.* **38**, 4778–4787 (2010).
 234. Zhao, Y. B. & Krishnan, J. Probabilistic Boolean Network Modelling and Analysis Framework for mRNA Translation. *IEEE/ACM Trans Comput Biol Bioinform* **13**, 754–766 (2015).
 235. Shmulevich, I., Dougherty, E. R., Kim, S. & Zhang, W. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **18**, 261–274 (2002).
 236. Kolb, S. J. & Kissel, J. T. Spinal Muscular Atrophy. *Neurologic Clinics* **33**, 831–846 (2015).
 237. Werdnig, G. From the Pathological-Anatomical Institute of Graz TWO EARLY INFANTILE HEREDITARY CASES OF PROGRESSIVE MUSCULAR ATROPHY SIMULATING DYSTROPHY, BUT ON A NEURAL BASIS. *Arch. Neurol.* **25**, 276–278 (1971).
 238. Tiryaki, E. & Horak, H. A. ALS and other motor neuron diseases. *Contin. Lifelong Learn. Neurol.* **20**, 1185–1207 (2014).
 239. d??Ydewalle, C. & Sumner, C. J. Spinal Muscular Atrophy Therapeutics: Where do we Stand? *Neurotherapeutics* **12**, 303–316 (2015).
 240. Lunn, M. R. & Wang, C. H. Spinal muscular atrophy. *Lancet* **371**, 2120–2133 (2008).
 241. Porro, F. *et al.* The wide spectrum of clinical phenotypes of spinal muscular atrophy with respiratory distress type 1: A systematic review. *Journal of the Neurological Sciences* **346**, 35–42 (2014).
 242. Finkel, R. S. *et al.* Observational study of spinal muscular atrophy type I and implications for clinical trials. *Neurology* **83**, 810–817 (2014).
 243. Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy-determining gene [see comments]. *Cell* **80**, 155–165 (1995).
 244. Gabanella, F. *et al.* SMN affects membrane remodelling and anchoring of the protein synthesis machinery. *J. Cell Sci.* **129**, 804–16 (2016).
 245. Béchade, C. Subcellular distribution of survival motor neuron (SMN) protein: Possible involvement in nucleocytoplasmic and dendritic transport. *Eur. J. Neurosci.* **11**, 293–304 (1999).
 246. Sanchez, G. *et al.* A novel function for the survival motoneuron protein as a translational regulator. *Hum. Mol. Genet.* **22**, 668–684 (2013).
 247. Kim, E. & Jung, H. Local protein synthesis in neuronal axons: Why and how we study. *BMB Reports* **48**, 139–146 (2015).
 248. Jung, H., Gkogkas, C. G., Sonenberg, N. & Holt, C. E. Remote control of gene function by local translation. *Cell* **157**, 26–40 (2014).
 249. Bernabò, P. *et al.* In vivo translome profiling reveals early defects in ribosome biology

- underlying SMA pathogenesis. *bioRxiv* 103481 (2017).
250. MacDonald, C. T. & Gibbs, J. H. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Biopolymers* **7**, 707–725 (1969).
 251. Reuveni, S., Meilijson, I., Kupiec, M., Ruppin, E. & Tuller, T. Genome-scale analysis of translation elongation with a ribosome flow model. *PLoS Comput Biol* **7**, e1002127 (2011).
 252. Gilchrist, M. A. Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol Biol Evol* **24**, 2362–2372 (2007).
 253. Park, J. H. *et al.* Preferential use of minor codons in the translation initiation region of human genes. *Hum. Genet.* **136**, 67–74 (2017).
 254. Wilson, D. N., Arenz, S. & Beckmann, R. Translation regulation via nascent polypeptide-mediated ribosome stalling. *Current Opinion in Structural Biology* **37**, 123–133 (2016).
 255. Ishimura, R. *et al.* RNA function. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science* **345**, 455–9 (2014).
 256. Rooijers, K., Loayza-Puch, F., Nijtmans, L. G. & Agami, R. Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. *Nat. Commun.* **4**, 2886 (2013).
 257. Raveh, A., Margalio, M., Sontag, E. D. & Tuller, T. A Model for Competition for Ribosomes in the Cell. *J. R. Soc. Interface* **13**, 1508.02408 (2015).
 258. Warner, J. R. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* **24**, 437–440 (1999).
 259. Lauria, F. *et al.* RiboAbacus: A model trained on polyribosome images predicts ribosome density and translational efficiency from mammalian transcriptomes. *Nucleic Acids Res.* **43**, (2015).
 260. Ooi, H. K. & Ma, L. Modeling cell-to-cell stochastic variability in intrinsic apoptosis pathway. in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 5498–5501 (2012). doi:10.1109/EMBC.2012.6347239
 261. Gillespie, D. T. Stochastic Simulation of Chemical Kinetics. *Annu. Rev. Phys. Chem.* **58**, 35–55 (2007).
 262. Nedialkova, D. D. & Leidel, S. A. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell* **161**, 1606–1618 (2015).
 263. Gao, X. *et al.* Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods* **12**, 147–153 (2014).
 264. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).
 265. Holmqvist, E. *et al.* Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. *EMBO J.* **35**, e201593360 (2016).
 266. Uren, P. J. *et al.* Site identification in high-throughput RNA-protein interaction data.

- Bioinformatics* **28**, 3013–3020 (2012).
267. Corcoran, D. L. *et al.* PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.* **12**, R79 (2011).
 268. Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct. Mol. Biol.* **20**, 1434–42 (2013).
 269. Akaike, H. Information theory and an extension of the maximum likelihood principle. *Int. Symp. Inf. theory* 267–281 (1973). doi:10.1007/978-1-4612-1694-0
 270. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **6**, 461–464 (1978).
 271. Fisher, R. *Statistical methods for research workers. Biological monographs and manuals* (1925). doi:10.1056/NEJMc061160
 272. Fuller, H. R. & Morris, G. E. SMN complexes of nucleus and cytoplasm: A proteomic study for SMA therapy. *Transl. Neurosci.* **1**, 261–267 (2010).
 273. Fallini, C., Donlin-Asp, P. G., Rouanet, J. P., Bassell, G. J. & Rossoll, W. Deficiency of the Survival of Motor Neuron Protein Impairs mRNA Localization and Local Translation in the Growth Cone of Motor Neurons. *J. Neurosci.* **36**, 3811–20 (2016).
 274. Hsieh-Li, H. M. *et al.* A mouse model for spinal muscular atrophy. *Nat. Genet.* **24**, 66–70 (2000).
 275. Chen, J., Tsai, A., O’Leary, S. E., Petrov, A. & Puglisi, J. D. Unraveling the dynamics of ribosome translocation. *Current Opinion in Structural Biology* **22**, 804–814 (2012).
 276. Budkevich, T. V. *et al.* Regulation of the mammalian elongation cycle by subunit rolling: A eukaryotic-specific ribosome rearrangement. *Cell* **158**, 121–131 (2014).
 277. Han, Y. *et al.* Ribosome profiling reveals sequence-independent post-initiation pausing as a signature of translation. *Cell Res.* **24**, 842–851 (2014).
 278. Sako, H., Yada, K. & Suzuki, K. Genome-Wide Analysis of Acute Endurance Exercise-Induced Translational Regulation in Mouse Skeletal Muscle. *PLoS One* **11**, (2016).
 279. Sugimoto, Y. *et al.* {hiCLIP} reveals the in vivo atlas of {mRNA} secondary structures recognized by Staufen 1. *Nature* **519**, 491–494 (2015).
 280. Schueren, F. & Thoms, S. Functional Translational Readthrough: A Systems Biology Perspective. *PLoS Genetics* **12**, (2016).
 281. Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479–90 (2007).
 282. Floquet, C., Hatin, I., Rousset, J. P. & Bidou, L. Statistical analysis of readthrough levels for nonsense mutations in mammalian cells reveals a major determinant of response to gentamicin. *PLoS Genet.* **8**, (2012).
 283. Manuvakhova, M. *et al.* Aminoglycoside antibiotics mediate context-dependent suppression of termination codons in a mammalian translation system. *RNA* **6**, 1044–1055 (2000).
 284. McCaughan, K. K., Brown, C. M., Dalphin, M. E., Berry, M. J. & Tate, W. P. Translational termination efficiency in mammals is influenced by the base following the stop codon.

- Proc. Natl. Acad. Sci. U. S. A.* **92**, 5431–5435 (1995).
285. Sulima, S. O. *et al.* Eukaryotic rpl10 drives ribosomal rotation. *Nucleic Acids Res.* **42**, 2049–2063 (2014).
 286. Shoemaker, C. J. & Green, R. Kinetic analysis reveals the ordered coupling of translation termination and ribosome recycling in yeast. *Proc Natl Acad Sci* **108**, E1392-8 (2011).
 287. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128 (2013).
 288. Workman, E., Kalda, C., Patel, A. & Battle, D. J. Gemin5 binds to the survival motor neuron mRNA to regulate SMN expression. *J. Biol. Chem.* **290**, 15662–15669 (2015).
 289. Tuller, T., Waldman, Y. Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3645–50 (2010).
 290. Zhou, Z. *et al.* Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *Proc. Natl. Acad. Sci.* **113**, E6117--E6125 (2016).
 291. Gerashchenko, M. V. & Gladyshev, V. N. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.* **42**, (2014).
 292. Andreev, D. E. *et al.* Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Research* **45**, 513–526 (2017).
 293. Ramírez, V. *et al.* Loss of a Conserved tRNA Anticodon Modification Perturbs Plant Immunity. *PLoS Genet.* **11**, (2015).
 294. Harrison, B. R., Yazgan, O. & Krebs, J. E. Life without RNAi: noncoding RNAs and their functions in *Saccharomyces cerevisiae*. *Biochem. Cell Biol.* **87**, 767–779 (2009).
 295. Doma, M. K. & Parker, R. Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation. *Nature* **440**, 561–4 (2006).
 296. Buchan, J. R. & Stansfield, I. Halting a cellular production line: responses to ribosomal pausing during translation. *Biol. Cell* **99**, 475–487 (2007).
 297. Bonnin, P., Kern, N., Young, N. T., Stansfield, I. & Romano, M. C. Novel mRNA-specific effects of ribosome drop-off on translation rate and polysome profile. *PLoS Comput. Biol.* **13**, (2017).
 298. Janosi, L. *et al.* Evidence for in vivo ribosome recycling, the fourth step in protein biosynthesis. *EMBO J.* **17**, 1141–1151 (1998).
 299. Young, D. J., Guydosh, N. R., Zhang, F., Hinnebusch, A. G. & Green, R. Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3'UTRs In Vivo. *Cell* **162**, 872–884 (2015).
 300. Guydosh, N. R. & Green, R. Dom34 rescues ribosomes in 3' untranslated regions. *Cell* **156**, 950–962 (2014).
 301. Bernabo, P. *et al.* In Vivo Translatome Profiling in Spinal Muscular Atrophy Reveals a Role for SMN Protein in Ribosome Biology. *Cell Rep.* **21**, 953 (2017).

302. Valášek, L. S. *et al.* Embraced by eIF3: structural and functional insights into the roles of eIF3 across the translation cycle. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx805
303. Fallini, C. *et al.* Dynamics of survival of motor neuron (SMN) protein interaction with the mRNA-binding protein IMP1 facilitates its trafficking into motor neuron axons. *Dev. Neurobiol.* **74**, 319–332 (2014).
304. Fallini, C. *et al.* The survival of motor neuron (SMN) protein interacts with the mRNA-binding protein HuD and regulates localization of poly(A) mRNA in primary motor neuron axons. *J. Neurosci.* **31**, 3914–25 (2011).
305. Fallini, C., Bassell, G. J. & Rossoll, W. Spinal muscular atrophy: The role of SMN in axonal mRNA regulation. *Brain Research* **1462**, 81–92 (2012).
306. Rossoll, W. *et al.* Smn, the spinal muscular atrophy-determining gene product, modulates axon growth and localization of ??-actin mRNA in growth cones of motoneurons. *J. Cell Biol.* **163**, 801–812 (2003).
307. Donlin-Asp, P. G., Rossoll, W. & Bassell, G. J. Spatially and temporally regulating translation via mRNA binding proteins in cellular and neuronal function. *FEBS Lett.* (2017).
308. Liu-Yesucevitz, L. *et al.* Local RNA translation at the synapse and in disease. *J. Neurosci.* **31**, 16086–93 (2011).
309. Wang, E. T. *et al.* Dysregulation of mRNA Localization and Translation in Genetic Disease. *J. Neurosci.* **36**, 11418–11426 (2016).
310. Robinson, MD, McCarthy, DJ, Smyth, G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
311. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–7 (2012).
312. Kuleshov, M. V *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-7 (2016).

Acknowledgements

I am truly indebted and thankful to Dr. Gabriella Viero for mentoring me throughout my PhD experience, for her unlimited patience and for the enthusiasm she was able to communicate and transmit to me.

Massive thanks go to Dr. Toma Tebaldi for guiding me through the mysteries of computational biology, for always advising me and for his kind and never-ending support.

I am very grateful to Dr. Guido Sanguinetti for supervising the mathematical side of the project and for giving me the opportunity to join his group in Edinburgh for three months.

A sincere thanks to my co-workers Paola and Marta for the good time spent together, for answering my countless questions – not always concerning biology – and for their constant company.

Thanks to all the present and former colleagues and students that made these four years highly enjoyable. Special thanks to Max, Primoz, Matteo, Elena and Francesca.

I would like to thank all the scientists who contributed to this project providing and sharing materials, notions and thoughts.

Thanks to the AXonomIX research project funded by the Provincia Autonoma di Trento for the financial support for the project. I am also grateful to the European Molecular Biology Organization for awarding me a short-term fellowship funding my period abroad.

Finally, thanks to my family and friends for their support over the years.