Insights into translational regulation from ribosome profiling data

Katarzyna Anna Chyżyńska

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2019

UNIVERSITY OF BERGEN

Insights into translational regulation from ribosome profiling data

Katarzyna Anna Chyżyńska

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 15.10.2019

Year: 2019

Title: Insights into translational regulation from ribosome profiling data

Name: Katarzyna Anna Chyżyńska

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

This thesis was prepared at the Computational Biology Unit, Department of Informatics, University of Bergen in fulfilment of the requirements for acquiring a Ph.D. degree.

All the work was funded by UiB and carried out under the supervision of Sushma Grellscheid, Ph.D. and co-supervision of Eivind Valen, Ph.D.

Acknowledgements

First of all, I would like to thank the Internet, particularly the *Google* search engine, *Stack Overflow* and *Seinfeld* reruns, which provided me with much needed knowledge, inspiration and distractions essential to completing this thesis.

Secondly, I would like to thank my supervisor, Sushma Grellscheid, for taking me on as her student and doing her very best in the limited time we had.

Thirdly, I would like to thank the administrative staff at the Department of Informatics for making it a fair and supportive workplace.

Last but not least, I would like to thank my colleagues and friends for both academic and non-academic support. Particularly, TT for encouragement through thick and thin, KL for giving ear to many crazy ideas and endless frustrations, PT for library writing sessions. AH, for the good vibes and cancelled balloon rides. MG, for PLK and so much more.

Abstract

Ribosomes carry out protein synthesis from mRNA templates by a highly regulated process called translation. Within the four phases of translation - initiation, elongation, termination and recycling - the focus of translation regulation studies has traditionally fallen on the initiation as the rate-limiting step in protein production. Recent evidence, however, points to the profound importance of regulatory control of elongation during development, neurologic disease, cell stress and even cancer.

Ribosome profiling provides an unprecedented means of studying translational regulation on a global level. It is based on deep sequencing of ribosome-protected mRNA fragments, capturing snapshots of genome-wide translation. However, as with any new experimental technique, biases inherent in the ribosome profiling method are gradually being explored and understood, and serve to inform further refinement of the technique.

In the first part of this thesis I provide a comprehensive overview of the current state of knowledge on translation and its regulation, particularly at the elongation phase. I describe the ribosome profiling technique, data processing and applications to studying translational regulation. Afterwards, I go on to present the results in the form of two scientific papers. First paper tackles the challenge of ribosome profiling data processing, setting the ground work for second paper. The second paper uses improved processing to explore ribosome stalling and its potential regulatory functions.

The first paper presents *Shoelaces*, a tool for processing and visualization of ribosome profiling data. Here, I demonstrate how streamlining and standardizing processing steps can contribute to better quality and comparability of data for downstream analyses. At the core of this are (1) filtering genuine translating footprints from noise based on periodicity and (2) determining a specific codon being translated by the ribosome thanks to length-dependent offset calculation. *Shoelaces* automatically selects footprint lengths and offsets, offering a user-friendly graphical interface as well as command line interface for batch processing. By reanalyzing 79 human libraries, I show that *Shoelaces* retains more quality data than the original manual analyses.

In the second paper, I investigate regulation of translation elongation by ribosome stalling. Utilizing the robust processing technique developed in the first paper, I apply it to process 20 ribosome datasets form yeast, fruit fly, zebrafish, mouse and human. Hypothesising that deep conservation of translation machinery would exist also for biologically significant stall sites, I detect 3293 of these conserved in at least two organisms. I find that proline and negatively charged amino acids are the main contributors to stalling. Furthermore, many of the stall sites are found in RNA processing genes, suggesting that stalling might play a conserved regulatory role in RNA metabolism. The project provides a rich resource for further in-depth studies on conserved stalling and suggests its possible roles in regulation of translation elongation.

Finally, the last part of this thesis consists of conclusive remarks an critical reflection on the impact these projects brought into the field. Here, I point out possible directions for future investigations. Additionally, I include a related paper, on the use of ribosome profiling data of initiating ribosomes in re-annotation of bacterial genomes.

Overall, this thesis demonstrates how mining ribosome profiling data can result in biologically meaningful discoveries pertaining to regulation of translation.

Abbreviations

mRNA	messenger RNA
poly(A)	poly-adenosine
rRNA	ribosomal RNA
SSU	small ribosomal subunit
LSU	large ribosomal subunit
tRNA	transfer RNA
aa-tRNA	aminoacyl-tRNA
A-site	aminoacyl entry site
P-site	peptidyl transferase site
E-site	exit site
ORF	open reading frame
UTR	untranslated region
RBS	ribosome binding site
SD	Shine-Dalgarno
IF	initiation factor
eIF	eukaryotic initiation factor
PIC	preinitiation complex
NMD	nonsense-mediated decay
CDS	coding sequence
TMD	transmembrane domain
SRP	signal recognition particle
RQC	ribosome-associated protein quality control
NGD	no-go decay
NSD	nonstop decay
PTC	premature termination codon
EJC	exon-junction complex
ncRNA	non-coding RNA
TIS	translation initiation site
lncRNA	long non-coding RNA
CHX	cycloheximide
uORF	upstream open reading frame
aSD	anti Shine-Dalgarno
TF	targeting factor
dORF	downstream ORF
miRNA	micro RNA

List of papers

Papers included in this thesis:

1. Åsmund Birkeland[†], **Katarzyna Chyżyńska**[†] and Eivind Valen, *Shoelaces: an interactive tool for ribosome profiling processing and visualization*, BMC Genomics 19:543, 2018.
2. **Katarzyna Chyżyńska**, Carl Jones, Kornel Labun, Eivind Valen[†] and Sushma Grellscheid[†], *Deep conservation of ribosome stall sites across RNA processing genes* (manuscript in preparation)

[†] Contributed equally

Related paper, included in the appendix:

3. Adam Giess, Veronique Jonckheere, Elvis Ndah, **Katarzyna Chyżyńska**, Petra Van Damme and Eivind Valen, *Ribosome signatures aid bacterial translation initiation site identification*, BMC Biology 15:76, 2017.

Contents

Chapter 1

Introduction

The fundamental story of all living things is that of interplay between genes, in the form of DNA, and proteins which they encode. Over 60 years ago, Francis Crick formulated the Central Dogma, depicting the basic flow of information between the two (Figure 1.1).

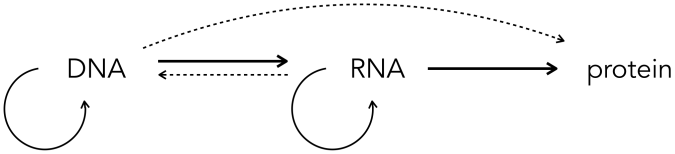


Figure 1.1: **The Central Dogma.**

For Crick, four types of information transfer were clear: DNA to DNA (DNA replication), DNA to messenger RNA (transcription), mRNA to protein (translation) and RNA to RNA (RNA viruses copying themselves), as well as potentially RNA to DNA (as it turned out later, some retroviruses use reverse transcriptase to transform RNA into DNA) and DNA to protein (proved possible *in vitro*, but not in living organisms) [39].

To study these processes in depth, one needed the ability to read the sequence of nucleic acids and proteins. Around that time, first DNA sequencing efforts took place. After their development, in the 80s and 90s, the efforts focused on sequencing the human genome, resulting in Human Genome Project [87], completed in 2003. On the other hand, the application of mass spectrometry to study proteins became popularized in the 80s, greatly facilitating research on proteins. Thus the focus was on starting and end points of protein synthesis, ignoring the middle.

In the early 2000s, RNA sequencing launched as well. This caused the spotlight to fall on transcription and its regulation, which led to discovery of intricate mechanisms regulating gene expression and the role of transcriptional control in disease [104]. All the while, translation has been neglected in comparison. While the levels of mRNA and proteins correlated to some extent, that did not explain everything. Only after the year 2009, which brought on the development of ribosome profiling, the high-throughput sequencing of ribosome protected mRNA fragments [84], the translational regulation came into light.

Translational control, the last step of protein synthesis, is the topic I will focus on in this thesis. While a lot has been researched on regulatory mechanisms of translation in the recent years [77], the field is still fairly fresh and unploughed.

1.1 mRNA translation

Unlike prokaryotic bacteria, which can translate an mRNA directly after transcription, an eukaryotic transcript undergoes several maturation steps. In eukaryotic organisms transcription and translation happen in separate cell compartments, nucleus and cytoplasm. Before a mRNA is ready to be translated into a protein, it has already been processed by (1) capping, an attachment of 7-methylguanosine residue to the 5' terminal of the transcript (5'cap), (2) polyadenylation, an addition of poly-adenosine (Poly-A) tail to the 3'end of the transcript, (3) RNA splicing, the removal of non-coding RNA introns and joining together the exons to form the mature mRNA and (4) optional modifications, altering the chemical composition of the ribonucleic acid. The cap and tail are used for stability, the first serving as an attachment point of the ribosomes, the latter guiding the transcript so that it can find its way out of the nucleus. Some genes can be alternatively spliced, producing different mature mRNAs, which is especially common in higher organisms. RNA modifications have the potential to alter function and stability of the transcript.

Translation is performed by a ribosome, made up of ribosomal RNA (rRNA) and a set of distinct ribosomal proteins, arranged in two ribosomal subunits: small (SSU) and large (LSU). Historically, the size of ribosomal parts has been measured in *Svedberg* units, a measure of sedimentation rate. The size of bacterial ribosome is 70S where the small subunit is of size 30S and large 50S. The main rRNAs are 16S, located in the SSU and 5S and 23S in the LSU. Eukaryotes have 80S ribosomes, with the sizes 40S and 60S for SSU and LSU, respectively. 40S subunit contains 18S rRNA, while 60S has 5S, 28S and 5.8S rRNAs [4]. In addition to rRNAs, the bacterial ribosomes contain 52 and eukaryotic 82 ribosomal proteins, although we now know that the number can vary [173].

The large subunit of the ribosome contains three active sites where the translation occurs, capable of binding transfer RNA (tRNA) molecules, carrying amino acids which will form a peptide. A-site binds aminoacyl-tRNA (aa-tRNA, a tRNA with amino acid attached to it), peptidyl bond between added amino acid is formed at P-site (peptidyl transferase center) and E-site (exit) binds free tRNA before it exits the ribosome. The peptide moves through the exit tunnel, which spans from the P-site to the cytoplasmic surface of the LSU. It is about 100 Å and can accomodate up to 40 amino acids-long nascent chain before it emerges from the ribosome, depending on the co-translational folding of the protein which happens already inside the tunnel [142, 161]. The ribosome is schematically represented in the Figure 1.2.

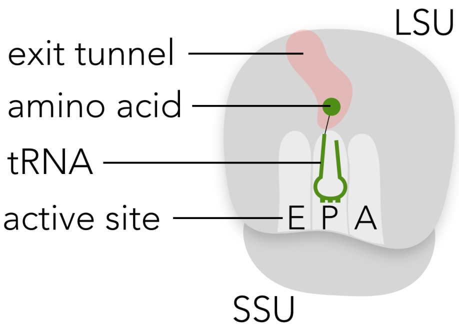


Figure 1.2: **Schematic representation of a ribosome with a bound aa-tRNA.** The ribosome consists of two subunits, small (SSU) and large (LSU). The tRNA binds to active sites of the ribosome (A, P and E), and the nascent peptide is formed in the exit tunnel spanning from the P-site to the cytoplasmic surface of the LSU.

1.2 Phases of translation

The process of translation is composed of four main phases: initiation, elongation, termination, and ribosome recycling. The core aspects of translation are highly conserved between bacteria and eukaryotes, with some substantive differences between the two. The generalized model, applicable for both, is shown in the Figure 1.3. Every mRNA transcript contains one or more open reading frame (ORF), consisting of triplets of nucleotides, codons. It is initiated with a start codon (usually AUG) and terminated with a stop codon (UAA/UAG/UGA). Eukaryotic genes typically contain stretches of nucleotides upstream of the start site and downstream of the termination site, the so-called untranslated regions, 5'UTR and 3'UTR (5' being present in some bacterial genes as well). The ribosome translates ORFs - decoding the codons by pairing with tRNAs, which contain anticodon sequence, complementary to the codon being translated. tRNAs bear amino acids coded for the codon (in total there are 64 codons, 61 ′sense′ codons, encoding amino acids, and 3 stop codons). When the stop codon is reached, the ribosome releases the nascent protein, afterwards it dissociates into subunits and is available for another round of initiation. At each phase, ribosomes form transient complexes with auxiliary translation factors that facilitate protein synthesis.

1.2.1 Initiation

During translation initiation, ribosome recruits an mRNA transcript and finds the start codon of the ORF, typically AUG (in some cases near-cognate start codons [68, 97]) coding for methionine (Figure 1.3 A). The process of finding the start codon is quite different in bacteria and eukaryotes.

In bacteria, initiation occurs co-transcriptionally, with the ribosome and the RNA polymerase interacting with each other. The ribosome binds to the ribosome binding site (RBS) on an mRNA, as soon as it emerges from the polymerase. The most commonly studied mRNAs are those containing Shine-Dalgarno (SD) sequence located 8-

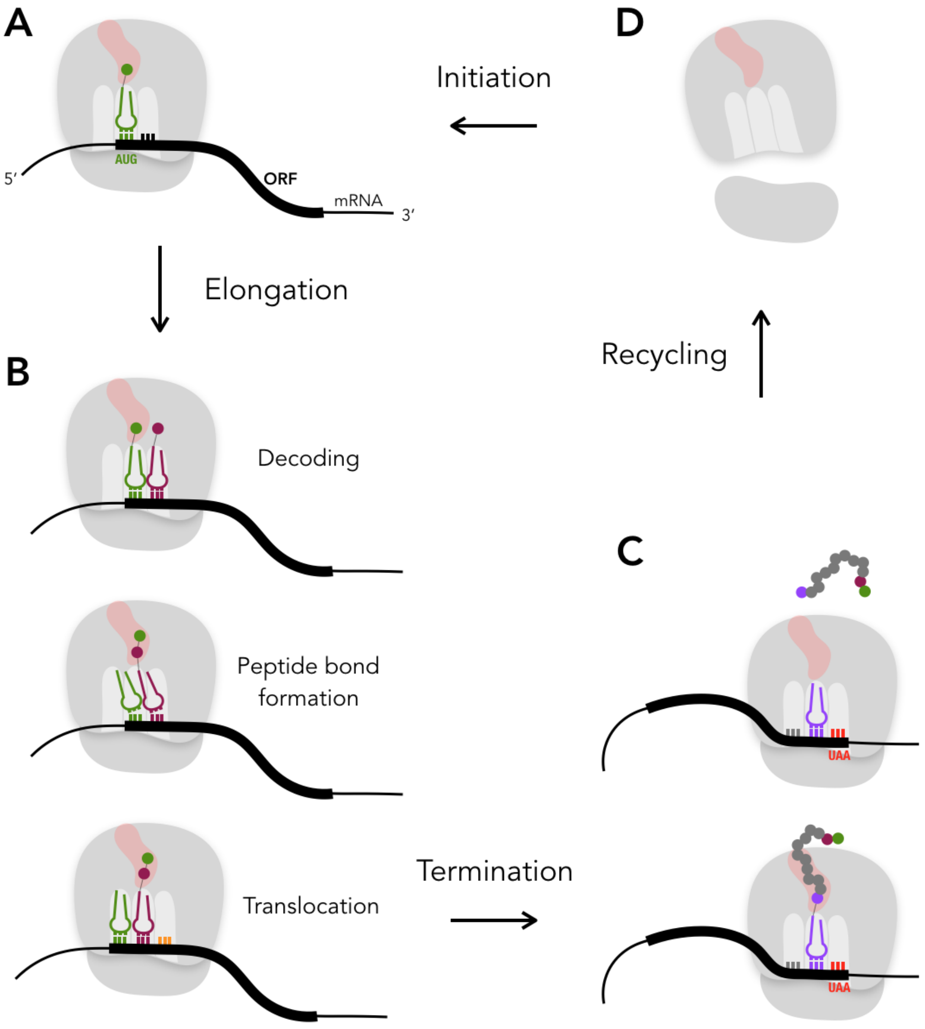


Figure 1.3: **Phases of translation.** (**A**) Translation begins with initiation, where the mRNA (black), ribosomal subunits (gray), Met-tRNA (green) and initiation factors (not shown) assemble at the start codon of an ORF (thick black line). (**B**) Next, elongation involves repetitive steps of decoding (docking of a tRNA with anticodon complementary to the codon in the A-site), peptide bond formation (between aa-tRNAs in the P and A-sites) and translocation (the free tRNA exits at the E-site, the decoded one moves to a P-site and the next codon on mRNA is ready to be decoded in the A-site). (**C**) When the ribosome reaches stop codon in the A-site (red), it is recognized by release factors (not shown), which trigger the release of the peptide. (**D**) Finally, the ribosome subunits are dissociated and recycled for the next round of initiation.

10 nt upstream of the start codon (on an extended 5'UTR region), which ensures correct positioning of the start codon on the small subunit. The whole RBS spans nucleotides -20 to +15 around the start codon. The 30S SSU is recruited to the site through interactions between the SD and anti-SD sequence on 16S rRNA. The initiation is promoted by initiation factors, IF1, IF2 and IF3. Although the most studied, not all bacterial transcripts contain SD sequence (or a 5'UTR), with its prevalence varying from around 12% to 90% depending on the genome [28].

In typical, SD-led mRNAs, the SSU with three initiation factors and Met-tRNA recruit the mRNA and recognize start codon, forming 30S initiation complex. Joining of the 50S LSU triggers dissociation of initiation factors, settlement of Met-tRNA in the P-site and formation of 70S initiation complex, which is ready for elongation.

The initiation is said to be the limiting step in translation. It can modulate translation efficiency by the type of start codon (whether AUG or near-cognate GUG or UUG, where the binding is weaker), the strength of SD sequence and its distance to start codon, mRNA secondary structure near the start site and A/U rich elements in the mRNA, bound by the largest ribosomal protein, bS1, required for binding and unfolding mRNA structure [51, 68, 142].

Initiation in eukaryotic organisms typically follows the scanning mechanism. The 40S SSU loaded with Met-tRNA and eukaryotic initiation factor 2 (eIF2) form 43S preinitiation complex (PIC), in a reaction promoted by initiation factors eIF1, eIF1A, eIF5, and eIF3. The eukaryotic initiation factors display activities that resemble bacterial ones, but there is very little sequence homology among them [142]. The complex attaches to the 5'end of mRNA, preactivated by association with eIF4F at the cap and poly(A)-binding protein bound to the poly(A)-tail, forming 48S PIC. The 48S PIC then scans 5'UTR base by base for complementarity to the anticodon of Met-tRNA, as successive triplets of nucleotides enter the P-site. Upon encountering the first AUG codon, the 60S LSU, stimulated by eIF5B, joins the complex, forming 80S initiation complex. However, if the context of the first start codon does not conform to the ′Kozak′ consensus, featuring a purine (A or G) at position -3 and guanine at position +4 relative to the AUG (at +1), the codon can be skipped and the scanning continues until it encounters a start codon in a favorable context. Once the 80S initiation complex is formed, the elongation phase commences [78].

1.2.2 Elongation

Elongation involves repetitive cycles of decoding, peptide bond formation, and translocation. It begins as soon as the second codon of the ORF becomes available for reading by amino acid-loaded tRNAs (aa-tRNAs) and ends when the ribosome arrives at the stop codon. The basic mechanism is very similar in bacteria and eukaryotes, and is facilitated by homologous elongation factors (EF-Tu/eEF1A, EF-G/eEF2, EF-P/eIF5A, SelB/ EFsec for bacteria and eukaryotes respectively) [44, 142] (Figure 1.3 B).

Decoding

During decoding, the ribosome translates the sequence of codons in an ORF into the sequence of amino acids, forming a protein. A codon, when exposed in the A-site, is recognized by anticodon sequence of aa-tRNA which is delivered to the ribosome in complex with elongation factors. The fidelity of aa-tRNA selection is high, yet errors do happen with frequencies 10^{-3} or less [5]. The accommodation of aa-tRNAs decoding rare codons is slower than of those that occur frequently in the genome, which is believed to be one of modulators of the speed of translation [133, 135].

Peptide bond formation

In the P-site, the peptidyl transferase center of the ribosome, the peptidyl-tRNA in the P-site and aa-tRNA in the A-site react to form a peptide bond. The peptide bond is formed by nucleophilic attack of the amino group of aa-tRNA on the carbonyl carbon of the ester bond in peptidyl-tRNA. The reactivities of amino acids in this reaction vary substantially [168]. A notable one is the synthesis of poly-proline stretches with three or more consecutive prolines or of XPPX sequences with two prolines flanked by specific amino acids [76, 132]. The low rate of peptide bond formation causes ribosome stalling, which can be alleviated by EF-P in bacteria and its homolog eIF5A in eukaryotes, specialized translation factors that enter the E-site and forces P- and A-site substrates towards catalitycally productive orientation [23].

Translocation

After peptide bond formation, the ribosomal subunits move relative to each other, from the non-rotated state with the two tRNAs bound at P- and A-sites the rotated state with the tRNAs bound in hybrid P/E and A/P states. The ribosome can adopt several intermediate conformations, with different positions of the tRNAs respective to the SSU and the A and P-site loops on the LSU. After unlocking the codon-anticodon complexes, the E-site tRNA moves further away from the P-site tRNA, accompanied by the loss of the E-site codon-anticodon interaction, and finally dissociates from the ribosome.

During translocation, secondary structure of mRNA coupled with slippery sequences (ones where codons in the 0- and 1-frame code for the same tRNAs), can lead to -1 frameshifting, causing decoding of the rest of the transcript in a wrong reading frame [44, 142].

1.2.3 Termination

When the elongating ribosome encounters a stop codon on an mRNA, termination occurs. There are no tRNAs recognizing stop codons, instead, this is the role for termination (or release) factors. In bacteria, these are RF1 and RF2, reading UAG/UAA and UGA/UAA codons, respectively. The third, RF3, facilitates turnover of the other two. In eukaryotes, all three stop codons are recognized by a single release factor, eRF1. The shape and size of release factors is similar to that of tRNAs, therefore it can bind to stop codon in the A-site. After recognition of stop codon, the ester bond of the peptidyl-tRNA hydrolyzes in the P-site and the release factors dissociate from the ribosome [142, 147] (Figure 1.3 C).

1.2.4 Ribosome recycling

After termination, the ribosomes release mRNA and tRNA. The ribosome subunits fully rotate and finally split, so that they can be reused in the next round of translation [142, 147] (Figure 1.3 D).

1.3 Translational control

For a cell to function properly, it has to synthesize proteins at the right time, place and in the right amount. The process of protein synthesis is energetically expensive for the cell. Therefore, to conserve energy, it needs to be highly regulated. The control of translation is a vital contributor to cell homeostasis, differentiation and proliferation, at the same time its dysregulation causes many disease states [140, 160]. Therefore it is essential to understand the mechanisms of translational control in depth.

The balance between protein synthesis and degradation rates determines the cellular level of proteins - at least to some degree. Transcript levels alone are not precise estimators of protein levels, as some mRNAs are not actively translated, or are translated poorly, being bound to ribonucleoproteins, or sequestered in stress granules or processing bodies [92]. Additionally, the rates of translation vary among different mRNAs as well as along individual transcripts [16, 85].

How many ribosomes translate a specific mRNA and how fast the peptide bonds are formed determine the synthesis rate of the given protein. The number of ribosomes in turn is determined by the number of active mRNAs and their ORF length, the rate of ribosome initiation, elongation, and termination/recycling. The initiation is traditionally believed to be the rate-limiting step, as it is the least energetically expensive (ribosomes require a lot of energy to be synthesised, so it makes sense to limit the number of ribosomes to be engaged with the mRNA in the first place). However, inhibition of elongation can rapidly decrease protein levels and conserve energy when needed [77, 166]. Similarly, degradation of transcripts coupled to or influenced by translation can regulate mRNA levels [74, 96]. The different control points throughout mRNA are summarized in Figure 1.4.

The factors influencing initiation rate include (1) mRNA secondary structures that affect interactions with the translational machinery, (2) trans-acting factors like proteins, small RNAs (microRNAs), riboswitches/ligands (riboswitches are regulatory segments of mRNA that bind a ligand, causing mRNAs to autoregulate their own activity) that bind specific mRNAs and enhance or inhibit recruitment of ribosomes [20, 48], (3) availability of ribosomes, initiation factors, and Met-tRNAs in the cell and (4) post-translational modifications of initiation factors (e.g. by phosphorylation) [136].

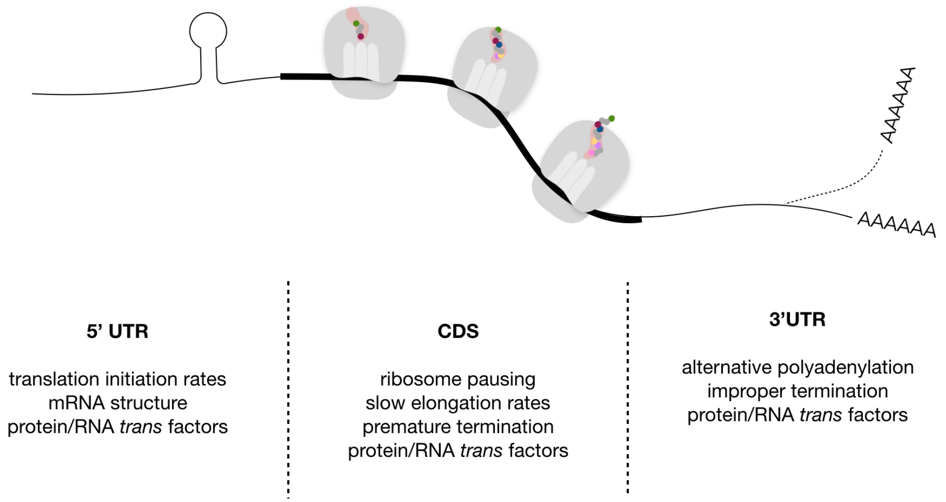

AAAAAA
AAAAAA
5' UTR
translation initiation rates
mRNA structure
protein/RNA *trans* factors
CDS
ribosome pausing
slow elongation rates
premature termination
protein/RNA *trans* factors
3'UTR
alternative polyadenylation
improper termination
protein/RNA *trans* factors

Figure 1.4: **Regulatory points throughout mRNA.** In the 5'UTR, initiation rates, RNA structure and trans-acting factors can regulate mRNA translation and decay. In the CDS (thick black line), events such as pausing of translating ribosomes, premature termination, slow elongation rates and protein/RNA trans factor interactions can alter elongation speed. The 3'UTR can modulate translation and decay by interactions with trans factors, different isoforms resulting from alternative polyadenylation, or improper termination, such as defective ribosome or lack of a stop codon. All these ensure accurate and efficient protein synthesis. Modified from [74].

Elongation is typically quite fast (around 5.6 amino acids per second [86]), although slowed elongation can reduce initiation as well, as the ribosomes bound at the start site must move at least five codons downstream to make space for a new initiating ribosome. Furthermore, adjustment of elongation rate by rare codon usage, tRNA levels, or phosphorylation of elongation factors [44, 133, 135, 136] can be used to control co-translational protein folding or targeting [77]. Regulation of elongation is described in more detail in section 1.4.

During termination in eukaryotes, if release factors are not available, the terminating ribosome may trigger nonsense-mediated decay (NMD). NMD is a mRNA surveillance pathway, trigged by inefficient translation termination, as well as degrading defective mRNAs containing premature termination codons [96]. Additionally, miRNAs bound in 3'UTRs of mRNAs can promote mRNA decay by recruiting deadenylases and decapping factors onto the target mRNAs, hence making them unavailable for translation [93]. Finally, different transcript isoforms can have 3'UTRs of varying lengths (alterntive polyadenylation). Shortening of the 3'UTRs can alter RNA structure or eliminate miRNA or protein binding sites, affecting mRNA stability and termination [74, 114].

1.3.1 Translational deregulation in human disease

Tight control of mRNA translation is vital in the regulation of gene expression in embryonic and adult tissues to ensure healthy development and functioning. Defects in the translation process, for both nuclear and mitochondrial genes, are deleterious for physiology and development of an organism. Translation-related human disorders can be categorized into four groups: those involving deregulated tRNA synthesis or function, ribosomopathies (caused by defects in ribosome biogenesis and function), deregulation of the integrated stress response pathway (which senses diverse cellular stresses and mediates changes in gene expression to adopt to stress) and mTOR pathway (intracellular signalling pathway important in regulating the cell cycle). The deleterious effects manifest itself as a wide range of diseases, such as immunodeficiency, metabolic disorders, neurological disorders and cancer, as well as during virus infection [160].

Cancer holds a special place among the mentioned diseases, as many oncogenes are regulated at the level of translation. The rapid and continuous proliferation of cancer cells requires rapid and continuous protein synthesis and increased ribosome content, increasing energy expenditure for protein synthesis. Most tumor cells are already under physiological stresses (which down-regulate mRNA translation in healthy cells), but also become uncoupled from regulation which further stresses the cell. Cancer cells hijack the translational machinery for their sustained proliferation, survival and metastasis (spread) to distant tissue sites [140].

All in all, why studying translational regulation is so important? A detailed understanding of translational control machinery will lead to thorough understanding of disease pathogenies and, hopefully, therapeutic opportunities.

1.4 Elongation in translational control

Of the four major phases of translation, research has mostly focused on initiation as the rate-limiting step of translational control. Nevertheless, growing evidence shows that all four phases are important for maintaining balance in protein production, folding, trafficking and degradation, with elongation being a central determinant to protein fate.

The regulation at elongation might be advantageous in that it (1) allows rapid response to stimulus and (2) protects mRNAs from nucleases. In the first case, loading an mRNA onto ribosomes and then stalling them in response to a certain stimulus would provide an instant on/off switch in protein production, which would be important in situations where immediate response is needed, such as e.g. neurons reacting to synaptic stimulation. Secondly, RNAs tend to be degraded when not associated with protein factors, therefore loading them onto translationally quiescent ribosomes might protect them from degradation by nucleases [139].

A translating ribosome can accelerate, slow down, pause and stall during elongation, resulting in non-uniform elongation rates [141]. Various factors influence the speed of translation elongation, and the modulation of speed has its own downstream effects (Figure 1.5). Each of these is described in detail below.

1.4.1 Upstream determinants of elongation rate regulation

The coding region of the mRNA may contain regulatory signals defining local elongation rates, leading to translation bursts and pauses. Here the terms ′pause′ and ′stall′ are used interchangeably, meaning a ribosome pile-up at certain codons. Such bursts and pauses may be caused by a variety of factors, such as codon-specific rates of cognate aa-tRNA delivery to the ribosome and their abundance, codon context, aminoacyl moieties attached to the tRNAs in the P- and A-site attenuating the rates of peptide bond formation, amino acids in the nascent peptide interacting with ribosome exit tunnel or mRNA secondary structure blocking the translating ribosomes in their way [141, 156].

Codon usage and tRNA abundance

The genetic code is degenerate, with the same amino acids being encoded by different synonymous codons - 61 sense codons coding for 20 amino acids. The frequencies of synonymous codons can vary between rare and common codons by over an order of magnitude [26], as well as between species, defining codon usage of a given organism.

Codon usage dictates the dynamics of elongation mainly by its interdependence with tRNA abundance [26, 81, 137]. Hundreds (or even thousands, e.g. in zebrafish) of tRNA genes make up for a significant variation in the number of tRNAs that can decode a given codon. Some codons can be decoded only by one cognate tRNA, others have dozens of possible isoacceptor tRNAs (same anticodon, but variation elsewhere in the tRNA sequence), and still others do not have a cognate tRNA with exact Watson-Crick base pairing and require wobble interactions at the third base of the codon. Thus

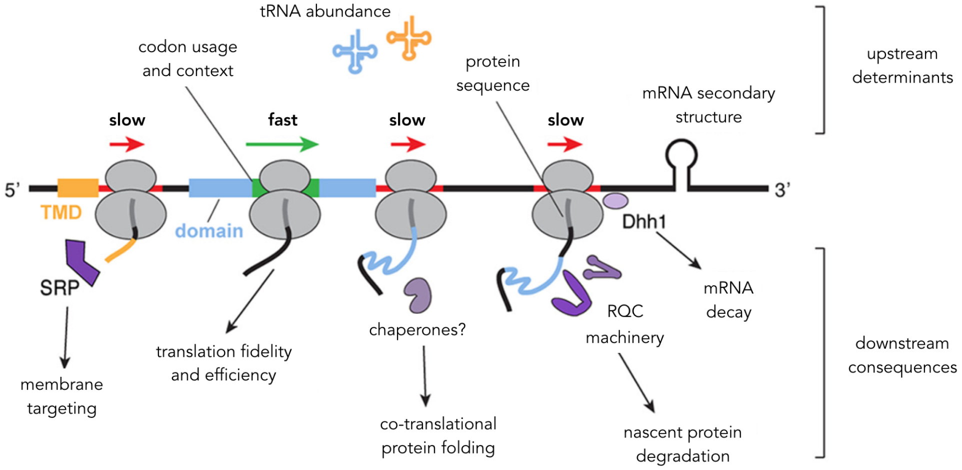


Figure 1.5: **Factors regulating translation elongation and their downstream consequences.** Multiple upstream variables modulate translation speed, including codon usage and tRNA abundance, protein sequence of the nascent peptide and mRNA secondary structure. Translation speed in turn modulates downstream pathways that determine the fate of nascent proteins. While fast it increases translation fidelity and efficiency, while slow - facilitates subsequent processes. These include membrane targeting, after emergence of targeting signals such as transmembrane domains (TMD) from the ribosome tunnel, as well as recruitment of chaperones for co-translational protein folding, RQC machinery for nascent peptide degradation or initiation of mRNA decay. Modified from [156].

the speed of decoding depends on the abundance of cognate tRNAs. In general, an aa-tRNA that is cognate to a given non-optimal codon constitutes only a small fraction of the total aa-tRNA pool, therefore near- and non-cognate aa-tRNAs can compete with the cognate ones for the initial binding to the ribosome until the A-site codon is recognized. Conversely, optimal codons are decoded by abundant tRNAs causing translation to proceed more efficiently [21, 57, 141, 156, 169]. Of note, the composition of the tRNA pool can change in response to altered cellular conditions and disease [62, 65, 145].

Arrest peptides

Nascent peptide sequence can regulate the kinetics of translation as well [167]. First, the efficiency of addition of incoming amino acids influences the rate of elongation. Proline is both a poor acceptor of a peptidyl moiety in the A-site as well as poor donor in the P-site, slowing down translation when present in these positions [112, 123, 127, 168]. This effect is particularly pronounced for proline-rich motifs, promoting ribosome stalling during elongation and termination and necessitating the recruitment of elongation factor eIF5A for progression [70, 148].

Second, poly-basic amino acid stretches have been shown to contribute to ribosome slowdown in one study [30]. This has been observed when positively charged peptide segments were situated 10-20 Å away from the peptidyl-transferase center where the ribosomal proteins constrict ribosome exit tunnel, likely because of the interaction with the negatively charged walls of the tunnel [111]. However, other studies have disputed against these findings [9], found only a subtle effect and only in the absence of translation inhibitors [138] or even found negatively charged amino acids contributing to stalling in certain conditions [144].

Third, stretches of consecutive AAA codons, encoding for lysine, have shown yet another effect on translation. Peptide bond formation between the two consecutive Lys is very slow, resulting in ribosome stalling. Stalling in turn may trigger ribosome sliding on multi-A motif, such that when the ribosome resumes translation, it may shift into a different reading frame [100]. Around 2% of human genome may be regulated in this way [8].

mRNA secondary structure

One of the least studied factors influencing translation rate are the mRNA secondary structure elements. Structure at the initiation sites highly correlates with efficiency [135], yet its role along the CDS is less clear. It has been shown that codons cognate to high-abundance tRNAs are preferentially used in highly structured regions, while regions composed of rare codons contain little structure, thus smoothing the overall translation rate [66].

Chemical probing of *in vivo* structure revealed that only a small fraction of transcripts are structured in the cell, compared to *in vitro* [143]. On a global level, translation guides structure rather than structure guiding translation, with the ribosome being the major remodeler [13]. Additionally, thermodynamically stable mRNA secondary structure elements, such as stem-loops and pseudoknots are known to slow down ribosomes at sequences that cause programmed ribosome frameshifting [24, 32, 99]. The transcript structure also depends on the interactions with RNA-binding proteins, which may stabilize or disrupt secondary structure elements [141].

1.4.2 Downstream consequences of elongation rate regulation

Regulation of elongation rate has several biological consequences. On the one hand, fast translation of protein structural elements ensures high fidelity and increases translational efficiency. On the other hand, transient ribosome pausing allows time for recruitment of various machinery and facilitates subsequent processes, such as membrane targeting or co-translational protein folding. If the ribosome stalls due to aberrant translation, it may trigger recruitment of machinery to degrade the nascent peptide and/or trigger mRNA decay.

Translation fidelity and efficiency

The rate of elongation affects the translation efficiency and fidelity in several ways. First, increased translation speed allows for faster ribosome turnover and efficient loading of the ribosomes onto the mRNA. Conversely, stalling at the beginning of the transcript may result in ribosome queuing and inhibit translation initiation. Second, regulatory signals enhancing or slowing translation rate modulate the local translation rates and facilitate co-translational processes. Third, the local changes in elongation rate may alter fidelity of translation and affect the quality of the protein product, resulting in incorrect or misfolded proteins, that need to be degraded by the quality control pathways. It is very difficult to point to the exact determinants of the observed effects, as most of the effects can be caused by parallel confounding factors [141, 156].

Membrane targeting

The slowdown of translation may allow time for the ribosome to sense the nature of the nascent protein chain and recruiting components that would assist in its maturation and targeting [130]. For example, decreased translation rate promotes recruitment of the signal recognition particle (SRP) [128]. A rare codon cluster is present 35-40 codons downstream of the targeting signals, thus slowing translation as the signal recognised by the SRP emerges from the exit tunnel. Once bound, the SRP co-translationally targets secretory proteins to the endoplasmic reticulum [31, 128].

Co-translational protein folding

For many proteins, folding begins while they are still being translated, upon emergence of the nascent peptide from the ribosome exit tunnel. Co-translational folding is modulated by the factors determining elongation rate, particularly codon optimality [156]. In general, translation needs to slow down to facilitate protein folding. In return, folding exerts force on the nascent chain [40, 64], which supports a ′tug-of-war′ hypothesis in which translation slows down to facilitate folding, in turn successful folding pulls the nascent chain, relieving stalling and accelerating translation [156].

One way in which slower translation facilitates protein folding is by regulating synthesis of structural elements. The distribution of synonymous codons along mRNA is not uniform and non-random. Clusters of optimal and non-optimal codons are conserved in equal measure and are position-dependent. In particular, optimal codons are often enriched in regions where accurate translation is required, such as highly conserved parts of structural domains [105, 175]. In contrast, non-optimal codon clusters tend to be located at structural boundaries, such as linker regions downstream of protein structural domains or separating smaller secondary structure motifs within the larger domain [26, 27, 129]. Slowing down translation at these points would allow time for the domains to fold into lower-energy folding intermediates, before more of the protein is synthesised [156].

Small structural elements can fold while the nascent peptide is still within the exit tunnel. In the wider part of the tunnel near the exit port, the nascent chain can adopt α-helical structure or fold small protein domains such as zinc-finger domain [113, 124]. As for larger proteins domains, upon emergence from the ribosome the chaperones need to be recruited to provide sheltered folding environment, facilitate native contacts and prevent the formation of non-native contacts. The rate of translation may modulate the folding landscape to influence such contacts [141, 163].

Moreover, it has been demonstrated that enrichment of optimal codons and conserved structural elements may reduce the dependence of nascent proteins on chaperones for proper folding [59, 164]. This provides additional evidence that co-translational protein folding requires local translation slowdown.

Ribosome-associated protein quality control

The cell benefits from removing errors as soon as they are detected. If a ribosome will never succeed in reaching the correct termination codon, the resulting protein product will be truncated and likely defective, or even toxic to the cell [88]. Thus it is advantageous to tag it for degradation while still engaged on the ribosome to ensure rapid elimination and minimize inappropriate interactions in the cytosol. Hence, ribosome-associated quality control (RQC) senses the state of translation and targets stalled ribosomes before they reach stop codon, subsequently eliminating the partially synthesised peptide [19].

Abnormal translation or aberrant mRNA can cause ribosome stalling - the kind which will not resume translation afterwards. Such errors include chemical damage of the mRNA, mRNA cleavage, translation of the 3'UTR or poly(A) tail due to lack of recognised stop codon, excessive mRNA secondary structure or insufficient amounts of particular amino acids or tRNAs. Recognition of the damage by RQC leads to tagging of the nascent chain by e.g. ubiquitylation [14] and subsequent degradation. RQC may trigger mRNA decay as well (described below) [19, 94, 156].

mRNA decay

The three main forms of co-translational mRNA surveillance include nonsense-mediated decay (NMD), no-go decay (NGD) and nonstop decay (NSD). NMD targets transcripts containing a premature termination codon (PTC), NSD targets those lacking a stop codon whatsoever and NGD targets mRNAs containing a range of stall-inducing sequences [152].

In a simple organisms like yeast, NMD is proposed to be induced by recognition of a stop codon upstream of an extended 3'UTR [6, 122]. In more complex organisms, containing introns, NMD tends to depend on the presence of an exon-junction complex (EJC) in the proximity of PTC. EJCs are placed near exon junctions during pre-mRNA splicing in the nucleus. As an authentic stop codon would be located in the 3'exon of the spliced mRNA, the presence of an EJC downstream of the stop codon marks the transcript as a suspect. Once targeted for degradation, the NMD transcripts undergo decay from both $5'$ and $3'$ directions [119, 121, 152]. NMD regulates an estimated 10% of all eukaryotic genes [73, 115], modulating the stability of alternatively spliced mRNAs, mRNAs containing upstream open reading frames and those derived from transposons, pseudogenes or frameshifts [29]. Additionally, NMD is strongly connected to human disease, with some 30% of inherited genetic disorders involving mutations that cause PTCs [55].

NSD eliminates mRNAs that lack a stop codon. The missing stop codon might be because of mRNA truncation, in which case the ribosome simply runs to the end of the transcript, while in those that do not have a stop codon but do contain a poly(A) tail, the translation of consecutive positively charged lysine molecules would stall translation by interaction with the negatively charged ribosome exit tunnel [91]. Following endonucleolytic cleavage (in $5'$-$3'$ direction) upstream of the stall site, an upstream ribosome reaching the cleavage site comes to a secondary stall, targeting the truncated mRNA for additional rounds of decay [74, 152].

NGD targets mRNAs with features that cause the ribosomes to $'$no go$'$ - permanently stall at sense codons. As in the case of NSD, such stalling results in upstream endonucleolytic cleavage, causing secondary stall as in the case of NSD and subjecting the mRNA to additional rounds of mRNA surveillance [74, 152].

1.4.3 Ribosome pausing in health and disease

Regulation of translation elongation has been found to play a role in early development, neural function and cancer. Early development examples have been shown in fruit fly, where *nanos* and *oskar* mRNAs are regulated at translation elongation stage [18, 37]. Stalling has been implicated in regulation of synaptic plasticity (caused by phosphorylation of eEF2 factor or using stalled polyribosomes to bypass the rate-limiting step of translation initiation) [67, 159], fragile X syndrome - a form of autism (caused by stalling a translation repressor, lack of which causes synaptic dysfunction) [155] and neurodegeneration (caused by mutation in a specific tRNA) [90]. Disruption of RQC can lead to protein aggregation and neurodegeneration as well [34, 35]. Finally, stalling has a role in oncogenic transfromation. Some tumor cells can adapt to starvation conditions by upregulating the tumorigenesis/caloric-restriction-induced stress pathway to inhibit polypeptide elongation and promote cell survival [139].

1.5 Ribosome profiling

Ribosome profiling provides an experimental means to studying translational regulation. The basis of this approach is the fact that a ribosome covers tightly around 30 nucleotides of translated mRNA, roughly centred at the P-site. This ′footprint′ is protected from ribonuclease (RNase) activity, which degrades the unprotected mRNA fragments, leaving the footprint intact. The initial studies recovering these fragments were limited to studying one mRNA at the time [157, 170], but coupling it to high-throughput sequencing in 2009 led to development of a transformative technique enabling global analyses of *in vivo* translation [84, 85].

1.5.1 The method

In brief, the experimental protocol for ribosome profiling begins with cell lysis and harvesting under conditions that should preserve the ribosomes in their positions. The lysates contain polysomes, clusters of ribosomes attached to a strand of mRNA which they are translating. Treatment with nuclease digests the parts of mRNAs unprotected by ribosomes, and the ribosomes with short mRNA footprints are recovered by RNA extraction. The purified footprints are reverse transcribed, amplified by PCR and subject to deep sequencing [82]. The sequencing data, after processing and aligning back to transcriptome, result in a ribosome profile over mRNAs. The simplified scheme of experimental protocol is represented in Figure 1.6.

1.5.2 Data processing

The preprocessing and alignment of ribosome profiling sequencing data can be done with tools available for RNA-seq analysis [38]. The main steps would include (1) quality control of raw reads (with e.g. FastQC [1]), with subsequent trimming of adaptors and poor-quality bases (using e.g. FASTX-Toolkit [2] or Trimmomatic [17]), (2) removing rRNA and other short non-coding RNA (ncRNA) contamination (by excluding reads aligning to known ncRNA sequences), (3) transcriptome and/or genome alignment, with aligners suited to mapping short reads (such as Bowtie2 [101] which is also a part of TopHat2 [98], STAR [46] or BWA [108]), and optionally (4) selection of footprint lengths originating from translating ribosomes and (5) getting the data in single nucleotide resolution for sub-codon analyses. The last two steps are specific to ribosome profiling data processing and described in more detail below.

Determining bona fide ribosome protected footprint lengths

Ribosome profiling typically results in a multitude of sequencing reads of varying lengths and origin. Fragment length distribution depends on chosen experimental con-

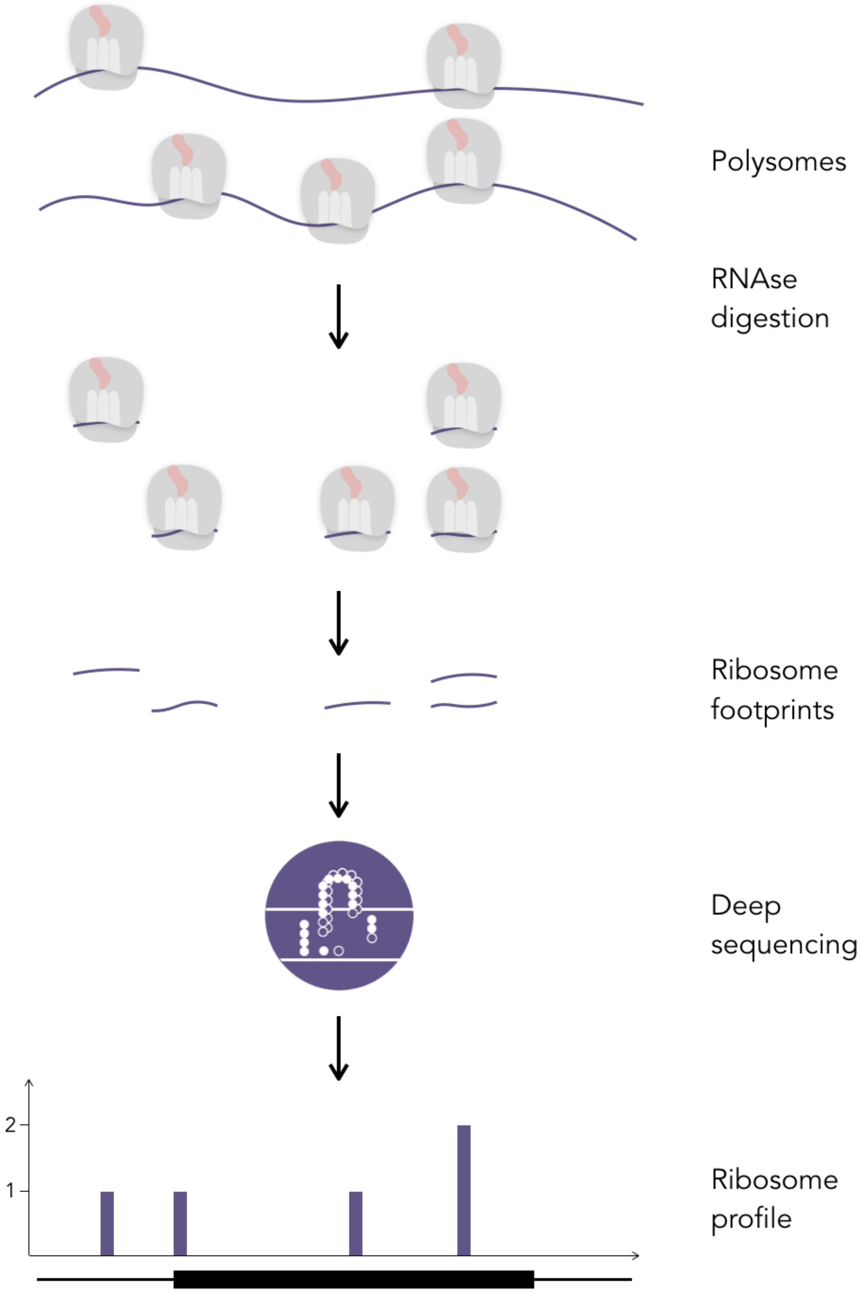


Figure 1.6: **Steps in a ribosome profiling experiment.** Polysomes are isolated from cells and treated with RNAse, digesting mRNA fragments not protected by ribosomes. The remaining footprints are subject to high-throughput sequencing. Aligning the sequenced reads back to transcriptome produces a quantitative profile of ribosome occupancy. Modified from [85].

ditions, such as digestion time and type of nuclease used [118], use of elongation inhibitors [86] and lysis buffer salt conditions [82]. It is also influenced by distinct ribosome conformational state due to stage in translation elongation cycle [102] or stalling on a truncated mRNA [71], position on the transcript [118], and whether reads originate from nuclear CDSs vs mitochondrial or non-coding, e.g. ribosomal or small nucleolar RNAs [83]. Genuine ribosomal footprints display a property of phasing, a strong 3 nucleotide periodicity of the reads stemming from coding regions [12, 84, 117], which can be utilised to isolate them from noise.

Getting sub-codon resolution

Footprint alignments can be assigned to P- and A-site nucleotides by calibrating the offsets from 5' (or 3') ends of the footprints. For each fragment length, footprints overlapping the starts and ends of CDSs are aligned, and the distance from majority of their 5'ends to the start and stop codons reveals the correct offset [84, 171]. Each footprint is then assigned to one nucleotide on the transcripts, shifting the 5'end by its length-specific offset. This sort of processing allows for sub-codon analyses, particularly important in studies on codon decoding rates [42, 43], ribosomal pausing [107, 171], stop codon readthrough [50] or frameshifting [117].

1.5.3 Applications

As mentioned earlier, mRNA and protein levels correlate to some degree. Yet in some cases, levels of some mRNAs were found to anti-correlate with their protein products [33]. Ribosome occupancy provides much better proxy for protein synthesis, as it correlates better with protein abundance than mRNA levels do [33, 110]. This additional, biologically relevant information captured by ribosome profiling can and has been examined to address a wide diversity of questions pertaining to mRNA translation and its regulation (Figure 1.7).

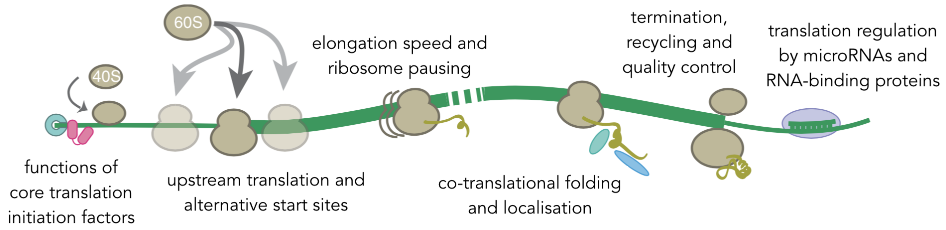

60S
40S
elongation speed and ribosome pausing
termination, recycling and quality control
translation regulation by microRNAs and RNA-binding proteins
functions of core translation initiation factors
upstream translation and alternative start sites
co-translational folding and localisation

Figure 1.7: **Insights into translational regulation from ribosome profiling.** Ribosome profiling served as a basis for studying many aspects of protein synthesis and translational control. Modified from [85].

Translation initiation

Translation initiation is slow compared to elongation, therefore is a rate limiting step. Critical aspect of applying ribosome profiling to study initiation is using drugs that freeze the initiating ribosomes. These include (1) harringtonine, which binds to LSU and forms ribosomal complex with initiator tRNA but blocks aa-tRNA binding in the A-site and so subsequent peptide formation [54], (2) puromycin, inducing premature termination of elongating ribosomes, which results in relative increase in ribosome density at the few first codons of an ORF and (3) lactimidomycin, binding to the initiation complex after its assembly at the start site.

Data produced with the use of such inhibitors would produce prominent peaks at initiation sites, visible in the ribosomal profile. Identification of such peaks with machine learning predictors has been used to map translation initiation sites (TISs) in a variety of tissues and organisms (e.g. [56, 86, 103, 158]). All of these studies found a variety of short open reading frames (sORFs) different from canonical CDSs, the most pervasive being upstream ORFs (uORFs), present in 5'UTRs. These can often repress and sometimes promote translation of the canonical CDS [153] and most of them initiate at non-AUG codons [103] (usually near-cognate, like CUG). Similarly, many alternative protein isoforms have been detected (both extensions and truncations) [56, 86, 103]. Short ORFs have also been found on non-coding RNAs (mainly long, lncRNAs), many of them conserved [10, 12, 25, 53, 116] and few of the resulting ′micropeptides′ have been shown to be functional [7, 45, 126]. The main non-canonical translation events detected by ribosome profiling are shown in Figure 1.8.

Translation elongation

To get an accurate picture of translation elongation, the recovered footprints need to reflect the positions on ribosomes as closely as possible. To capture the exact positions of ribosomes at the moment of experiment, the cells are treated with translation elongation inhibitors to immobilize polysomes prior to cell lysis and nuclease digestion. Nearly all experiment to date have used cycloheximide (CHX) [63], although simple liquid nitrogen freezing or other antibiotics such as emetine (in eukaryotes) and chloroamphenicol (in bacteria) have also been used [84, 86, 125].

Such ribosome profiling data provide many insights into the dynamics of elongation. One of the most important is the speed of ribosomes itself. The estimations of ribosome speed concluded that they progress on mRNAs at an average rate of around 5.6 codons per second [86]. However, the speed is not uniform. The ′snapshots′ of translation reveal regions of higher and lower ribosome density, determining parts of transcripts where the translation is slower and faster, respectively (Figure 1.9).

Many features have been suggested to influence ribosome velocity, including codon bias, tRNA abundance, mRNA folding, amino acid charge of the nascent peptide chain, G:U wobble base-pairing of tRNAs to codons (instead of G:C) [30, 41, 135, 137, 154, 165] but also distinct stages of translation elongation cycle, which are represented by short footprints (20-22 nucleotides instead of typical 28-30), corresponding to rotated conformation of ribosomes [102]. The accumulation of footprints is especially pronounced at specific ′pause′ sites. In addition to the above mentioned, additional factors were suggested to influencing dramatic pausing: proline stretches [9] and specifically in bacteria internal anti Shine-Dalgarno (aSD) sequences on mRNAs and glycines in

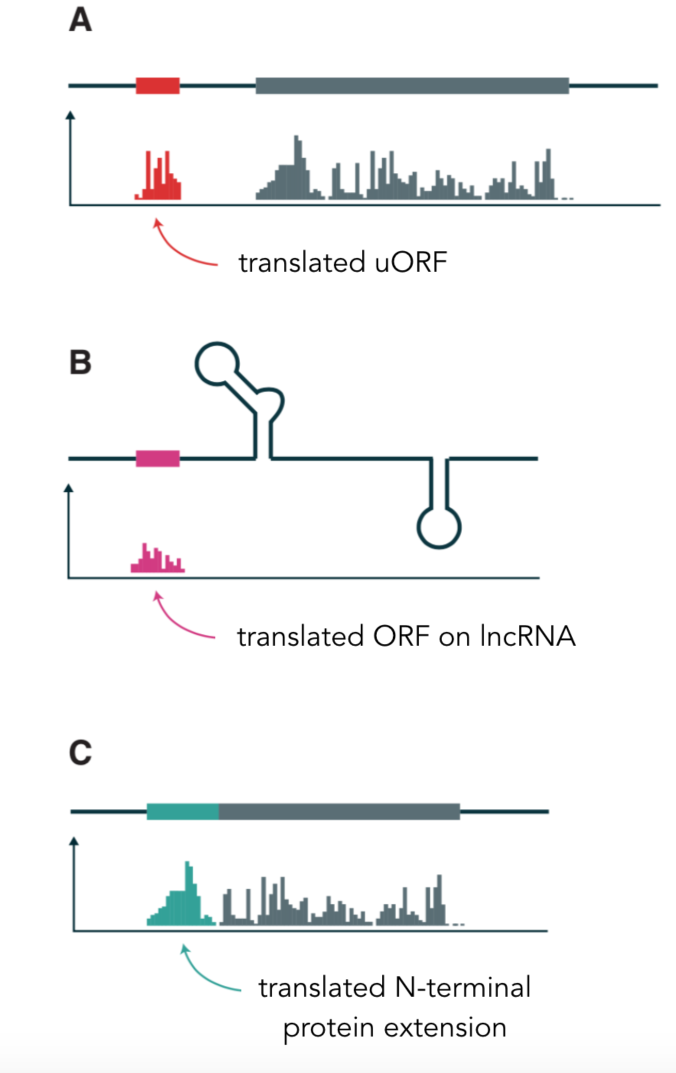


Figure 1.8: **Annotating proteome with ribosome profiling.** The ribosome coverage outside of annotated regions is a sign of translation. The schematics of transcripts (top) and their ribosome profiles (bottom) show most common non-canonical translation events. (**A**) uORFs (red) can regulate downstream canonical CDSs, (**B**) short ORFs found on long non-coding RNAs are often conserved and some code for functional micropeptides, (**C**) N-terminal extensions (and truncations) result in alternative protein isoforms. Modified from [85].

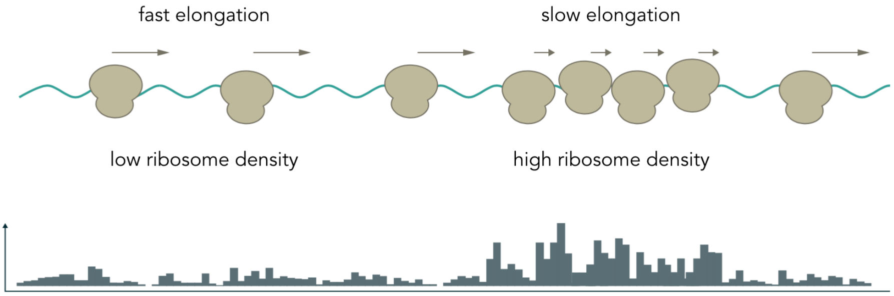


Figure 1.9: **Inferring elongation speed of ribosomes.** The schematics of polysome (top) and their ribosome profiles (bottom). Regions of fast elongation accumulate fewer ribosomes than regions of slow elongation, where the ribosome density is higher. Modified from [85].

the E-site [120].

Co-translational processes

As soon as a nascent chain emerges from the ribosomal exit tunnel, it is engaged by a series of processing enzymes, targeting factors (TF), and molecular chaperones. ′Selective ribosome profiling′, a variation of ribosome profiling coupled technique combined with affinity purification, allows to obtain information on ribosomes that are engaged by specific chaperones or TFs [125]. It has revealed engagement patterns in bacteria (TFs bind to the nascent chain after around a hundred codons have been translated) [125] and eukaryotes [47]. Adaptation of the technique to profile the nascent peptides directly revealed specific pausing at the 5th codon downstream from the start codon, attributed to the geometry of the exit tunnel [72]. Indeed, protein assembly has been shown to be directly coupled to the translation process and involving a multiprotein complex [151]. The co-translational assembly could be coupled to degradation of monomers lacking proteins to form the complex [89] to maintain the stoichiometry of the proteome [106].

Translation termination and beyond

Termination sites are clearly visible in ribosome profiling data of elongating ribosomes, as the latter accumulate at stop codon waiting for dissociation of elongation complex. Although usually 3'UTRs are usually devoid of ribosome profiling reads, there exist some downstream ORFs (dORFs) [117]. Stop codon readthrough causes an accumulation of in-frame footprints downstream of the stop codon [50]. On the other hand, defects in ribosome recycling [75] cause the unrecycled ribosomes to enter 3'UTRs and reinitiate translation in a different reading frame, producing out-of-frame footprints relative to CDS [174].

microRNA-mediated repression and epitranscriptomic modifications

Ribosome profiling, together with measurements of mRNA and protein levels, provided insights into mechanisms of microRNAs (miRNAs) mediating repression and promoting mRNA decay [48]. The most widely supported model is that of miRNA-mediated inhibition of cap-dependent translation initiation. miRNAs were also reported to repress translation post-initiation and before mRNA deadenylation, but the majority of repression seems to be attributable to mRNA decay [48, 52, 69]. How that happens exactly is a topic of current study.

Similarly, ribosome profiling has distinguished the translational effects of mRNA modifications (e.g. adenosine N6 methylation). Epitranscriptomic modifications seem to influence every phase of mRNA translation, as well as alter mRNA levels and their subcellular localization [131].

Profiling of stress and disease conditions

Environmental stressors, be it extremes in temperature, exposure to toxins, or mechanical damage cause cellular stress response. The extent of cell damage and whether it can be repaired or not, depend on severity and duration of stress encountered. If it is short-lasting and mild, cells can re-establish cellular homeostasis to the former state. Less severe cell stress may change cellular responses to subsequent environmental signals, while persistent stress often enhances susceptibility to cancer and ageing-associated diseases [134].

Ribosome profiling has been applied to study changes in protein synthesis in response to amino acid starvation, detecting changes in translational level in a third of analyzed genes [84]. Similarly, in response to oxidative stress the number of genes whose expression changed increased with prolonged stress [61]. Heat shock in turn was found to cause accumulation of ribosomes in the first 200 nucleotides of ORFs [150], while proteotoxic stress (misfolding of proteins) resulted in pausing near the site where nascent peptides emerge from the exit tunnel [109].

mTOR is a kinase involved in cap-dependent initiation. The mTOR pathway is dysregulated in many diseases, particularly in cancer, where the dysregulation causes uncontrollable growth of tumors. A number of genes regulating mTOR are tumor suppressors and oncogenes, therefore of great interest as potential agents for cancer therapy [15, 36, 140, 172]. Ribosome profiling has been employed to characterize translational regulation mediated by mTOR [79, 149, 162] and expose targets for therapeutic intervention.

Chapter 2

Aim of the study

The wealth of ribosome profiling data accumulated in the recent years provides a rich, yet underused resource of genomic information. Given the large amount of biases present in such data, different studies often reach conclusions that are contrary to each other. Therefore it is necessary to account for biases and standardize the processing steps, to increase the amount of quality data and facilitate comparison of different experiments. More quality data in turn means a bigger scope of analyses and, hopefully, better quality results on a more global spectrum.

One of the phenomena that could benefit from increased coverage and scrutiny in analysis is ribosome stalling. So far mainly studied on gene-by-gene basis and with few attempts to analyze it genome-wide, the mechanism remains an enigma in many ways.

The aim of this thesis is to investigate translation regulation by bioinformatic analyses of ribosome profiling data. In particular, it is to develop a processing method for making data from various experiments comparable, and to answer whether ribosome stalling might be a genome-wide regulatory mechanism.

The specific objectives are to:

1. develop a tool for ribosome profiling data processing based on state-of-the-art knowledge
2. process multiple publicly available ribosome profiling datasets and determine the best strategy for obtaining maximum quality, comparable data
3. use the tool to process multi-species data for downstream analyses of genome-wide ribosome stalling and create a comprehensive resource for future in-depth research on stalling

Chapter 3

Introduction to the papers

Paper I: *Shoelaces: an interactive tool for ribosome profiling processing and visualization*

In this paper we tackle the challenge of ribosome profiling data processing, as described in section 1.5.2. To date, every study based on ribosome profiling data has used their own processing, typically using manual and somewhat arbitrary selection of bona fide translating footprints and offsets revealing the active sites of ribosome. *Shoelaces* is an attempt to standardize processing steps, so that the resulting data are devoid of experiment-specific biases and can be compared among each other. All in a user-friendly, visual, interactive manner for people with little to no programming experience, and featuring command line tools and advanced options for integration into automated pipelines for more seasoned bioinformaticians.

Paper II: *Deep conservation of ribosome stall sites across RNA processing genes*

In this paper we take on the phenomenon of ribosome stalling. The sites of major slowdown in translation elongation can be observed as peaks in ribosome profiling data. However, library-specific biases lead to a staggering amount of local variability, which can make stall sites predictions unreliable and the regulatory factor confounded in a large amount of bias. To address these challenges, we (1) process 20 ribosome profiling datasets from five model organism: yeast, fruit fly, zebrafish, mouse and human with *Shoelaces*, to make them comparable and (2) taking advantage of the deep conservation of translation machinery, we check for multi-species conservation of peak positions, thus identifying potentially functionally important stall sites. We find 3293 stall sites that are conserved in at least two organisms. We analyze these present in human and at least one other organism further in terms of both known determinants (section 1.4.1) and consequences (section 1.4.2), as well as potential novel causes and implications. We find proline, glycine and negatively charged amino acids being the main contributors to conserved stalling. Many of the conserved stall sites are found in RNA processing genes, suggesting that stalling might have a regulatory effect on RNA metabolism. Overall, the results of these study provide a rich resource for further in-depth studies of conserved stalling, and indicate possible roles of stalling in translation regulation.

Chapter 4

Scientific results

Paper I

4.1 Shoelaces: an interactive tool for ribosome profiling processing and visualization

Åsmund Birkeland[†], **Katarzyna Chyżyńska**[†] and Eivind Valen

BMC Genomics, 19:543 (2018)

[†] Contributed equally

Birkeland *et al. BMC Genomics* (2018) 19:543
https://doi.org/10.1186/s12864-018-4912-6

BMC Genomics

SOFTWARE

Open Access

Shoelaces: an interactive tool for ribosome profiling processing and visualization

Åsmund Birkeland[1†], Katarzyna Chyżyńska[2†] and Eivind Valen[2,3*]

Abstract

Background: The emergence of ribosome profiling to map actively translating ribosomes has laid the foundation for a diverse range of studies on translational regulation. The data obtained with different variations of this assay is typically manually processed, which has created a need for tools that would streamline and standardize processing steps.

Results: We present Shoelaces, a toolkit for ribosome profiling experiments automating read selection and filtering to obtain genuine translating footprints. Based on periodicity, favoring enrichment over the coding regions, it determines the read lengths corresponding to bona fide ribosome protected fragments. The specific codon under translation (P-site) is determined by automatic offset calculations resulting in sub-codon resolution. Shoelaces provides both a user-friendly graphical interface for interactive visualisation in a genome browser-like fashion and a command line interface for integration into automated pipelines. We process 79 libraries and show that studies typically discard excessive amounts of quality data in their manual analysis pipelines.

Conclusions: Shoelaces streamlines ribosome profiling analysis offering automation of the processing, a range of interactive visualization features and export of the data into standard formats. Shoelaces stores all processing steps performed in an XML file that can be used by other groups to exactly reproduce the processing of a given study. We therefore anticipate that Shoelaces can aid researchers by automating what is typically performed manually and contribute to the overall reproducibility of studies. The tool is freely distributed as a Python package, with additional instructions, tutorial and demo datasets available at https://bitbucket.org/valenlab/shoelaces.

Keywords: Ribosome profiling, Bioinformatics, Genomics, Python, Tool

Background

Ribosome profiling provides the first opportunity to monitor the behavior of translating ribosomes on a transcriptome-wide scale. Since its development [1], the technique has been widely adopted and inspired a diverse range of studies on translational regulation. While the assay itself has been partially standardized, the processing of data has not. A significant bottleneck is that of reproducibility and interpretation. In particular, most studies rely on manual selection of read lengths and manual P-site determination. The choices made are highly variable between studies, biasing the sub-codon resolution or discarding excessive amounts of data, which makes it challenging to compare results in the literature.

The consistent processing of such data necessitates that two major challenges are met: (1) separating signal from noise, i.e. distinguishing footprints of translating ribosomes from reads originating from other processes and (2) determining the specific codon being translated by the ribosome which the read fragment originates from (a P-site offset). While some software tools have been developed for analyzing ribosome profiling data (for an overview see [2]), few address these challenges directly. Instead, tools typically rely on manual selection of read lengths and offsets [3, 4] or perform selection as part

*Correspondence: eivind.valen@uib.no
†Åsmund Birkeland and Katarzyna Chyżyńska contributed equally to this work.
[2]Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway
[3]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway
Full list of author information is available at the end of the article

of an integrated pipeline for open reading frame prediction, with no option to export ribosome coverage after processing [5].

Here, we introduce Shoelaces, a software tool for processing ribosome profiling data. Shoelaces addresses the processing challenges by (1) utilizing a property of phasing, a strong 3-nucleotide periodicity of the reads stemming from coding regions [1, 6, 7] to filter genuine translating footprints and (2) calibrating P-site offsets based on metagene profiles over start or stop codons, stratified by footprint length [1, 8]. Shoelaces automatically selects these lengths and offsets, as well as offers batch-mode for processing multiple libraries in bulk.

The tool can be run in two modes: either using a graphical or command line interface. The graphical interface is accessible to users of all levels and guides the user through each processing step, allows for interactive adjustments and offers a range of extra visualization features on both gene/transcript or global level. The command line interface offers the same functionality as the graphical interface, without the interactivity, and can be easily integrated into automated processing pipelines.

Implementation

Shoelaces is implemented in Python3 and designed to run on Linux and MacOS operating systems. It relies on OpenGL for rendering graphics and PyQt5 for cross-platform graphical user interface. GUI is composed of a set of windows that user can easily rearrange by drag-and-drop to customize layout. The plots are interactive making the processing easily adjustable to specific needs. While primarily designed for the visualization features, Shoelaces can be also run in command line, making it easy to incorporate into processing pipelines. Shoelaces operates on common genomic formats (BAM, GTF, BED, wiggle), and stores settings in XML files, for maximum ease of use and reproducibility of analyses.

Results and discussion

Data processing workflow

The workflow of Shoelaces is shown in Fig. 1. Shoelaces accepts standard genomic formats requiring alignment of ribosome profiling reads (BAM) and corresponding gene/transcript annotations (GTF or BED). Shoelaces then guides the user through three main steps: (1) read filtering, (2) footprint identification and (3) P-site determination.

In the initial step Shoelaces filters reads from noise regions. Here, users can optionally include an additional annotation file with regions (such as e.g. ribosomal RNA or repetitive elements) which will be masked from all further analyses. Specific genes can also be deselected during this step if certain outliers are undesired.

In the following step, Shoelaces automatically determines the correct footprint lengths. This is based on the intrinsic 3-nucleotide periodicity characteristic of ribosome-derived fragments as opposed to reads originating from other processes [7]. The periodicity is detected using discrete Fourier transform (see below) over the coding regions (CDS) of annotated genes. Lengths displaying periodicity are selected for further analysis. The rest is classified as noise but is available for further analysis by the user.

Finally, for each footprint Shoelaces determines the codon that is actively translated. A length dependent P-site offset is calibrated using change point analysis (see below) over the distribution of footprints surrounding start and stop codons of annotated genes. Based on this, Shoelaces will automatically suggest offsets and provide plots of the summed footprints over start and stop codons of all genes. In addition, ribosome footprints are known to map preferentially to the first nucleotide in the codon [1] and Shoelaces therefore displays the fraction of reads falling into each reading frame. Manual adjustment is also possible if deemed necessary by the user.

After confirming the selection of the suggested footprint lengths and offsets, the user can export the ribosome coverage into flat file format (wiggle) for further downstream analysis, either in genomic or transcriptomic coordinates. Optionally, different footprint lengths can be exported into separate files. Separation by length can be useful for more specific analysis, such as e.g. detection of conformational changes of ribosome at certain positions [9, 10].

To aid the researcher, the GUI produces summary statistics and counts for individual genes and transcripts, as well as for the whole library. It provides an overview over how many reads of a given length fall into different genomic regions (CDSs, 5, leaders, 3, trailers and introns) as well as how many footprints are found over non-coding transcripts or mapping to multiple positions in the genome. Users can update the statistics after read length and offset selection to see how they change. Together, these give an indication of the quality of the library and how well the reads represent genuine ribosome protected fragments.

Additionally, Shoelaces can produce expression tables for ribosome profiling data normalized to reads per kilobase of exon per million mapped reads (RPKM). Optionally, if additional RNA-seq data is loaded, Shoelaces calculates translational efficiency per gene as well.

Automatic selection of read lengths and offsets

An ideal-case scenario is presented in Fig. 2: the given footprint length is periodic (Fig. 2d), the metagene pro-

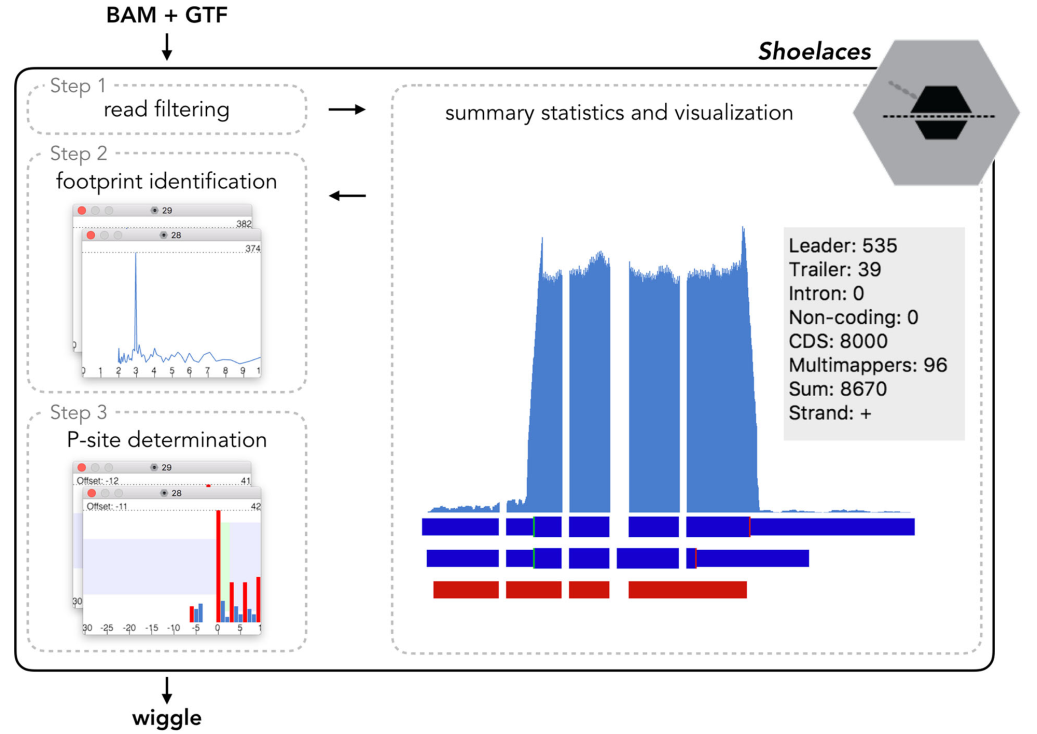


Fig. 1 Shoelaces workflow. The toolkit accepts BAM and GTF files as input, filters reads, identifies translating lengths, determines P-site offsets and outputs tracks into wiggle format. Visual representation and summary statistics aid the processing steps

files have distinct peaks over start and before stop codons (Fig. 2a, b) and reads preferentially map to the first reading frame (Fig. 2c). However, library-specific biases can result in varying distributions of coding footprint lengths, as well as varying offsets (for various examples see Additional file 1: Figures S1-S3). To take these biases into account, as well as to make processing large amounts of ribosome profiling data easy for the user, Shoelaces automatically suggests read lengths and offsets to be used.

Selection of periodic lengths

For each fragment length, the 5, ends of footprints mapping to the first 150 nucleotides of CDSs (by default from top 10% of protein-coding genes with highest coverage) are summed together. As the reads map preferentially to the first nucleotide of every codon, the periodic pattern will be conserved. The resulting vector is subject to discrete Fourier transform, and the fragment lengths whose highest amplitude corresponds to a period of 3 are considered to be periodic.

P-site determination

For each fragment length, the distribution of 5, ends of footprints surrounding start and stop codons (-30/+10 nucleotides) of protein-coding genes is calculated. The resulting window is subject to change point analysis, where for each adjacent position we calculate the difference of means. The maximum shift in means is assumed to correspond to the 5, end of the footprints of initiating ribosomes and the distance from these to the P-site becomes the offset for that fragment length.

Stratification per footprint length covers all different assignment strategies [1, 8], as the effective position of the P-site will be the same, whether calibrated from 5, end or 3, end of the footprint of a given length (see Additional file 1: Figure S5). This accounts for biases in different ribosome profiling libraries, which can have uniform offsets from 5, ends of reads (Additional file 1: Figure S2), or changing in increments of one nucleotide with increasing footprint length from 5’ends, thus having uniform 3, end offsets (with minor variations, Additional file 1: Figures

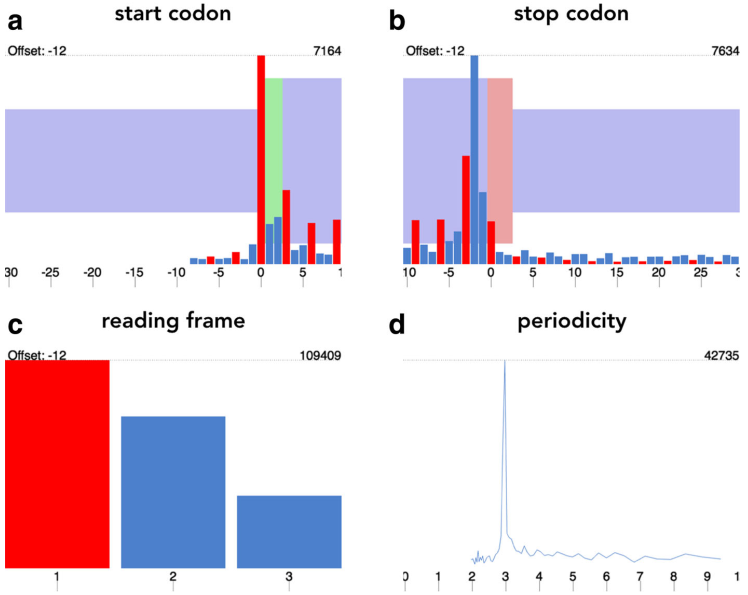


Fig. 2 Read length and offset selection. In an ideal case scenario, the 3-nucleotide periodicity determines if the footprint length is coding (**d**), the peaks over start (**a**) and the last codon before stop (**b**) codons are used to calibrate offsets and most of the reads map to the first reading frame (**c**). Here, the plots demonstrate length 28 in human ribosome profiling sample (SRR493747, [15]). For more plots and datasets see Additional file 1

S1 and S3). Shoelaces offers calibration over both start and stop codons, accounting for libraries where there is no clear peak defined over either end (shorter footprint lengths in Additional file 1: Figures S2 and S3).

Visualization

Shoelaces also allows for visual inspection of coverage over individual genes (or group of genes) of interest. Users can manually zoom in/out to adjust the view, inspect the summary statistics with and without using offsets, and export high quality figures and tracks for further analysis and visualization.

Large-scale processing

For processing multiple libraries in bulk, a batch mode is available. For instance, for a number of same-batch libraries, one can be inspected visually, processing steps stored in an XML file and applied to the others. This additionally makes the processing easily reproducible later on. The processing can also be performed and fully automated from the command line allowing Shoelaces to be a part of a more comprehensive pipeline.

Analysis of human ribosome profiling data

We analyzed 79 libraries of human ribosome profiling data from 12 studies [11–22] and compared our read selection to the original, where applicable. Shoelaces retains up to 32% more data mapping to the coding regions of the genome: CDSs and 5, leaders (see Additional file 1: Table S1) than when originally processed, simultaneously decreasing the relative frequency of non-translating footprints, such as those that map primarily to 3, trailers, suggesting that they might originate from e.g. mRNA-binding proteins, abundant in 3, trailers, secondary structure or other sources of noise (see Additional file 1: Figure S4).

Conclusions

Shoelaces aims for an intuitive and streamlined processing of libraries from different studies and treatments, making them comparable and analysis easily reproducible. The precision in bringing the data to sub-codon resolution is especially important in studies on translational efficiency of different codons [23, 24], but also allows for detection of translational events such as ribosomal pausing [25], stop codon readthrough [3] or frameshifting [6]. The automation and batch processing facilitate dealing with large amounts of data, while visualization features add to user-friendliness and allow for more specific analyses. As we demonstrate on human ribosome profiling data, Shoelaces retains more reads mapping to coding regions than arbitrary manual processing. Overall, Shoelaces is a comprehensive tool for ribosome profiling data processing, and should prove useful to anyone interested in small or large-scale studies on ribosome profiling.

Availability and requirements

Project name: Shoelaces
Project home page: https://bitbucket.org/valenlab/shoelaces
Operating systems: Linux and MacOS
Programming language: Python3
Other requirements: Python3 packages: pysam, numpy, pyqt5, pyopengl
License: MIT license

Additional files

Additional file 1: Analysis examples. **Figures S1-S3**. Three different examples of offset selection (PDF file) for human ribosome profiling datasets: SRR493747 [15], treated with harringtonine and cycloheximide; SRR1039861 [22], treated with cycloheximide; SRR592961 [20], no drug. **Table S1**: Comparison of selected footprint lengths as originally in human ribosome profiling studies and Shoelaces. **Figure S4**: Comparison of reads mapping to different parts of transcript as selected by Shoelaces and the original manual selection (SRR493747 [15]). (PDF 8213 kb)

Abbreviations

CDS: Coding sequence, the coding part of a messenger RNA; RPKM: Reads per kilobase of exon per million mapped reads

Acknowledgements

We would like to acknowledge Guo-Liang "Chewie" Chew for his suggestion of the name Shoelaces.

Funding

This work was supported by the Bergen Research Foundation and the Norwegian Research Council (#250049). The funding bodies did not have any role in design or execution of the study.

Availability of data and materials

The datasets analyzed in the current study are available in the Sequence Read Archive with accession numbers SRP038695 [11], SRP031501 [12], SRP002605 [13], SRP010679 [14], SRP012648 [15], SRP045257 [16], SRP014629 [17], SRP017263 [18], SRP053402 [19], SRP016143 [20], SRP029589 [21], SRP033369 [22]. The demo dataset is available together with the pipeline at https://bitbucket.org/valenlab/shoelaces.

Authors' contributions

ÅB designed and implemented the software. KC implemented the algorithms for the method, tested the software, analyzed the data and wrote the manuscript. EV conceived the pipeline, guided the design and made critical revisions to the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable. This is a tool building on previously published, public data.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

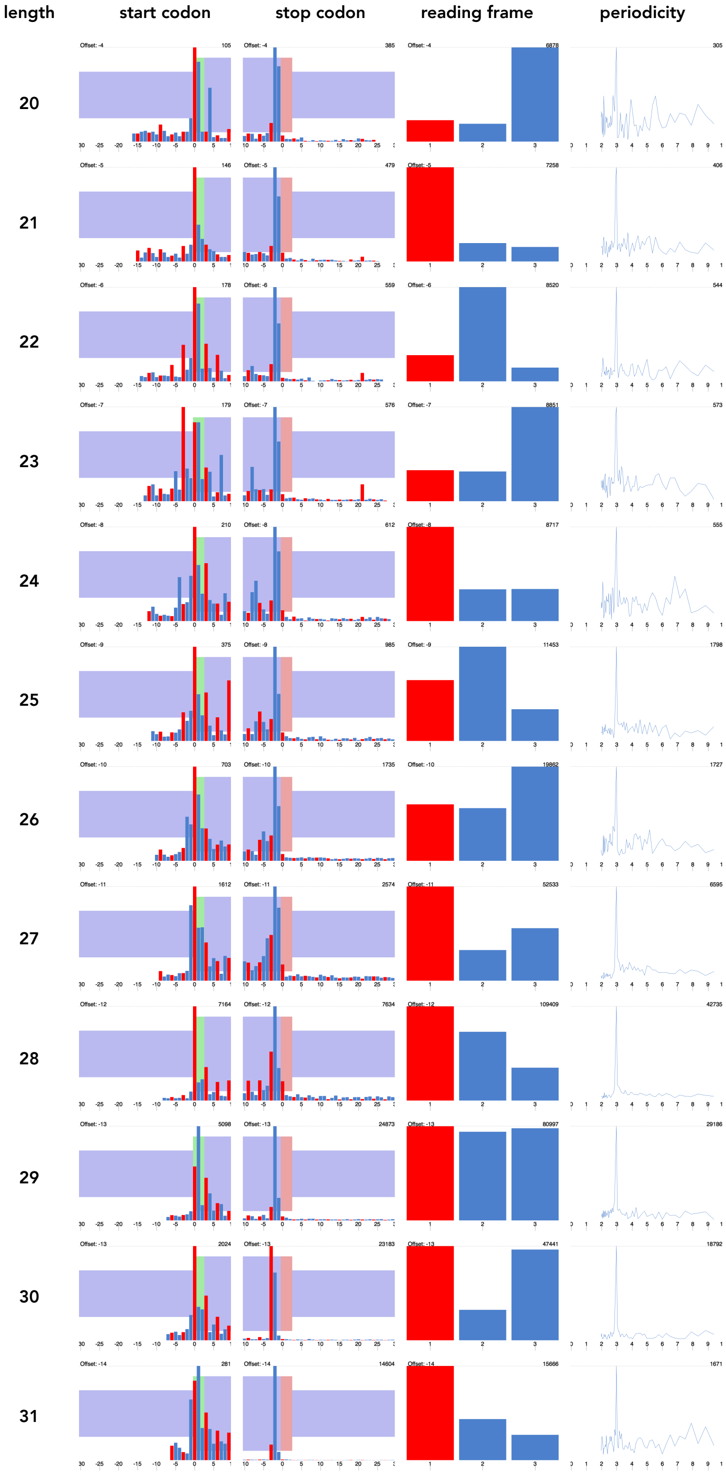
[1]Department of Informatics, University of Bergen, 5008 Bergen, Norway. [2]Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway. [3]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway.

Received: 22 May 2018 Accepted: 2 July 2018
Published online: 18 July 2018

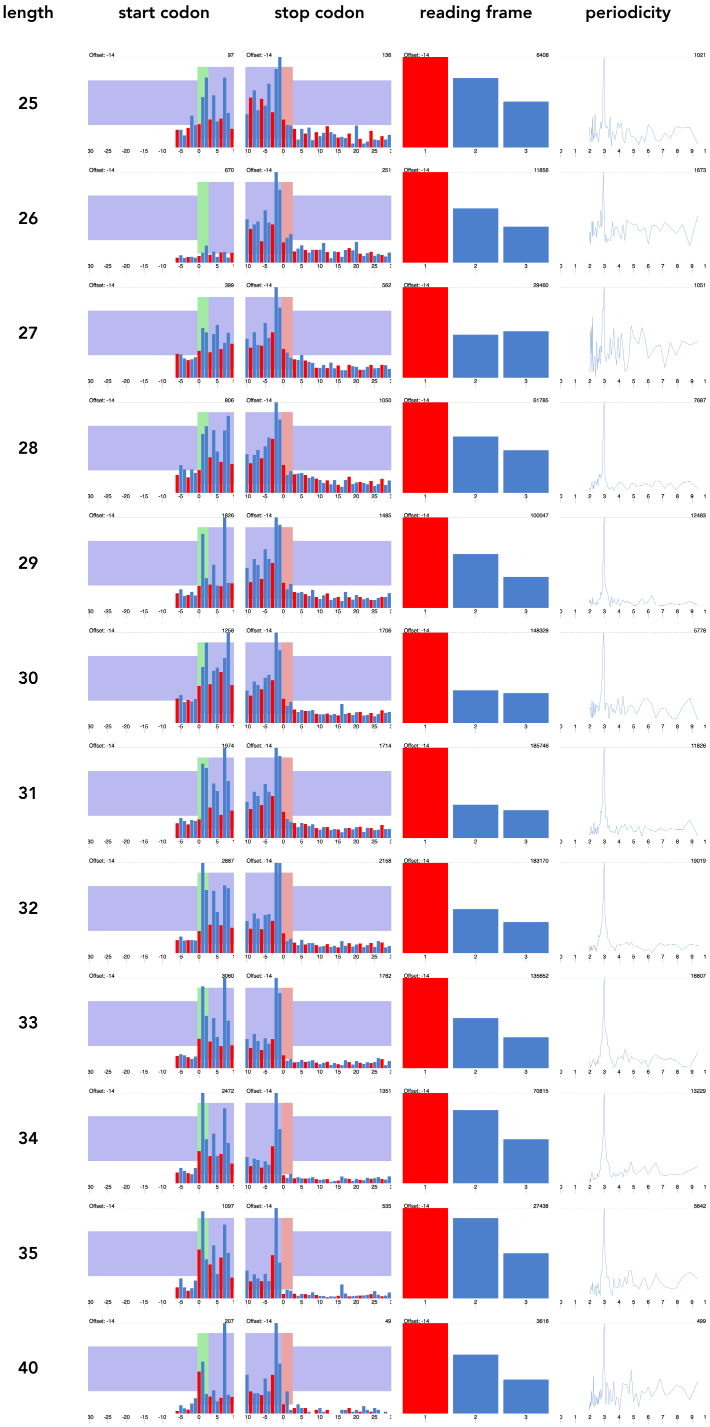
References

1. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324(5924):218–23.
2. Wang H, Wang Y, Xie Z. Computational resources for ribosome profiling: from database to web server and software. Brief Bioinform. 2017. https://doi.org/10.1093/bib/bbx093.
3. Dunn JG, Weissman JS. Plastid: nucleotide-resolution analysis of next-generation sequencing and genomics data. BMC Genomics. 2016;17(1):958.
4. Popa A, Lebrigand K, Paquet A, Nottet N, Robbe-Sermesant K, Waldmann R, Barbry P. Riboprofiling: a bioconductor package for standard ribo-seq pipeline processing. F1000Res. 2016;5:1309.
5. Malone B, Atanassov I, Aeschimann F, Li X, Grosshans H, Dieterich C. Bayesian prediction of rna translation from ribosome profiling. Nucleic Acids Res. 2017;45(6):2960–72.
6. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012;22(11):2219–29.
7. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, Vejnar CE, Lee MT, Rajewsky N, Walther TC, Giraldez AJ. Identification of small orfs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014;33(9):981–93.
8. Woolstenhulme CJ, Guydosh NR, Green R, Buskirk AR. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking efp. Cell Rep. 2015;11(1):13–21.
9. Giess A, Jonckheere V, Ndah E, Chyzynska K, Van Damme P, Valen E. Ribosome signatures aid bacterial translation initiation site identification. BMC Biol. 2017;15(1):76.
10. Lareau LF, Hite DH, Hogan GJ, Brown PO. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mrna fragments. Elife. 2014;3:01257. NLM: Original DateCompleted: 20140612.
11. Andreev DE, O'Connor PBF, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, Cormican P, Morris DW, Shatsky IN, Baranov PV. Translation of 5′ leaders is pervasive in genes resistant to eif2 repression. Elife. 2015;4:03971.
12. Gonzalez C, Sims JS, Hornstein N, Mela A, Garcia F, Lei L, Gass DA, Amendolara B, Bruce JN, Canoll P, Sims PA. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. J Neurosci. 2014;34(33):10924–36.
13. Guo H, Ingolia NT, Weissman JS, Bartel DP. Mammalian micrornas predominantly act to decrease target mrna levels. Nature. 2010;466(7308):835–40.
14. Hsieh AC, Liu Y, Edlind MP, Ingolia NT, Janes MR, Sher A, Shi EY, Stumpf CR, Christensen C, Bonham MJ, Wang S, Ren P, Martin M, Jessen K, Feldman ME, Weissman JS, Shokat KM, Rommel C, Ruggero D. The translational landscape of mtor signalling steers cancer initiation and metastasis. Nature. 2012;485(7396):55–61.
15. Ingolia NT, Brar GA, Rouskin S, McGeachy AM, Weissman JS. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mrna fragments. Nat Protoc. 2012;7(8):1534–50.
16. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, Wills MR, Weissman JS. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 2014;8(5):1365–79.
17. Lee S, Liu B, Lee S, Huang S-X, Shen B, Qian S-B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2012;109(37):2424–32.

18. Liu B, Han Y, Qian S-B. Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. Mol Cell. 2013;49(3):453–63.
19. Sidrauski C, McGeachy AM, Ingolia NT, Walter P. The small molecule isrib reverses the effects of eif2alpha phosphorylation on translation and stress granule assembly. Elife. 2015;4:1–16.
20. Stern-Ginossar N, Weisburd B, Michalski A, Le VTK, Hein MY, Huang S-X, Ma M, Shen B, Qian S-B, Hengel H, Mann M, Ingolia NT, Weissman JS. Decoding human cytomegalovirus. Science. 2012;338(6110):1088–93.
21. Stumpf CR, Moreno MV, Olshen AB, Taylor BS, Ruggero D. The translational landscape of the mammalian cell cycle. Mol Cell. 2013;52(4): 574–82.
22. Subtelny AO, Eichhorn SW, Chen GR, Sive H, Bartel DP. Poly(a)-tail profiling reveals an embryonic switch in translational control. Nature. 2014;508(7494):66–71.
23. Dana A, Tuller T. The effect of trna levels on decoding times of mrna codons. Nucleic Acids Res. 2014;42(14):9171–81.
24. Dana A, Tuller T. Mean of the typical decoding rates: A new translation efficiency index based on the analysis of ribosome profiling data. G3. 2015;5(1):73–80.
25. Li G-W, Oh E, Weissman JS. The anti-shine-dalgarno sequence drives translational pausing and codon choice in bacteria. Nature. 2012;484(7395):538–41.

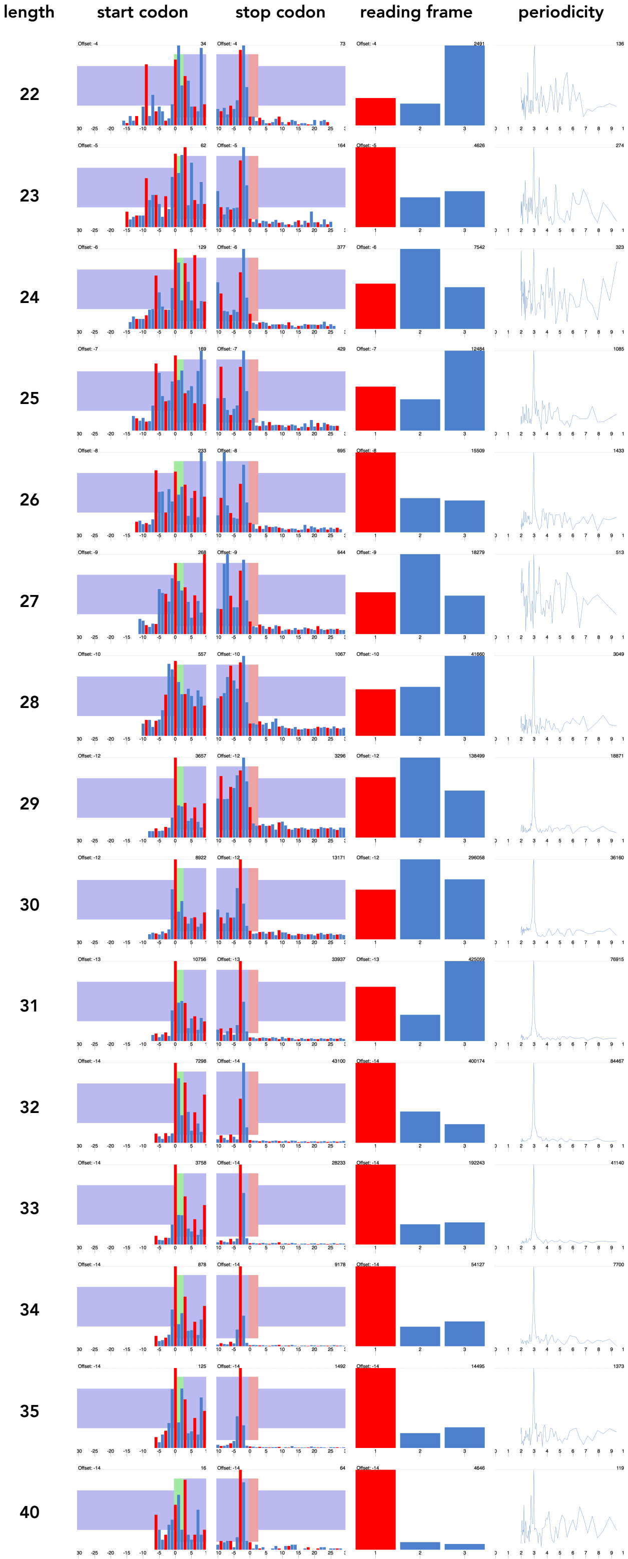


Additional figure 1
Analysis of human ribosome profiling, SRR493747 (Ingolia *et al.*, 2012). The libraries were treated with harringtonine, which causes accumulation of ribosomes at translational initiation sites and cycloheximide, which inhibits translation. The fragments of lengths 20-31 exhibit periodicity. The peaks are clear on start and before stop codons and the offsets change in increments of one nucleotide with increasing footprint length.



Additional figure 2

Analysis of human ribosome profiling, SRR1039861 (Subtelny *et al.*, 2014). The libraries were treated with cycloheximide, which inhibits translation. The footprints of lengths 25-35 and 40 are periodic (there are no footprints of lengths 36-39 in the library). The peaks are somewhat ambiguous on start, but clear before stop codons and all lengths map preferentially to the first reading frame. The offset from 5′ end of footprint is uniform for all lengths.



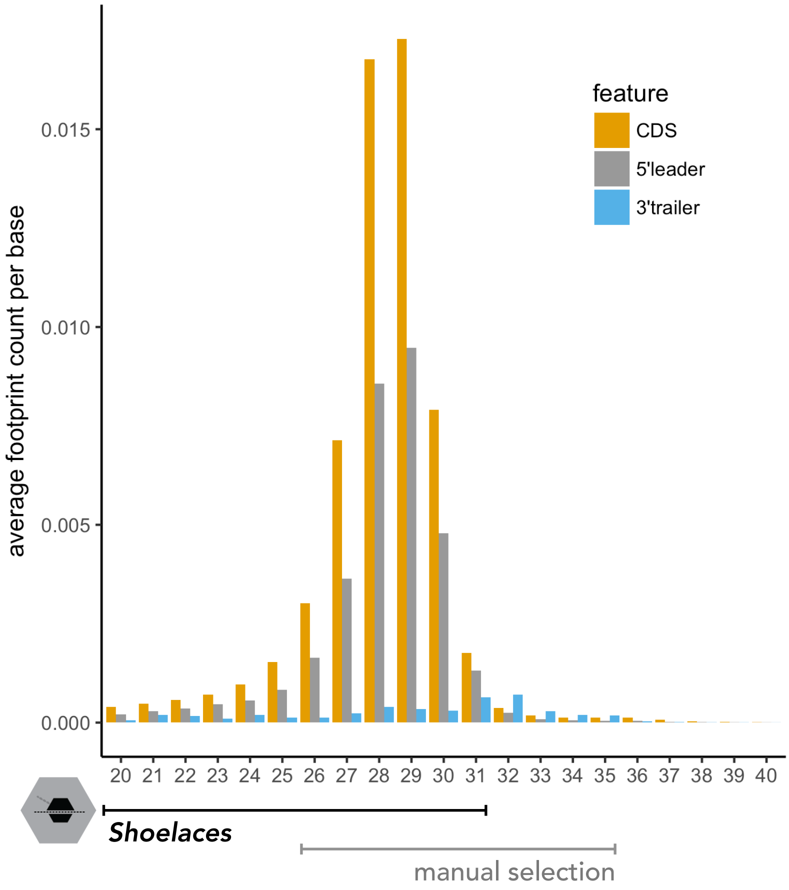
Additional figure 3

Analysis of human ribosome profiling, SRR592961 (Stern-Ginossar *et al.*, 2012). There were no translational inhibitors used in the library preparation. The footprints of lengths 22-35 and 40 are periodic (there are no footprints of lengths 36-39 in the library). The metagene profiles over start codon are unclear for shorter fragment lengths, but clear for the last codon of the coding region. The offsets change in increments of one nucleotide for shorter fragment lengths, while are more uniform for longer footprints. Longer footprints also map preferentially to the first reading frame.

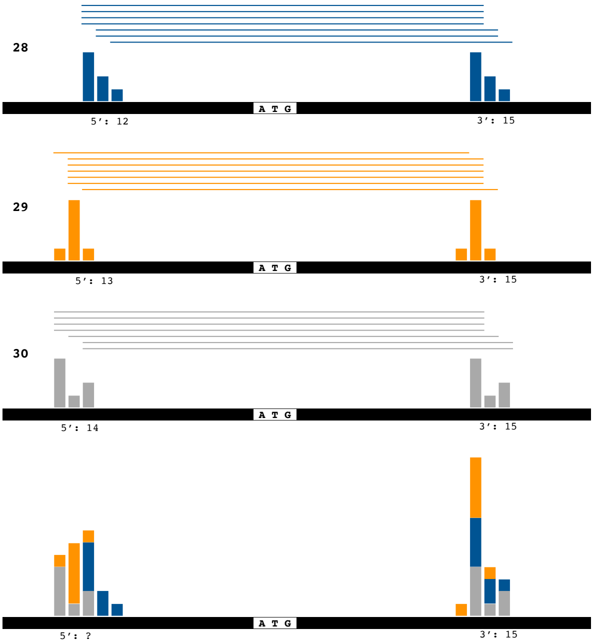
Additional table 1

Comparison of selected footprint lengths as in original study (blue) and based on periodicity in Shoelaces (**S**). Shoelaces captures up to 32% more reads mapping to CDSs and 5′leaders compared to arbitrary manual selection.

Study	Study ID	Run ID	Footprint length																					% gain in reads over CDSs and 5’leaders
			20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	
Andreev *et al.*, 2015	SRP038695	SRR1173905	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	n/a	n/a	n/a	**0.1**
		SRR1173907	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	n/a	n/a	n/a	**0.1**
		SRR1173909								S	S	S	S	S	S	S	S	S	S			n/a	n/a	**9.6**
		SRR1173910								S	S	S	S	S	S	S	S	S	S		n/a	n/a	n/a	**9.1**
		SRR1173913								S	S	S	S	S	S	S	S				n/a	n/a	n/a	**15.9**
		SRR1173914	n/a	n/a						S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	n/a	n/a	**15.5**
Gonzalez *et al.*, 2014	SRP031501	SRR1562539	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S					n/a	n/a	**13.4**
		SRR1562540	S	S	S	S	S	S	S	S	S	S	S	S	S					n/a	n/a	n/a	n/a	**4.1**
		SRR1562541	S	S	S	S	S	S	S	S	S	S	S	S	S				n/a	n/a	n/a	n/a	n/a	**5.4**
		SRR1562542	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	n/a	n/a	n/a	n/a	n/a	**6.5**
		SRR1562543		S	S	S	S	S	S	S	S	S	S	S	S			n/a	n/a	n/a	n/a	n/a	n/a	**4.5**
Guo *et al.*, 2010	SRP002605	SRR057511						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**19.5**
		SRR057512						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**19.9**
		SRR057516						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**19.2**
		SRR057517						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**19.2**
		SRR057521						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**16.7**
		SRR057522						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**16.7**
		SRR057526						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**22.1**
		SRR057529						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**23.0**
		SRR057532						S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**23.0**
		SRR065774	n/a	n/a	n/a			S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**8.2**
		SRR065775	n/a					S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**8.0**
		SRR065779	n/a	n/a	n/a			S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**10.1**
		SRR065780	n/a					S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	n/a	n/a	n/a	n/a	**10.0**
Hsieh *et al.*, 2012	SRP010679	SRR403883	S	S	S	S	S	S	S	S	S	S	S	S					n/a	n/a	n/a	n/a		**1.9**
		SRR403885	S	S	S	S	S	S	S	S	S	S	S	S					n/a	n/a	n/a	n/a	n/a	**3.1**
		SRR403887	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	**5.1**
		SRR403889	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a		**1.1**
		SRR403891	S	S	S	S	S	S	S	S	S	S	S	S	S				n/a	n/a	n/a	n/a		**1.6**
		SRR403893			S	S	S	S	S	S	S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	**0.9**
Ingolia *et al.*, 2012	SRP012648	SRR493747	S	S	S	S	S	S	S	S	S	S	S	S								n/a	n/a	**7.2**
		SRR493748	S	S	S	S	S	S	S	S	S	S	S	S	S							n/a	n/a	**6.5**
		SRR493749	S	S	S	S	S	S	S	S	S	S	S	S	S							n/a	n/a	**8.1**
Ingolia *et al.*, 2014	SRP045257	SRR1536302		S	S	S	S	S	S	S	S	S	S			S	S	S	S	S	n/a	n/a	n/a	**30.9**
		SRR1536303		S	S	S	S	S	S	S	S	S	S			S	S	S		n/a	n/a	n/a	n/a	**25.9**
		SRR1536304	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			n/a	n/a	n/a	**7.3**
		SRR1536305	S	S		S	S	S	S	S	S	S	S	S	S	S			n/a	n/a	n/a	n/a	n/a	**6.6**
Lee *et al.*, 2012	SRP014629	SRR618770			S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	n/a	n/a	n/a	n/a	**32.4**
		SRR618771	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	n/a	n/a	n/a	**28.6**
		SRR618772	n/a	n/a	n/a	S	S	S	S	S	S	S	S	S				n/a	n/a	n/a	n/a	n/a	n/a	**28.0**
		SRR618773	n/a	S	S	S	S	S	S	S	S	S	S	S				n/a	n/a	n/a	n/a	n/a	n/a	**24.7**
		SRR964946	n/a	n/a			S	S	S	S	S	S	S	S				n/a	n/a	n/a	n/a	n/a	n/a	**17.3**
Liu *et al.*, 2013	SRP017263	SRR619082	S	S	S	S	S	S	S	S	S	S	S	S	S			n/a	n/a	n/a	n/a	n/a	n/a	**5.6**
		SRR619083		S	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	n/a	n/a	n/a	**2.0**
		SRR619084	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	n/a	n/a	**5.6**
		SRR619085	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	n/a	**1.1**
		SRR619086	S	S	S	S	S	S	S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	n/a	n/a	n/a	**4.1**
		SRR619087				S	S	S	S	S	S	S	S	S	S		n/a	n/a	n/a	n/a	n/a	n/a	n/a	**3.0**
		SRR619088	n/a	n/a		S	S	S	S	S	S	S	S	S	S	S		S	S		S		n/a	**1.0**
		SRR619089				S	S	S	S	S	S	S	S	S	S	S		S	S		S			**0.9**
		SRR619090					S	S	S	S	S	S	S	S	S	S		S	S	S	S	S		**0.4**
		SRR619091					S	S	S	S	S	S	S	S	S	S		S	S	S	S	S		**0.4**
		SRR619092	S			S	S	S	S	S	S	S	S	S	S				S			S		**-2.8**
		SRR619093				S	S	S	S	S	S	S	S	S	S			S	S	S	S	S		**-1.4**
		SRR619094	S	S		S	S	S	S	S	S	S	S	S	S	S	S	S	S					**2.2**
		SRR619095				S	S	S	S	S	S	S	S	S	S	S								**0.8**
Sidrauski *et al.*, 2015	SRP053402	SRR1795425	n/a	n/a	n/a	S	S	S	S	S	S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	n/a	**11.9**
		SRR1795426			S	S	S	S	S	S	S	S	S	S	S	S	S	S					n/a	**11.1**
		SRR1795427			S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			n/a	**4.5**
		SRR1795428			S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	**4.5**
		SRR1795429			S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	**4.3**
		SRR1795430			S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	**4.3**
		SRR1795431			S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			n/a	**3.1**
		SRR1795432			S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			n/a	**3.2**
		SRR1795433				S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			n/a	**3.8**
		SRR1795434				S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			n/a	**4.0**
		SRR1795435				S	S	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	**5.0**
		SRR1795436				S	S	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	n/a	**5.0**
		SRR1795437	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			**19.8**
		SRR1795438	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			**17.9**
		SRR1795439	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			**10.7**
		SRR1795440	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			**10.0**
Stumpf *et al.*, 2013	SRP029589	SRR970490	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			**0.8**
		SRR970538	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S		n/a	**1.9**
		SRR970561	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			**5.4**
		SRR970565	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S			**0.4**
		SRR970587	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S				n/a	n/a	**4.0**
		SRR970588	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S						**0.1**
Subtelny *et al.*, 2014	SRP033369	SRR1039861	n/a					S	S	S	S	S	S	S	S	S	S	S	n/a	n/a	n/a	n/a	S	**18.8**



Additional figure 4
Average count of 5′ends of footprints per base in CDSs, 5′leaders and 3′trailers, stratified by fragment length for SRR493747 (Ingolia *et al.*, 2012). The fragments of lengths 20-31 exhibit periodicity (see Additional figure 1), and were therefore selected by Shoelaces, while in the original selection the reads of lengths 26-35 were kept. The selection performed by Shoelaces increases the count of translating footprints (CDSs and 5′leaders) and decreases the count of long footprints mapping to 3′trailers, which could be stemming from mRNA binding proteins, secondary structure or other sources of noise.



Additional figure 5
Calibrating offsets for each fragment length separately bypasses the need to consider both 5′ and 3′ assignment strategies, as both types of assignments are mathematically identical. Here, for the three footprint lengths of 28 (blue), 29 (orange) and 30 nucleotides (grey), the 5′ offsets to the first base of start codon are 12, 13 and 14 nt long respectively, while the 3′ offset is uniform and has length 15 nt. Have all the footprint lengths been aggregated (bottom panel), the profile of 5′ ends of footprints would have been ambiguous (Woolstenhulme *et al.*, 2015).

Paper II

4.2 Deep conservation of ribosome stall sites across RNA processing genes

Katarzyna Chyżyńska, Carl Jones, Kornel Labun, Eivind Valen[†] and Sushma Grellscheid[†]

manuscript in preparation

[†] Contributed equally

Chapter 5

Discussion and perspectives

The work presented in this thesis aimed to discover and annotate genome-wide translational regulation by ribosome stalling using publicly available ribosome profiling data. The first paper demonstrated that a standardized, library-specific processing decreases the amount of bias in the data and prepares it for analyses at the nucleotide level. The second paper showed that ribosome stalling is a genome-wide, conserved mechanism. It determined that proline, glycine and negatively charged amino acids in a proper context are significant contributors to conserved stalling. Furthermore, it revealed that conserved stall sites tend to be present in RNA processing genes, suggesting regulation of RNA metabolism via stalling. Finally, it created a resource for further in-depth studies on conserved stalling. Here, I discuss the lessons learnt from the research.

Large-scale, standardized ribosome profiling data analyses to improve research reproducibility

While analyzing multiple ribosome profiling datasets, it quickly becomes obvious just how large the extent of local variability is. The footprints vary both in size (section 1.5.2) and location, with the latter being especially sensitive to use of translation inhibitors. Cycloheximide is the most common inhibitor used in the vast majority of ribosome profiling studies to date. It is used to halt translational elongation, effectively ′freezing′ ribosomes on mRNA. While this is very beneficial and aids the experimental protocol considerably, its use has turned out to be highly controversial. As sequencing data and resolution of analysis has improved, there are increasing numbers of studies reporting a bias in ribosome profiling data generated using CHX, potentially caused by slowing of the ribosome rather than instant freezing, thus resulting in peaks that might falsely suggest increased ribosome occupancy. Many studies have pointed out abnormalities in ribosome profiling experiments with the use of CHX at the initiation and elongation stages [11, 49, 60, 80] and very recently hugely distorted measurements of mRNA levels and translation efficiency in such data [146]. All of them highlight the need for better standardization of the data and reexamination of conclusions from previous analyses.

The preparation and sequencing of new datasets are expensive, and a wealth of data is already deposited in databases. Large-scale bioinformatic analyses that account for biases in data are therefore preferred to producing new, costly datasets. Proper standardized processing, accounting for the known systematic artefacts and including a large amount of data from various sources and treatments are ways to increase confidence in the findings. The first paper addressed the first part of this problem. Where the datasets were typically manually processed on an *ad hoc* basis, our method *Shoelaces* introduces standardizing the selection and processing of footprints to improve the quality and resolution of data at the nucleotide level. Secondly, systematic CHX-induced biases, like a ramp at the beginning of coding sequences can be easily excluded from analysis. Finally, the second paper is an application of the principles from the first study to a specific problem in translation: whether translational stalling is widespread and a potential regulatory mechanism for gene expression. We demonstrated how increasing the amount of data both from humans as well as inclusion of data from other organisms can help elucidate a regulatory mechanism. With the knowledge we have now, I strongly support the claim that many of the studies relying on ribosome profiling data should be revisited.

Footprint lengths encode valuable information

One of the ways that bias shows itself in ribosome profiling data is footprint length distribution. The experimental protocol in the vast majority of cases includes size-selection, with the mean footprint size oscillating around 28-30 nucleotides, typical of the non-rotated, elongating ribosomes. Afterwards, the footprints are further subset for bioinformatic analyses. In such two-step selection, a lot of information is lost. This has been demonstrated to lead to erroneous conclusions in studies on stalling. In one bacterial study, preferential selection for longer footprints largely overestimated the strength of SD pauses, and when corrected to include the whole range of fragment lengths, found that SD-associated pausing was no longer observable [120]. In other studies, it has been shown that lengths one might expect in stalling include longer footprints of around 80 nucleotides, as a result of closely stacked di-ribosomes [71], or shorter ones of 20-22 nucleotides representing the rotated ribosome conformation [102]. In *Shoelaces*, we addressed this by keeping maximum range of footprints in each library, as long as they represent genuine translating ribosomes. In the second paper, we demonstrated that keeping the different lengths can detect the shift in fragment length distribution at terminating ribosomes on a metalevel. With higher content of non-standard footprints, it might be possible to use it as a feature in detecting termination at the gene level. Finally, as demonstrated in the third paper (included in the Appendix), the read length distribution can be used for re-annotation and discovery of novel translation initiation sites.

Precise determination of codon under translation for precise analyses

Precise assignment of reads to positions being actively translated is necessary for many codon-level analyses (section 1.5.2). The often manually implemented and arbitrary selection of read offsets can lead to vastly variable results in such analyses. In *Shoelaces*, we implemented a strategy for precise determination of codon under translation, and applied it to the determination of conserved stall sites in the second paper. Comparison of the stalling results to previous studies, however, needs to take into account the sensitivity of offset selection. For instance, we found glycine to be a stalling determinant at the P-site. However, a previous bacterial study found it at the E-site [120]. This discrepancy can be attributed to the offset selection strategy. The bacterial study used 3′end assignment strategy, not taking into account differing footprint lengths. Even though not a large difference in this case, an understanding of possible mechanisms causing stalling depends on knowing the correct position where the stalling occurs.

Regulatory mechanisms are likely to be deeply conserved

The biases do not pose the only obstacle to reproducible ribosome profiling studies. The datasets produced are - rightly - tailored to the specific research question. However, they too can be used for global analyses in two ways. Firstly, nearly every study includes wild type, unstressed control libraries. These can be analyzed in search for global patterns in translational regulation. Secondly, as translation is highly conserved among species, so will be the systems regulating it. The second paper demonstrates that stall sites are deeply conserved across phylogenetically diverse organisms. Such conservation is likely to exist for other regulatory systems as well. Therefore, multi-species comparison of conservation in regulatory patterns is a good starting point for discovery of valuable information.

Determinants of stalling, revisited

Translation stalling has several previously reported causes (section 1.4.1). Proline is a well-known cause of stalling, whether alone or in doublets [9, 132]. It is both a poor donor (peptidyl-tRNA) and acceptor (aminoacyl-tRNA) of peptide bond formation [112], slowing down translation, and if present in stretches, makes the nascent chain incompatible with the ribosome exit tunnel and destabilizes peptidyl-tRNA to the same effect. We did indeed find proline as a contributor to around 15% of conserved stalling cases. Additionally, we found a contribution from glycine (12%), which was previously reported in bacteria [120], but never in eukaryotes. The study attributed it to the action of chloroamphenicol, a type of antibiotic used for translation inhibition. As none of the data analyzed in the second paper were treated with it, it is unlikely to be the case.

The charge of amino acids and its role in stalling has been highly disputed. We found positively charged amino acids present in a bit over 2% of conserved stall sites, confirming the effect is there, but is not as major as previously reported [30], not only attributable to biases as disputed in another study [9], not only evident in the absence of inhibitors as claimed elsewhere [138] and is caused by lysine stretches known to cause stalling [8, 100]. Most interestingly however, the largest proportion of conserved stall sites can be explained by negatively charged amino acids - 17% by aspartate at P-site, 10% by glutamate at A-site (additional 7% has Pro/Gly/Asp at P-site) and both aspartate and glutamate at position -2 (additional 7%). The mechanism of these is not clear. Those in the P- and A-site might slow down peptide bond formation, while the others might interact with the exit tunnel. Possibly, upon encountering the narrow, negatively charged entrance to the tunnel, the charges of amino acids and the tunnel might act as repellants, and thus slow down translation. Most importantly, the amino acids to not stall translation alone, but their context seems to be detrimental, possibly in the same way as Kozak sequence is detrimental for translation initiation[3]. Kozak sequence flanks the AUG start codon and helps tether the ribosome to the initiation site by interaction with 18S rRNA. In a similar way, mRNA sequence flanking the stall sites could interact with 40S subunit components, causing the ribosome to pause.

Other hypothesized causes of stalling include mRNA secondary structure and tRNA abundance. mRNA structure blocking translation, although theoretically sound, has never been proven to cause stalling on a global level. A recent chloroplast study reported detection of structure downstream of a small sample of stall sites [58]. However in our second paper we found no evidence of strong structures downstream of conserved stall sites. Additionally, given the small sample size the results in that study were likely accidental. Given the limitations of structure prediction methods and a largely different folding of *in vivo* transcripts [143] due to the presence of RNA binding proteins, a large-scale analysis of data from *in vivo* structure probing experiments might be more suited to determine to what extent - if any - structure may lead to ribosome stalling. Similarly, experimental studies measuring tRNA abundance would be needed in addition to ribosome profiling data, in order to determine possible connections between tRNA availability, codon abundance and ribosome stalling.

Known and novel consequences of stalling

Known consequences of stalling include membrane targeting, co-translational protein folding and, if due to aberrant translation, degradation of nascent peptide and mRNA template. By analysing the predicted protein secondary structure around conserved stall sites, we found evidence for possible co-translational folding as coiled coil domains were more prevalent around conserved stall sites, suggesting that pausing or stalling of the ribosome might give more time for correct protein folding to occur.

While we did not find enough direct support for membrane targeting being a significant cause of conserved stalling, we did observe an enrichment by gene ontology analysis for mRNAs containing stall sites coding for transmembrane proteins. Most likely, membrane targeting it is not a major consequence of conserved stalling.

Another interesting gene ontology term that was strikingly enriched was RNA metabolism, specifically RNA binding proteins involved in alternative splicing and translation regulation. This might imply a possible self-regulation mechanism, where stalling both regulates protein synthesis and is regulated by the synthesized proteins. How this links to alternative splicing that occurs mostly in the nucleus is not intuitive. In recent years there has been a growing interest in the role of whether alternative splicing factors play independent roles in the cytoplasm, usually in translation control [22]. In addition to this, there is indeed some splicing activity occurring in the cytoplasm, as in the case of *XBP1* retained intron. In this context it is also worth mentioning that increasing numbers of studies are reporting retained introns that are transported to the cytoplasm [95]. The fate of these (m)RNAs is not known- are they spliced in the cytoplasm like *XBP1*? Are they degraded by a quality control mechanism before translation, or simply translated to include an additional domain? These are open questions in the alternative splicing field and given that cytoplasmic splicing is highly controversial, it was beyond the scope of this study to delve into these in any detail. On inspection of the known *XBP1* CSS which was also predicted in our study and the location of the known retained intron in *XBP1* that is cytoplasmically spliced, it is not clear how the two processes could be related because the CSS is downstream of the retained intron, if we assume as would be expected, that splicing occurs before translation. We briefly investigated whether RNA binding proteins that we found to contain stall sites are also predicted to bind those mRNAs in an auto-regulatory mechanism. This investigation was very preliminary and yet inconclusive, but remains an interesting hypothesis worthy of further investigation.

In terms of whether stalling is part of a quality control mechanism in translation regulation, we found that a few of the conserved stall sites might lead to aberrant translation triggering nonsense-mediated decay or no-go decay, but the effect was not widespread.

The mechanistic basis for the very strong overrepresentation of mRNAs for RNA binding proteins in our stall site predictions may be a subject worth investigating in detailed molecular biology studies in the future. In addition to the known processes possibly regulated by stalling, we found conserved stall sites overrepresented in genes with a wide-range of functions and processes involved in RNA metabolism. Yet to explain it further, future investigations, extending beyond the scope of ribosome profiling are needed. A catalogue of genes with stall sites and their functions that we created is a good starting point for advanced research.

Appendix A

Appendix

A.1 Ribosome signatures aid bacterial translation initiation site identification

Adam Giess, Veronique Jonckheere, Elvis Ndah, **Katarzyna Chyżyńska**, Petra Van Damme and Eivind Valen

BMC Biology, 15:76 (2017)

Giess *et al. BMC Biology* (2017) 15:76
DOI 10.1186/s12915-017-0416-0

BMC Biology

RESEARCH ARTICLE Open Access

Ribosome signatures aid bacterial translation initiation site identification

Adam Giess[1], Veronique Jonckheere[2,3], Elvis Ndah[2,3,4], Katarzyna Chyżyńska[1], Petra Van Damme[2,3*†] and Eivind Valen[1,5*†]

Abstract

Background: While methods for annotation of genes are increasingly reliable, the exact identification of translation initiation sites remains a challenging problem. Since the N-termini of proteins often contain regulatory and targeting information, developing a robust method for start site identification is crucial. Ribosome profiling reads show distinct patterns of read length distributions around translation initiation sites. These patterns are typically lost in standard ribosome profiling analysis pipelines, when reads from footprints are adjusted to determine the specific codon being translated.

Results: Utilising these signatures in combination with nucleotide sequence information, we build a model capable of predicting translation initiation sites and demonstrate its high accuracy using N-terminal proteomics. Applying this to prokaryotic translatomes, we re-annotate translation initiation sites and provide evidence of N-terminal truncations and extensions of previously annotated coding sequences. These re-annotations are supported by the presence of structural and sequence-based features next to N-terminal peptide evidence. Finally, our model identifies 61 novel genes previously undiscovered in the *Salmonella enterica* genome.

Conclusions: Signatures within ribosome profiling read length distributions can be used in combination with nucleotide sequence information to provide accurate genome-wide identification of translation initiation sites.

Keywords: Ribosome profiling, Bacterial translation initiation, Machine learning, N-terminal proteomics, Proteogenomics

Background

Identification of translated open reading frames (ORFs) is a critical step towards gene annotation and the understanding of a genome. The rapid advances in sequencing have resulted in a deluge of new genomes, making manual annotation intractable and the development of accurate automated methods a necessity. In prokaryotes, ORF delineation is particularly challenging since genes are often tightly packed and frequently overlapping. Whole genome ORF identification in prokaryotes is most commonly performed in silico, using a variety of sequence features, such as guanine-cytosine (GC) codon bias, and motifs, such as the ribosomal binding site or Shine–Dalgarno (SD) sequence [1–3], to differentiate those ORFs that are thought to be functional from those that occur in the genome by chance. While these techniques are able to identify genomic regions containing ORFs with a high accuracy [3], predicting translation initiation sites (TISs), and thus the exact beginning of a protein coding sequence (CDS), is substantially more challenging. In addition to providing functional information via the peptide sequence, regulatory and targeting information are often contained within protein N-termini [4, 5], making accurate identification of the beginning of ORFs essential. This has led to the development of a number of in silico-based TIS identification methods relying on a variety of sequence features [6–9], typically applied after initial ORF annotation in order to re-annotate the often erroneously predicted TIS.

* Correspondence: petra.vandamme@vib-ugent.be; eivind.valen@gmail.com
†Equal contributors
[2]VIB-UGent Center for Medical Biotechnology, B-9000 Ghent, Belgium
[1]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen 5020, Norway
Full list of author information is available at the end of the article

High throughput proteogenomics has the potential to enable identification of protein N-termini, and by extension TISs, from an entire proteome. In practice, however, variation in protein expression levels, physical properties, MS-incompatibility and the occurrence of protein modifications limit the number of detectable protein N-termini [10, 11]. In prokaryotes, N-terminal proteomics typically captures the corresponding peptides of hundreds to the low thousands of genes [11]. For example, a recent study identified N-terminal peptides of 910 of the 4140 (22%) annotated genes in *Escherichia coli* [12]. Although falling short of providing full genome annotation, such datasets provide an effective means of experimental TIS validation.

Significantly higher coverage of TISs can be achieved with sequencing-based technologies. By specifically focusing on ribosome protected reads, ribosome profiling (ribo-seq) [13] infers which parts of the transcriptome are actively undergoing translation. Briefly, ribo-seq aims to capture, select and sequence mRNA reads that are associated with ribosomes, typically reads of 26–34 nt (eukaryotic) [14, 15] or 20–40 nt (prokaryotic) [16, 17] in length. These reads are then commonly assigned to a fixed offset [15, 17–20], or a read length-dependent offset [14, 21, 22], in order to resolve the translated codon represented by each read. In this way, ribo-seq has been used to demonstrate translation of many RNAs and regions that were not thought to be associated with ribosomes [14, 18, 21, 23–26]. Being able to identify translation on a transcriptome-wide scale has obvious application to ORF annotation and a number of methodologies have been developed for prediction of translated ORFs [15, 18, 21, 22, 27]. These methods rely on a number of features, like codon periodicity, read context and read lengths, to distinguish footprints indicative of translation from other, non-translating footprints frequently observed in ribo-seq data. While progress has been made on finding translated regions, delineating their exact boundaries has received less attention. Antibiotic treatment can be used to stall and capture footprints from the initiating ribosome [14, 28, 29], but finding a suitable compound has been elusive in prokaryotes, with only one dataset available to date [30].

Here, we present a generally applicable method that does not depend on specialised chemical treatment, but can take advantage of such data (Fig. 1a). Using N-terminal proteomics we demonstrate its high accuracy and show that it is consistent with other features linked to translation initiation. Applying the model, we predict numerous novel initiation sites in *Salmonella enterica* serovar Typhimurium and several novel genes.

Results

Ribosomal signatures of translation initiation

To investigate whether ribo-seq could aid in the accurate delineation of translated ORFs, we generated two ribo-

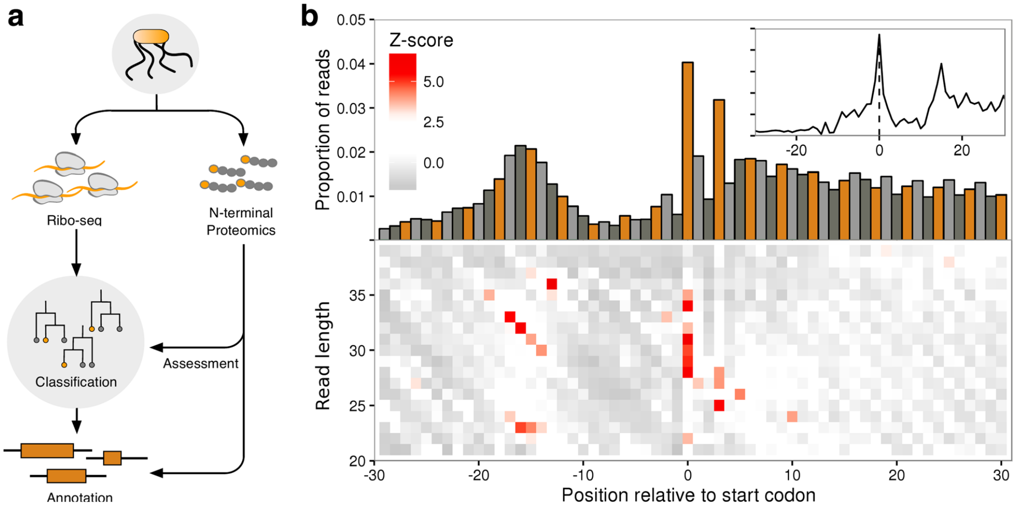


Fig. 1 Translation initiation site classification with ribo-seq read length patterns. **a** Schematic representation of the classification strategy. **b** Ribo-seq meta profiles in windows around start codons for all annotated coding sequences in the *S.* Typhimurium genome (monosome sample, n = 4187), contributions from each gene are scaled to a sum of one; (**upper panel**) proportion of 5′ ribo-seq read counts per nucleotide position, coloured by codon position; (**inset**) proportions of ribo-seq read counts per nucleotide position, after adjusting by read length offsets (see methods); (**lower panel**) heatmaps of z-scores of 5′ ribo-seq read counts per read length

seq libraries from monosome- and polysome-enriched fractions originating from *S*. Typhimurium. The similarities in the profiles of the two libraries (Additional file 1: Figure S1a–d), taken with current literature reports of similarities in the translational properties of polysome and monosome fractions [31], suggest that it is reasonable to consider these libraries sufficiently similar to serve as replicates for the purpose of initiation site prediction. The libraries were initially processed using a standard ribo-seq work-flow [14, 21, 22], trimmed footprints were aligned to a reference genome, and then adjusted based on 5' read profiles to determine the specific codon under active translation (Fig. 1b, inset). When exploring the processed reads, we discovered that, consistent with previous reports [20, 30], annotated start sites of prokaryotic ribosomes carry a specific signature around the initiating codon (Fig. 1b, inset). Examining the unprocessed reads, we observed that the pattern is a consequence of a specific distribution of read lengths (Fig. 1b), information which is typically lost in pipelines that pre-process the read signal by adjusting reads (Fig. 1b, inset). More specifically, heatmaps of 5' read profiles indicate that the pattern consists of an enrichment of longer reads (30–35 nucleotides (nt)) starting 14–19 nt upstream of the initiation codon (a diagonal pattern), but ending at the same location, 15 nt downstream of the initiation codon. A shorter set of reads (23–24 nt) are enriched in the same region, but have different end points, 7–9 nt downstream of the initiation codon. Finally, a strong enrichment of 5' ends of reads of length 28–35 nt can be observed exactly over the start codon itself (Fig. 1b). Looking at the compositions of these patterns, we observed a strong contribution from SD motifs (Additional file 1: Figure S2a, b), apparent as longer reads (30–35 nt) with a fixed 3' end and a 5' end dependent on the length of the SD motif from the TIS (Additional file 1: Figure S2c), as reported by O'Connor et al. [16] and Mohammad et al. [17]. Additionally, we discovered a smaller enriched read subset (24–26 nt) for which both 3' and 5' ends are dependent on the distance between the SD and the TIS (Additional file 1: Figure S2c). The SD sequence also impacts the distribution of reads immediately downstream, which show a depletion leading up to the TIS (Fig. 1b and Additional file 1: Figure S2 bar charts). Finally, we detected codon-specific enrichments of reads beginning at the first nucleotide of ATG and TTG codons (Additional file 1: Figure S3a, b) and, to a lesser extent, ending at the first codon position in GTG codons (Additional file 1: Figure S3c). These codon-specific enrichments could plausibly originate from experimental artefacts such as sequence-specific ligation or, perhaps more likely, from the sequence-specific digestion biases that have been reported to influence ribo-seq datasets [32, 33].

Ribosome profiling enables accurate annotation of TISs

We trained a random forest model on TISs from the top 50% translated ORFs (see methods) to recognise the patterns in 5' ribo-seq read lengths and sequence contexts in a −20 to +10 nt window around start codons. In addition, we encoded information about the start codon position within the ORF and the read abundance upstream and downstream of the start sites. The model was used to predict TISs from all in-frame cognate and near cognate start codons around annotated genes in the *S*. Typhimurium genome. Predictions on the two samples were highly accurate, with area under the curve (AUC) values of 0.9958 and 0.9956 on independent validation sets for the monosome and polysome samples, respectively (parameter importance for the models is summarised in Additional file 2: Tables S1, 2). In total, 4610 (monosome) and 4601 (polysome) TISs were predicted in the two sets. From these, we constructed a high confidence set from predictions common to both replicates. In total, this set contained 4272 predictions, representing an 86.50% agreement between the replicates. The discrepancies predominantly originate from genes with limited translation. Of the high confidence TISs, 3853 matched annotated ORFs, 214 represented extensions and 205 truncations. Representative examples of predicted extended, truncated and matching ORFs are shown in Fig. 2.

As expected, the predictions show the same start codon usage distribution (Additional file 1: Figure S4a) and carry the same read distribution signature as the annotated sites (Additional file 1: Figure S4b). Consistent with annotated initiation sites, an increase in ribosome protected reads can be seen downstream versus upstream of the predicted TIS (Fig. 3a). Furthermore, extended ORFs exhibit a shift in ribo-seq density downstream relative to the annotated TIS, consistent with the predicted extension. Conversely, truncated ORFs exhibit a shift in read density upstream relative to the annotated TIS and consistent with the predicted truncation.

To further assess the predictions we compared the newly predicted TISs with the previously, potentially erroneously, annotated TIS. A highly significant sequence feature of translation initiation sites is the SD sequence, which facilitates translation initiation in prokaryotes [34]. The consensus sequence GGAGG is located approximately 10 nt upstream of the start codon [35]. The predicted initiation sites show clear evidence of SD sequences centred 9–10 nt upstream of the start codon (Fig. 4a). Strikingly, the annotated TISs, in the same genes where our model has predicted novel sites, show an absence of the SD sequence (Fig. 4a). Since our model considers sequence context and SD-associated profiles, it is unsurprising that the predictions carry this

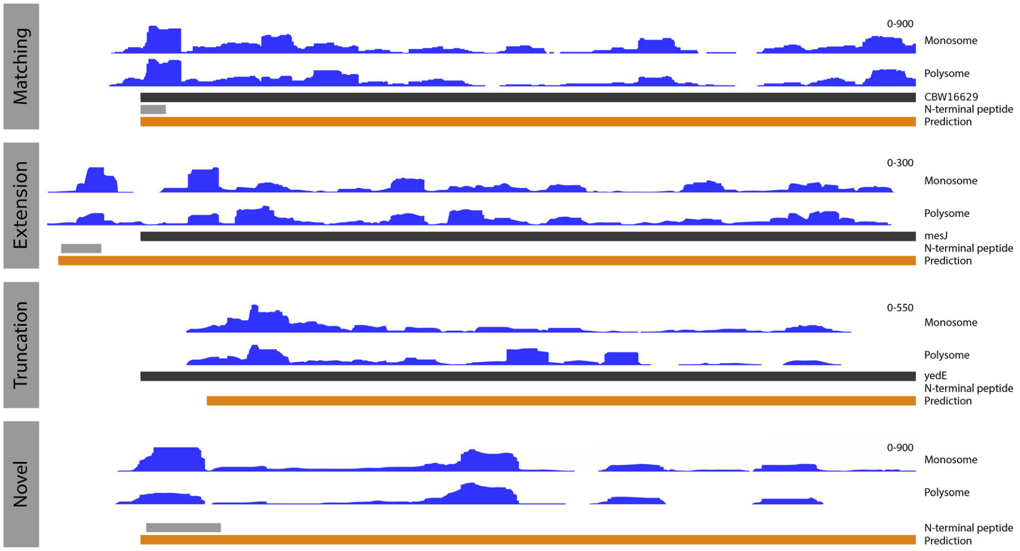


Fig. 2 Examples of predicted translated open reading frames (ORFs). Showing genomic tracks of unadjusted ribo-seq read coverage in blue (y axis scale on the right hand side), annotated genes in black, predicted ORFs in orange and N-terminal peptides in grey; (**upper**) a predicted ORF in agreement with the annotated ORFs, supported by ribo-seq coverage and N-terminal evidence; (**middle upper**) a predicted extension relative to the annotated ORF, with N-terminal evidence and ribo-seq coverage supporting the extended prediction; (**middle lower**) a predicted truncation relative to the annotated ORF, with support from ribo-seq coverage; (**lower**) a novel predicted ORF (Chromosome:2064124-2064558), supported by N-terminal evidence and ribo-seq coverage

signature, but the absence of these motifs around previously annotated start codons is notable.

Besides the presence of SD sequences, the GC content is commonly used to identify CDSs in prokaryotes. The overall GC content of a genome or genomic region is often highly optimised. In coding regions, this optimisation can be achieved via synonymous substitutions, predominantly at third codon positions [36], a finding explaining the pronounced bias in the GC content of third nucleotide positions observed in coding regions compared to the rest of the genome [37, 38]. Interestingly, at those annotated sites where our model predicts an extension, an increase in GC content upstream of the annotated start codon can clearly be observed consistent with the presence of an initiation site further upstream. Conversely, predicted truncations show a decrease downstream of the annotated start codon. In contrast, at predicted sites, both predicted extensions and truncations fit closely to the expected distribution (Fig. 4b upper). While our model has some potential to capture GC bias (at the nucleotides in positions –20 to +9 nt), the observable shifts in GC content, in relation to the annotated TIS where the model predicts an alternative TIS, argue against the validity of the annotated TIS.

Another significant feature of prokaryotic translation initiation is the absence of intrinsic structure in the region around the start codon, enabling easier access for ribosomes to bind [39]. We therefore calculated the average free energy over all predicted sites and compared them to the previous annotation in the same genes. Consistent with GC content patterns, ORFs where we predict an extension or truncation show lower free energy over the TIS at the previously annotated positions. For truncated ORFs, we also observe a higher propensity to form secondary structure downstream of the annotated start codon. In the predicted sites, these less-structured regions can clearly be observed directly over the start codon, highly indicative of true initiation sites (Fig. 4b lower).

Ribosomes translocate along mRNAs three nucleotides at a time, corresponding to one codon and an amino acid. Consequently, reads originating from bona fide translated regions also exhibit a three nucleotide periodicity in adjusted read counts, with a bias towards mapping to the first nucleotide in each codon [22]. At initiation sites, the read distribution therefore switches from a random distribution upstream to a periodic, biased distribution downstream, as demonstrated in Additional file 1: Figure S1a–d. While it has been argued that periodicity of ribosome profiling reads in prokaryotic genomes can be caused by the third codon GC bias

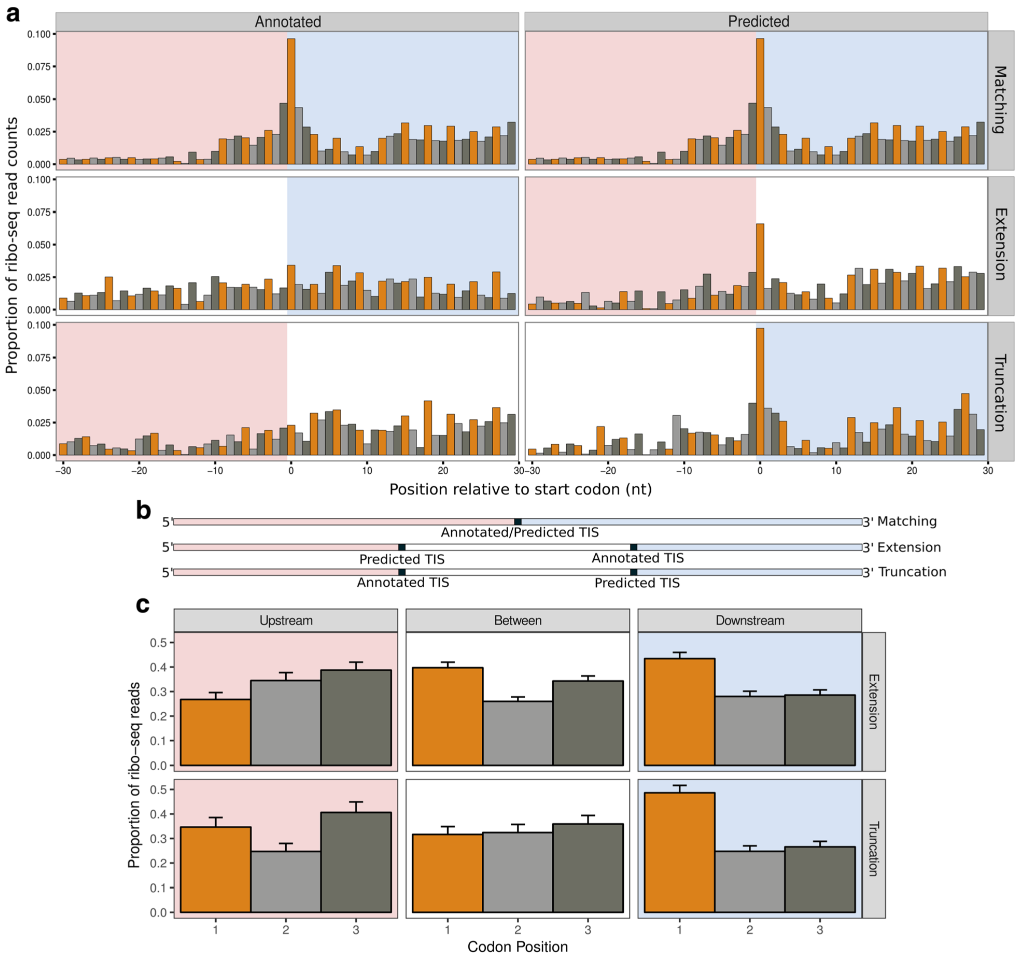


Fig. 3 Ribo-seq reads and periodicity are consistent with re-annotated translation initiation sites. Bar colour indicates codon position. Downstream regions are highlighted in pink, upstream regions are highlighted in light blue. **a** Meta-plots showing the proportion of scaled ribo-seq reads, after read length-specific adjustment, in relation to annotated or predicted translation initiation sites, for open reading frames matching annotated genes (n = 3853), predicted extensions (n = 214) or predicted truncations (n = 205). Contributions from each gene are scaled to a sum of one. Annotated translation initiation sites (TISs) show statistically significant increases in ribo-seq density upstream (extensions, Wilcoxon rank sum test W = 252,580, $P = 2.156 \times 10^{-6}$), or downstream of start codons (truncations, Wilcoxon rank sum test W = 293,200, $P = 1.139 \times 10^{-5}$). **b** Transcript models. **c** Bar plots with standard error of the mean, showing the proportion of scaled ribo-seq read counts, after read length-specific adjustment, in each codon position. For truncations, regions are 30 nt upstream of the annotated TIS, between the annotated and predicted TIS, and 30 nt downstream of the predicted TIS. For extensions, regions are 30 nt upstream of the predicted TIS, between the predicted and annotated TIS, and 30 nt downstream of the annotated TIS; 3 nt periodicity does not occur upstream of predicted TISs (truncations), but does occur upstream of annotated TISs (extensions)

(described above) [33], we observe that the periodicity is independent of third codon GC content (Additional file 1: Figure S5). Comparing the density of reads falling into each of the three codon positions, in extended ORFs, we observe increased read density at the first nucleotide position upstream of annotated, but not of predicted, TISs. Similarly, at truncated ORFs we see a decrease in the density of reads at the first nucleotide position downstream of the annotated TIS (Fig. 3c, “Between”), but not downstream of predicted TIS (Fig. 3c, “Downstream”).

Taken together, the patterns in read distribution, SD motifs, GC bias, unstructured regions and three

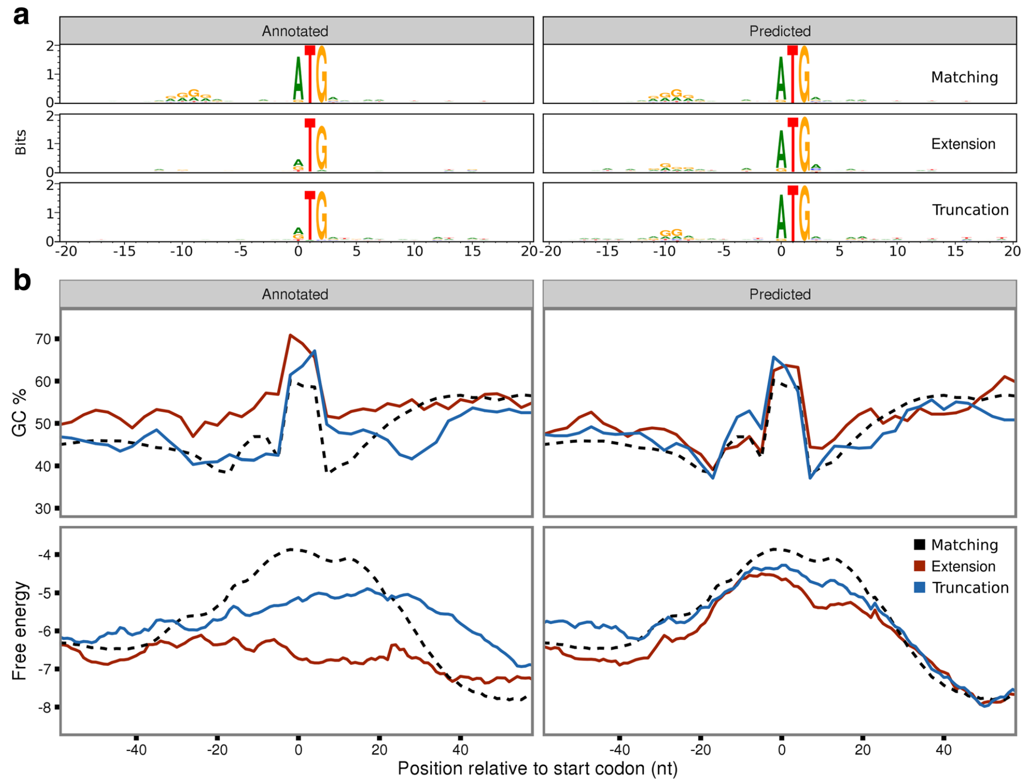


Fig. 4 Sequence and structure features support re-annotation of translation initiation sites (TISs). **a** Sequence motifs relative to annotated or predicted TISs in the same genes. 'Matching' (n = 3853) are identical, while predicted extensions (n = 214) and truncations (n = 205) have stronger Shine-Dalgarno sequences than their annotated counterparts. **b** Meta-profiles relative to annotated or predicted TISs, with lines representing open reading frames matching annotated genes (dashed black), predicted extensions (red) and predicted truncations (blue). (**upper**) Mean guanine-cytosine (GC) content at third codon positions, averaged over 9 nt sliding windows. Predicted TISs match the expected profile more closely than annotated positions (Wilcoxon rank sum test W = 463,640, *P* = 0.001665 for extensions, W = 453,510, *P* = 0.0001546 for truncations), showing an increase in GC content immediately after the start codon, whereas annotated extensions and truncations are less similar to the expected profile (Wilcoxon rank sum test W = 493,250, *P* = 1.226 × 10^{-9} for extensions, Wilcoxon rank sum test W = 460,810, *P* = 5.395 × 10^{-5} for truncations), showing shifts down or upstream in annotated TIS. Peaks over the zero position correspond to nucleotide biases in start codon selection. (**lower**) Meta-profiles of mean free energy averaged in 39 nt sliding windows. Peaks of low secondary structure potential, expected to occur over start codons, are centred over predicted TIS, but are clearly shifted down or upstream of annotated TIS, in predicted extensions and truncations

nucleotide periodicity, provide clear and consistent support that the TISs, which we re-annotate, show, on average, a higher agreement with features indicative of canonically translated prokaryotic ORFs, than their corresponding previously annotated counterparts.

N-terminal proteomics confirms predicted sites

In order to experimentally validate the accuracy of the predictions, positional proteomics analyses enriching for protein N-termini were performed. Blocked N-termini were identified matching 1040 *S.* Typhimurium ORFs, from which a high confidence subset of Nt-formylated initiator methionine (iMet)-starting N-termini was selected (see methods) and used to assess the accuracy of the model. In total, 114 high confidence N-termini were identified, supporting 102 annotated CDSs, 3 N-terminal CDS extensions and 9 N-terminal CDS truncations. Because genomic positions with N-terminal peptide support were excluded from the set used to train the random forest model, these high confidence TIS positions can be used to determine the accuracy of the predictions. Of the 102 N-terminally supported annotated genes, 97 were predicted by the model. Furthermore, two of the extensions and four of the truncations were captured (Fig. 5a, Additional file 2: Table S3). Assuming that none of these genes had multiple initiation sites, the sensitivity, specificity and positive predicted value of the model were estimated to be 0.9450, 0.9993 and 0.9537, respectively.

We found blocked N-terminal peptides matching the predicted start positions from three distinct novel regions (defined as ORFs at least 300 nt in length, in regions that were not overlapping with annotated genes or regions at least 999 bp upstream of annotated genes). Comparing the predictions to the blocked N-termini identified, we found support for 694 predictions that match annotated TISs, 23 predicted extensions, 22 predicted truncations and 3 novel ORFs (Fig. 5b, c, Additional file 2: Table S4).

In order to determine the contribution of sequence and ribosome read length features to the predictive performance, models were trained without including either of these feature sets or with just one of these feature sets. When either the ribo-seq read length or sequence information is excluded from the model, we observed a decrease in sensitivity (0.8785 and 0.3364 when excluding ribo-seq read lengths or sequence features, respectively, compared to 0.9450 when all features are used), while maintaining high specificity (0.9990 in both cases) (Additional file 2: Table S5). We observed that sequence features alone were able to correctly identify 85 of the 114 high confidence N-terminally supported TISs, and that ribo-seq read length information alone was able to correctly identify 36 of the 114 high confidence N-terminally supported TIS. Although sequence information had a larger impact on sensitivity than ribo-seq read lengths, the optimal values were only achieved when both sequence and read length features were used in combination (Additional file 2: Table S5).

TISs are predicted at novel genomic regions

In order to discover potential novel translated ORFs, we applied our prediction models to look for TISs in genomic regions outside annotated ORFs. Novel ORFs that were similar in size to known CDSs (>100 amino acids) and with ribo-seq coverage along a high proportion of the ORF (>75% coverage, see methods) were considered candidate translated novel ORFs. Of the 219 (monosome) and 193 (polysome) ORFs under consideration, 104 and 115 novel translated ORFs were predicted, respectively; 61 of these novel translated ORFs were common to both replicates (38.61% agreement) and used as a high confidence set of novel predictions. Unlike the annotated genes, these novel ORFs were not previously confirmed as translated regions and most had a significantly lower read density (mean fragments per kilobase per million mapped reads (FPKM) of 8) than annotated genes (mean FPKM of 126). The higher discrepancy between the two replicates is mainly a consequence of low-abundance ORFs that did not pass the coverage threshold in either of the replicates.

Read density plots over the novel ORFs revealed features consistent with protein coding regions, but with higher variance due to the low number of ORFs. Specifically, GC content increases downstream of the initiation codon, the regions around the initiation codon have less intrinsic structure potential and SD sequences were present upstream (Additional file 1: Figure S6). Additionally, three of the predicted novel translated ORFs were supported by N-terminal peptide evidence (Fig. 5c)

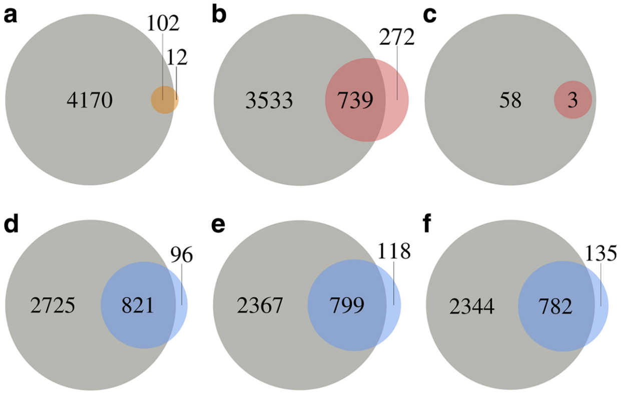


Fig. 5 Predicted open reading frames (ORFs) show high agreement with validation datasets. Venn diagram showing the agreement of predicted *S.* Typhimurium ORFs with (**a**) high confidence N-terminal peptides (orange) or (**b**) blocked N-terminal peptides (red), novel predicted *S.* Typhimurium ORFs (**c**) with blocked N-terminal peptides supporting novel ORFs (red) and predicted ORFs from the *E. coli* tetracycline (**d**), Li et al. (**e**) or Mohammad et al. (**f**) datasets with ecogene verified protein starts (blue)

(a representative example is shown in Fig. 2). A further 22 newly predicted ORFs showed high similarity to known protein sequences, four of which contained functional protein domains (Additional file 2: Table S6).

Tetracycline-treated samples improve classifier accuracy

While reads isolated from elongating ribosomes provide sufficient information to predict the majority of translation start sites, we set out to explore the full potential of our classifier in combination with data from initiating ribosomes. A recent study on *E. coli* [12] demonstrated the use of tetracycline as a translation inhibitor to enrich for footprints from initiating ribosomes in prokaryotes. Herein, the tetracycline datasets showed the expected pattern from initiating ribosomes as a range of read lengths starting 28–14 nt upstream of the initiation codon (5' data), but ending at the same positions 14–15 nt downstream of the initiation codon (3' data). An additional pattern of shorter read lengths was also observed, starting 26–18 nt upstream and ending 2 nt downstream of the initiation codon (Additional file 1: Figure S7a, b). Complementary datasets were selected from publically available *E. coli* ribo-seq libraries collected via flash freezing [17, 40]; in these datasets, the pattern was formed by an enriched set of reads lengths beginning 14 to 23 nt upstream of the initiation codon, ending directly at the initiation codons, and by a secondary set of read lengths starting at the initiation codons and ending 19–39 nt downstream (Additional file 1: Figure S7c–f).

We trained separate classifiers on tetracycline (initiating) libraries and flash frozen (non-specific) datasets, using two replicates for each dataset (Additional file 2: Table S7). Model performance was assessed with receiver operating characteristic curves using a validation dataset (see methods) for each replicate. The resulting AUC values of 0.9993 and 0.9994 in the tetracycline replicates were comparable to those of the flash frozen libraries (Li: 0.9998 and 0.9997, Mohammad: 0.9998 and 0.9993). The parameter importance in each of the models is shown in Additional file 2: Tables S8–S13. In the tetracycline dataset, a total of 3711 ORFs were predicted, with 86 extensions and 79 truncations (Additional file 2: Table S14). The flash frozen models predicted a total of 3269 and 3341 ORFs, including 48 and 73 extensions and 95 and 102 truncations in the Li and Mohammad libraries, respectively (Additional file 2: Table S15, S16).

E. coli predictions were assessed against the ecogene curated set of 923 experimentally verified protein starts [41]. Genes within this dataset were excluded from the sets used to train the random forest models in order to provide a means of assessing the accuracy of the ORF predictions. Five of the verified protein starts corresponded to pseudogenes without annotated CDSs; of the remaining 917 verified protein starts, 821 (89.53%) matched ORFs from the tetracycline predictions, with 24 (2.62%) predicted ORFs in disagreement with the curated set (11 extensions, 13 truncations) (Fig. 5d, Additional file 2: Table S14). In the Li flash frozen predictions, 782 (85.28%) were found to match ecogene start sites and 26 (2.84%) were found to be inconsistent (7 extensions, 18 truncations) with the verified protein starts (Fig. 5e, Additional file 2: Table S15). Of the Mohammad flash frozen predictions, 779 (84.95%) were found to match and 29 (3.16%) were found to disagree (12 extensions, 17 truncations) with ecogene verified protein starts (assuming genes do not have multiple TISs) (Fig. 5f, Additional file 2: Table S16). Based on the experimentally verified starts, the tetracycline-based classifier resulted in higher accuracy (sensitivity 0.9194, specificity 0.9996, positive predictive value 0.9716) than either of the flash frozen-based classifiers (sensitivity 0.8777/0.8773, specificity 0.9996/0.9995, positive predictive value 0.9678/0.9641 (Li/Mohammed)). Surprisingly, the difference was not substantial, arguing that using initiating ribosomes is not a prerequisite to obtain a good annotation of initiation sites.

Discussion

Our model shows that the distribution of ribo-seq footprint lengths can be used in conjunction with sequence features to accurately determine the translation initiation landscape of prokaryotes. These patterns are typically disrupted in standard ribo-seq analysis when reads are adjusted and merged to determine the specific codon under translation. The model is applicable across multiple organisms and experimental conditions, and can be augmented with data from initiating ribosomes. It exhibits high accuracy as assessed by cross-validation, N-terminal proteomics and independent sequence-based metrics such as potential to form RNA structures. Interestingly, the predicted TISs exhibited known features of translation initiation, while the previously matching annotations do not. In *S.* Typhimurium, our model provides evidence for 61 novel translated ORFs and the re-annotation of 419 genes. In particular, the current annotation includes 19 genes that lacked initiation codons, of which we were able to re-annotate 15 (Additional file 2: Table S4).

As expected, models based on initiating reads performed better than models based on non-specific ribo-seq reads, suggesting that an optimal strategy for TIS identification would favour the use of the more focused, initiating ribo-seq profiles. However, the degree of improvement between the models was relatively small, confirming the suitability of both non-specific and initiating ribo-seq libraries for the purposes of TIS and ORF detection.

Footprints containing SD sequences have been shown to produce longer ribo-seq reads, attributed to nuclease protection from RNA/anti-SD interactions [16, 17]. We observed enrichments of these longer read lengths at footprints overlapping the SD, but also found a second population of shorter SD associated reads, less prominent at internal SD sequences than those upstream of TISs (Additional file 1: Figure S2). It is interesting to note the importance the models place (Additional file 2: Tables S1, S2) on this shorter range of reads (23–25 nt) in the *S.* Typhimurium samples (Additional file 1: Figure S8). These shorter reads were also consistent with recent reports of ribosomal subunits in a variety of distinct configurations, observed from translation complex profiling in the eukaryote *Saccharomyces cerevisiae* [42]. Whether similar patterns of read length distributions can be observed in eukaryotic ribo-seq datasets remains to be determined, although conceptually the method and metrics described herein are fully extendable to eukaryotic datasets.

A key strength of this approach is that the model is able to build complicated rules incorporating multiple sets of features from changes in ribosome footprint density, sequence context and ribosomal profiling patterns indicative of RNA/ribosome interactions and initiation codons. Leaderless genes, for example, might not be expected to have an SD motif, but the flexibility in our model would allow these to be identified if other features were suggestive of TISs. For example, 3387 of the *S.* typhimurium TISs predicted do not have a strong SD sequence (defined as binding energy of ≤ −8 kcal/mol, see methods).

While the model relies on some pre-existing annotated ORFs for training, it does not require any prior knowledge, but rather detects the RNA/ribosome interactions of SD and initiation sites from patterns in the fragment length of protected reads. This may provide a fruitful avenue for exploring novel sequence features, for example, using patterns in protected read lengths as a proxy to identify ribosome/RNA interactions in species with alternative ribosome binding motifs or initiation mechanisms.

Conclusions

In conclusion, this study demonstrates the utility of ribo-seq fragment length patterns for TIS identification across multiple experimental conditions. These models provide a significant step forward in experimental TIS discovery, facilitating the move towards complete ORF annotation in both presumably well-annotated model organisms, as well as the ever growing list of newly sequenced genomes.

Methods

Preparation of ribo-seq libraries

Overnight stationary cultures of wild type *S.* Typhimurium (*Salmonella enterica* serovar Typhimurium, strain SL1344) grown in LB media at 37 °C with agitation (200 rpm) were diluted at 1:200 in LB and grown until they reached and OD600 of 0.5 (i.e. logarithmic (Log) phase grown cells). Bacterial cells were pre-treated for 5 min with chloramphenicol (Sigma Aldrich) at a final concentration of 100 μg/mL prior to collection by centrifugation (6000 × *g*, 5 min) at 4 °C. Collected cells were flash frozen in liquid nitrogen. The frozen pellet of a 50 mL culture was re-suspended and thawed in 1 mL ice-cold lysis buffer for polysome isolation (10 mM $MgCl_2$, 100 mM NH_4Cl, 20 mM Tris-HCl pH 8.0, 20 U/mL of RNase-free DNase I (NEB 2 U/μL), 1 mM chloramphenicol (or 300 μg/mL), 20 μL/mL lysozyme (50 mg/mL in water) and 100 μ/mL SUPERase.In™ RNase Inhibitor (Thermo Fisher Scientific, Bremen, Germany)), vortexed and left on ice for 2 min with periodical agitation. Subsequently, the samples were subjected to mechanical disruption by two repetitive cycles of freeze-thawing in liquid nitrogen, and 5 mM $CaCl_2$, 30 μL 10% DOC and 1 × complete and EDTA-free protease inhibitor cocktail (Roche, Basel, Switzerland) were added and the mixture was left on ice for 5 min. Lysates were clarified by centrifugation at 16,000 × *g* for 10 min at 4 °C.

For the monosome sample, the supernatant was subjected to MNase (Roche Diagnostics, Belgium) digestion using 600 U MNase (~1000 U per mg of protein). Digestion of polysomes proceeded for 1 h at 25 °C with gentle agitation at 400 rpm and the reaction was stopped by the addition of 10 mM EGTA. Next, monosomes were recovered by ultracentrifugation over a 1 M sucrose cushion in polysome isolation buffer without RNase-free DNase I and lysozyme, and by the addition of 2 mM DTT using a TLA-120.2 rotor for 4 h at 75,000 rpm and 4 °C.

For the selective purification of monosomes from polysomes (polysome sample), the supernatant was resolved on 10–55% (w/v) sucrose gradients by centrifugation using an SW41 rotor at 35,000 rpm for 2.5 h at 4 °C. The sedimentation profiles were recorded at 260 nm and the gradient fractionated using a BioComp Gradient Master (BioComp) according to the manufacturer's instructions. Polysome-enriched fractions were pooled and subjected to MNase digestion and monosome recovery, as described above.

Ribosome-protected mRNA footprints, with sizes ranging from 26 to 34 nt, were selected and processed as described previously [14], with some minor adjustments [43]. The resulting ribo-seq cDNA libraries of the monosome and polysome sample were duplexed and sequenced on a NextSeq 500 instrument (Illumina) to yield 75 bp single-end reads.

Ribo-seq data processing

Ribo-seq data were pre-processed with cutadapt (version 1.9.1) [44] to remove sequencing adaptors, discarding reads less than 20 nt in length after trimming. Trimmed reads were initially aligned to the SILVA RNA database version 119 [45], the remaining reads were then mapped to either *S. enterica* serovar Typhimurium, strain SL1344 (Assembly: GCA_000210855.2) or *E. coli* str. K-12 substr. MG1655 (Assembly: GCA_000005845.2). Alignments were performed with bowtie2 (version 2.2.4) [46]. Reads were brought to codon resolution by adjusting the 5’ position of each read by a fixed distance offset, specific to each read length, based on visual identification of periodicity meta plots of the read counts per read length (Additional file 1: Figure S9). In the *S.* Typhimurium dataset the following read lengths were selected and adjusted by the values in brackets, in the monosome sample 29 (13 nt), 30 (14 nt), 31 (15 nt), 32 (16 nt), 33 (17 nt), and in the polysome sample 29 (13 nt), 30 (14 nt), 31 (15 nt), 32 (16 nt), 33 (17 nt) and 34 (18 nt). Selected reads of the indicated lengths account for 39.98% and 48.69% of total reads for the monosome and polysome samples, respectively.

Recent publications reporting prokaryotic ribo-seq [12, 17, 20, 47] suggest that reads from libraries digested with micrococcal nuclease align more precisely to their 3’ rather than 5’ ends. Consistent with this, we observed a modest increase in the periodicity of meta profiles of the *S.* Typhimurium ribo-seq libraries when reads were brought to codon resolution from the 3’ end (Additional file 1: Figure S1a–d); however, this did not hold true for the tetracycline *E. coli* datasets, where the use of 3’ poly adenosine adaptors resulted in a loss of resolution at the 3’ end after read trimming (Additional file 1: Figure S7a, b), making the use of 5’ ends preferable. Since the protected read length patterns used in the input feature vectors for the classifier take both length and position into consideration, the classifier is unaffected by the alignment choice for generating positional data. However, to maintain consistency throughout this study, read counting for model predictors was performed using the 5’ alignments, while the periodicity plots, which are sensitive to read terminus choice, were calculated from either 3’ alignments or from ribo-seq reads after read length-specific read adjustment.

Read distributions and heatmaps

Ribo-seq read distributions were summarised over all annotated start codons in the *S.* Typhimurium and *E. coli* annotations, respectively. 5’ read counts were taken from regions 30 nt upstream to 60 nt downstream of the start codon (or −100 to −10 nt upstream of the stop codon), 3’ read counts were taken from the first nucleotide of the start codon up to 90 nucleotides downstream (or −70 upstream to 20 nt downstream of the stop codon). All reads with a MAPQ greater than 10, from the upper 90% of genes by total CDS expression, were included. Total counts were scaled to a sum of one per individual region, in order to not disproportionately favour profiles from highly expressed genes. Meta plots were then produced to show the proportion of read counts over the window across all genes. 3’ and 5’ heatmaps were generated from the scaled regions, showing the number of standard deviations from the row (read length) mean. ATG, GTG and TTG codons were taken from in-frame CDS regions, excluding annotated and predicted start codons. SD motifs were identified as sequences predicted to have a binding energy with the anti-SD sequence (AGGAGGTG) of −8 kcal/mol or lower. Energies were calculated in 8 nt overlapping windows across the whole genome with RNAsubopt (version 2.1.9) [48], and assigned to the first “A” of the anti-SD sequence. Upstream SD sequences were defined as those within 30 nt upstream of an annotated start codon. Downstream SD sequences, were defined as SD motifs within annotated CDS regions, excluding those within 50 nt of the start codon of a downstream annotated or predicted ORF. Third codon position GC content was calculated from nucleotide sequences 60–150 nt downstream of the start codon for the upper 90% of genes by total CDS expression. Three prime read distributions were plotted for the upper and lower 10% of sequences based on total third position GC content.

Model implementation

For each candidate TIS, a feature vector was defined as each nucleotide in a −20 to +10 nt window around the position, the ribo-seq 5’ FPKM between the current position and the next in-frame downstream stop codon, the count of in-frame cognate and near-cognate start sites from the nearest in-frame upstream stop codon to the current position, the proportion of 5’ ribo-seq reads upstream in a 20 nt window, the proportion of 5’ ribo-seq reads downstream in a 20 nt window, the ratio of 5’ ribo-seq reads up and downstream, and the proportion 5’ ribo-seq counts per read length for a fixed range of positions in relation to current site (selected from visual inspection of 5’ read length heatmaps (Additional file 1: Figure S1, S7)). In the *S.* Typhimurium samples, read lengths of 20–35 nt in positions −20 to −11 and 0 nt, were used. In the *E. coli* datasets for the tetracycline samples, read lengths of 20–35 nt at positions −25 to −16 nt were selected. Finally, for flash frozen samples, lengths of 20–35 nt at positions −20 to −11 and 0 nt were used. Stop-to-stop windows were defined for each annotated gene as all in-frame positions between the nearest in-frame upstream stop codon and the stop codon of the gene (with a maximum length cut-off 999 nt upstream).

The H2O random forest implementation (version 3.10.4.6) [49] was used and the models were trained with positive examples of randomly selected annotated start codons from the upper 50% of genes ranked by ribo-seq expression over the gene CDS. We additionally required that the positive examples were not among the genes supported by N-terminal peptides in the *S.* Typhimurium samples or included in the ecogene dataset for the *E. coli* samples, since these were retained for model accuracy assessment. Negative examples were randomly selected from in-frame codons in the stop-to-stop windows both upstream and downstream of the annotated TIS. The *S.* Typhimurium models were trained on 1500 positive and 6000 negative positions, with an independent validation set of 200 positive and 800 negative positions. Parameter tuning was performed for the number of trees in the random forest, using values from 50 to 1000 with a step size of 50, and selecting the value which produced the highest AUC values on the validation set (monosome: 600, polysome: 600) (see code for an example of automated parameter tuning). The *E. coli* models were trained on 1100 positive and 4400 negative positions, with an independent validation set of 200 positive and 800 negative positions for parameter tuning (number of trees: Li1: 550, Li3: 450, Mohammad1: 600, Mohammad2: 700, TET2: 650 and TET3: 700). Predictions were then run against all cognate and near cognate in-frame positions, in the stop-to-stop regions. Novel predictions were performed against all cognate and near cognate codons in stop-to-stop regions around ORFs of at least 300 nt in length, with a ribo-seq read coverage of 0.75 or more (ORF coverage was defined as the proportion of nucleotides in each predicted ORF that at least one ribo-seq read mapped to), that did not overlap with annotated CDSs. ORFs were delineated by extending each candidate TIS to the closest in-frame stop codon. For a given stop-to-stop region the model selected the TIS with the highest positive predicted score per sample. Predictions from the replicates for each of the datasets were then compared, discarding predictions that were unique to only one replicate.

N-terminal proteomics

Overnight stationary cultures of wild type *S.* Typhimurium (*S. enterica* serovar Typhimurium, strain SL1344) grown in LB media at 37 °C with agitation (200 rpm) were diluted at 1:200 in LB and grown until they reached an OD600 of 0.5 (i.e. logarithmic (Log) phase grown cells). Bacterial cells were collected by centrifugation (6000 × *g*, 5 min) at 4 °C, flash frozen in liquid nitrogen and cryogenically pulverized using a pestle and mortar cooled with liquid nitrogen. The frozen pellet of a 50 mL culture was re-suspended and thawed in 1 mL ice-cold lysis buffer (50 mm NH_4HCO_3 (pH 7.9) supplemented with a complete protease inhibitor cocktail tablet (Roche Diagnostics GmbH, Mannheim, Germany) and subjected to mechanical disruption by two repetitive freeze-thaw and sonication cycles (i.e. 2 minutes of sonication on ice for 20-s bursts at output level 4, with a 40% duty cycle (Branson Sonifier 250; Ultrasonic Convertor)). The lysate was cleared by centrifugation for 15 min at 16,000 × *g* and the protein concentration measured using the protein assay kit (Bio-Rad) according to the manufacturer's instructions. The guanidine hydrochloride (4 M f.c.) was added to the lysate and subjected to N-terminal COFRADIC analysis, as described previously [50]. Free amines were blocked at the protein level making use of an N-hydroxysuccinimide ester of (stable isotopic encoded) acetate (i.e. NHS esters of $^{13}C_2D_3$ acetate), which allows the distinction of in vivo and in vitro blocked N-terminal peptides [51]. The modified protein sample was digested overnight with sequencing-grade modified trypsin (1/100 (w/w trypsin/substrate)) at 37 °C and subsequent steps of the N-terminal COFRADIC procedure were performed as previously described [50].

LC-MS/MS analysis

LC-MS/MS analysis was performed using an Ultimate 3000 RSLC nano HPLC (Dionex, Amsterdam, the Netherlands) connected in line to an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). The sample mixture was loaded on a trapping column (made in-house, 100 μm ID × 20 mm, 5 μm beads C18 Reprosil-HD, Dr Maisch). After back flushing from the trapping column, the sample was loaded on a reverse-phase column (made in-house, 75 m ID × 150 mm, 5 μm beads C18 Reprosil-HD, Dr Maisch). Peptides were loaded in solvent A' (0.1% trifluoroacetic acid, 2% acetonitrile) and separated with a linear gradient from 2% solvent A" (0.1% formic acid) to 50% solvent B' (0.1% formic acid and 80% acetonitrile) at a flow rate of 300 nL/min followed by a wash reaching 100% solvent B'. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS/MS acquisition for the 10 most abundant peaks in a given MS spectrum. Full scan MS spectra were acquired in the Orbitrap at a target value of 1×10^6 at a resolution of 60,000. The 10 most intense ions were then isolated for fragmentation in the linear ion trap, with a dynamic exclusion of 20 s. Peptides were fragmented after filling the ion trap at a target value of 1×10^4 ion counts. Mascot Generic Files were created from the MS/MS data in each LC run using the Mascot Distiller software (version 2.5.1.0, Matrix Science, www.matrixscience.com/

Distiller.html). To generate these MS/MS peak lists, grouping of spectra was allowed with a maximum intermediate retention time of 30 s and a maximum intermediate scan count of 5. Grouping was performed with a 0.005 Da precursor tolerance. A peak list was only generated when the MS/MS spectrum contained more than 10 peaks. There was no de-isotoping and the relative signal-to-noise limit was set at 2.

The generated MS/MS peak lists were searched with Mascot using the Mascot Daemon interface (version 2.5.1, Matrix Science). Searches were performed using a 6-FT database of the *S.* Typhimurium genome combined with the Ensembl protein sequence database (assembly AMS21085v2 version 86.1), which totalled 139,408 entries after removal of redundant sequences. The 6-FT database was generated by traversing the entire genome across the six reading frames and searching for all NTG (N = A, T, C, G) start codons and extending each to the nearest in frame stop codon (TAA, TGA, TAG), discarding ORFs less than 21 nt in length. The Mascot search parameters were set as follows: heavy acetylation at lysine side-chains (Acetyl:2H(3)C13(2) (K)), carbamidomethylation of cysteine and methionine oxidation to methionine-sulfoxide were set as fixed modifications; and formylation, acetylation and heavy acetylation of N-termini (Acetyl:2H(3)C13(2) (N-term)) and pyroglutamate formation of N-terminal glutamine (both at peptide level) were set as variable modifications. Endoproteinase semi-Arg-C/P (semi Arg-C specificity with Arg-Pro cleavage allowed) was set as the enzyme, allowing for no missed cleavages. Mass tolerance was set to 10 ppm on the precursor ion and to 0.5 Da on fragment ions. Peptide charge was set to 1+, 2+ and 3+, and the instrument setting was switched to ESI-TRAP. Only peptides ranked the highest, had a minimum amino acid length of seven, scored above the threshold score set at 95% confidence, and belonged to the category of in vivo- or in vitro-blocked N-terminal peptides compliant with the rules of iMet processing [52] were withheld. More specifically, iMet processing was considered in the case of iMet-starting N-termini followed by any of Ala, Cys, Gly, Pro, Ser, Thr, Met or Val, and only if the iMet was encoded by ATG or any of GTG or TTG near-cognate start codons (Additional file 2: Table S17). While the occurrence of N-terminal protein acetylation (Nt-acetylation) and Nt-formyl retention are not trivial in bacteria (i.e. N-terminal protein acetylation and retention of the Nt-formyl group affected about 10% and 5% of uniquely identified protein in *E. coli*), the low degree of these N-terminal modifications at steady-state levels [53] – a finding in contrast to eukaryotic nascent protein N-termini – warrant caution to unequivocally assign bacterial protein N-termini as proxies of translation initiation. Because of the aforementioned reasons, we only selected Nt-formylated iMet-starting N-termini as a high confidence subset of TIS-indicative N-termini to experimentally validate the accuracy of the predictions (Additional file 2: Table S18).

Assessing model accuracy

Sensitivity, specificity and positive predictive values were calculated from all genes supported by either high confidence n-terminal peptides (*S.* Typhimurium) or experimentally verified protein starts (*E. coli*). Supported predicted ORFs were considered true positives, whereas predicted ORFs that disagreed with supported positions were classified as false positives. False negatives were assigned from supported genes where no ORF was predicted. All in-frame cognate and near cognate start codons in stop-to-stop regions of supported genes that were neither predicted nor supported were considered true negatives.

Further support for predicted ORFs

GC content was calculated at the third nucleotide positions for all annotated and predicted ORFs and mean GC values were summarised for each subgroup of predicted ORFs (matching annotations, truncations and extensions) in 9 nt sliding windows, over regions 57 nt upstream and 57 nt downstream of the annotated or predicted start sites.

Nucleotide sequences (−20 to + 20 nt) were extracted around the predicted and annotated TIS in the *S.* Typhimurium and *E. coli* genomes. Sequence logos were generated for each subgroup of matching annotations, truncations, extensions and novel genes, using the weblogo tool [54].

The minimum free energy of RNA secondary structure around predicted and annotated ORFs was estimated with RNAfold version 2.1.9 from the ViennaRNA package [48]. Mean free energy values were summarised for each ORF class in a 39-nt sliding window across regions 57 nt up and downstream of the start codon.

Read distributions were created for each subgroup of predicted ORFs (matching annotations, truncations, extensions and novel genes) and their corresponding annotated TIS. Distributions of ribo-seq reads adjusted to codon level resolution were summarised in regions 30 nt upstream and downstream of the first nucleotide of the initiation codon and total counts of each individual region were scaled to a sum of one in order to normalise profiles for differences in gene expression levels. Meta plots were then produced to show the proportion of reads over the window position from all predicted subgroups and their corresponding annotated start codons.

Amino acid sequences of novel ORF were compared to known proteins in the non-redundant protein database (Update date: 2016/12/15) and protein domains (cdd.v.3.15) using BLASTP [48, 55] (version 2.5.1+). Hits with the greatest coverage of query sequence and lowest e-value were selected. Hits were considered highly similar if they shared > 95% identity to a protein sequence over 100% of the novel ORF sequence.

Statistical analysis

Wilcoxon rank sum tests were performed on ribo-seq distributions in Fig. 3a. The ratio of mean ribo-seq counts, per gene, upstream (positions –30 to –1 nt) or downstream (positions 0 to 29 nt) for extended and truncated positions were compared to matching positions using Wilcoxon rank sum test with continuity correction. Third codon GC distributions in Fig. 4b (upper) were also assessed with Wilcoxon rank sum tests with continuity correction, comparing the difference between the mean ribo-seq counts in downstream regions (positions 18 to 57 nt) and upstream regions (positions –57 to 18 nt), per gene, for extended or truncated and matching genes. Regions were selected to exclude the bias caused by SD and start codon sequence composition (peaks in the –18 to 6 nt regions).

Additional files

Additional file 1: Figures S1–S9. Figure S1. Ribo-seq meta profiles at start and stop codons *S.* Typhimurium. **Figure S2.** Read length distributions at Shine–Dalgarno motifs. **Figure S3**: Codon-specific read length distributions. **Figure S4.** Additional prediction support. **Figure S5.** Third codon periodicity and GC content. **Figure S6.** Evidence for predicted novel translation initiation sites. **Figure S7.** Ribo-seq meta profiles at start codons for *E. coli*. **Figure S8.** Library read length distributions. **Figure S9.** Read length adjustments. (PDF 4700 kb)

Additional file 2: Tables S1–S18. Table S1. Variable importance in the *S.* Typhimurium monosome sample. **Table S2.** Variable importance in the *S.* Typhimurium polysome sample. **Table S3.** N-terminal support for *S.* Typhimurium predicted ORFs. **Table S4.** Predicted ORFs from the *S.* Typhimurium dataset. **Table S5.** Assessment of the contribution of parameter types to the predictive performance. **Table S6.** Support for novel predicted ORFs. **Table S7.** Ribo-seq sample info. **Table S8.** Variable importance in the *E. coli* TET2 sample. **Table S9.** Variable importance in the *E. coli* TET3 sample. **Table S10.** Variable importance in the *E. coli* Li1 sample. **Table S11.** Variable importance in the *E. coli* Li3 sample. **Table S12**: Variable importance in the *E. coli* Mohammad1 sample. **Table S13.** Variable importance in the *E. coli* Mohammad2 sample. **Table S14.** ORF predictions in the *E. coli* tetracycline libraries. **Table S15.** ORF predictions in the *E. coli* Li libraries. **Table S16.** ORF predictions in the *E. coli* Mohammad libraries. **Table S17.** Blocked N-terminal peptides. **Table S18.** High confidence N-terminal peptides. (XLSX 321 kb)

Abbreviations

AUC: area under curve; CDS: protein coding sequence; FPKM: fragments per kilobase per million mapped reads; iMet: initiator methionine; nt: nucleotide; ORF: open reading frame; ribo-seq: ribosome profiling; SD: Shine–Dalgarno sequence; TIS: translation initiation site

Acknowledgements

We thank Zoya Ignatova, Gerben Menschaert, Gunnar Schulze and Kornel Labun for valuable discussions.

Funding

We thank Prof. Kris Gevaert for financial support of this research (Research Foundation – Flanders (FWO-Vlaanderen), project number G.0440.10). PVD acknowledges support from the FWO-Vlaanderen, project number G.0269.13 N. AG and EV acknowledge support from the Bergen Research Foundation.

Availability of data and materials

S. Typhimurium ribo-seq sequencing data has been deposited in NCBI's Gene Expression Omnibus [56] and is accessible through GEO series accession number GSE91066. *S.* Typhimurium mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [57] partner repository with the dataset identifier PXD005579. The previously published *E. coli* ribo-seq datasets were downloaded from BioProject ID: PRJDB2960, and GEO accession numbers GSE53767 and GSE72899. Custom scripts used in this analysis are available at the following location: https://bitbucket.org/valenlab/giess-scripts/src/master/TIS_prediction/.

Authors' contributions

AG, EV and PVD conceived the study and wrote the manuscript. AG and KC performed the computational analysis. PVD performed the proteomics experiment. EN and PVD performed proteomics analysis. PVD and VJ prepared the ribo-seq libraries. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing financial interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen 5020, Norway. [2]VIB-UGent Center for Medical Biotechnology, B-9000 Ghent, Belgium. [3]Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium. [4]Lab of Bioinformatics and Computational Genomics, Department of Mathematical Modelling, Statistics and Bioinformatics, Faculty of Bioscience Engineering, Ghent University, B-9000 Ghent, Belgium. [5]Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway.

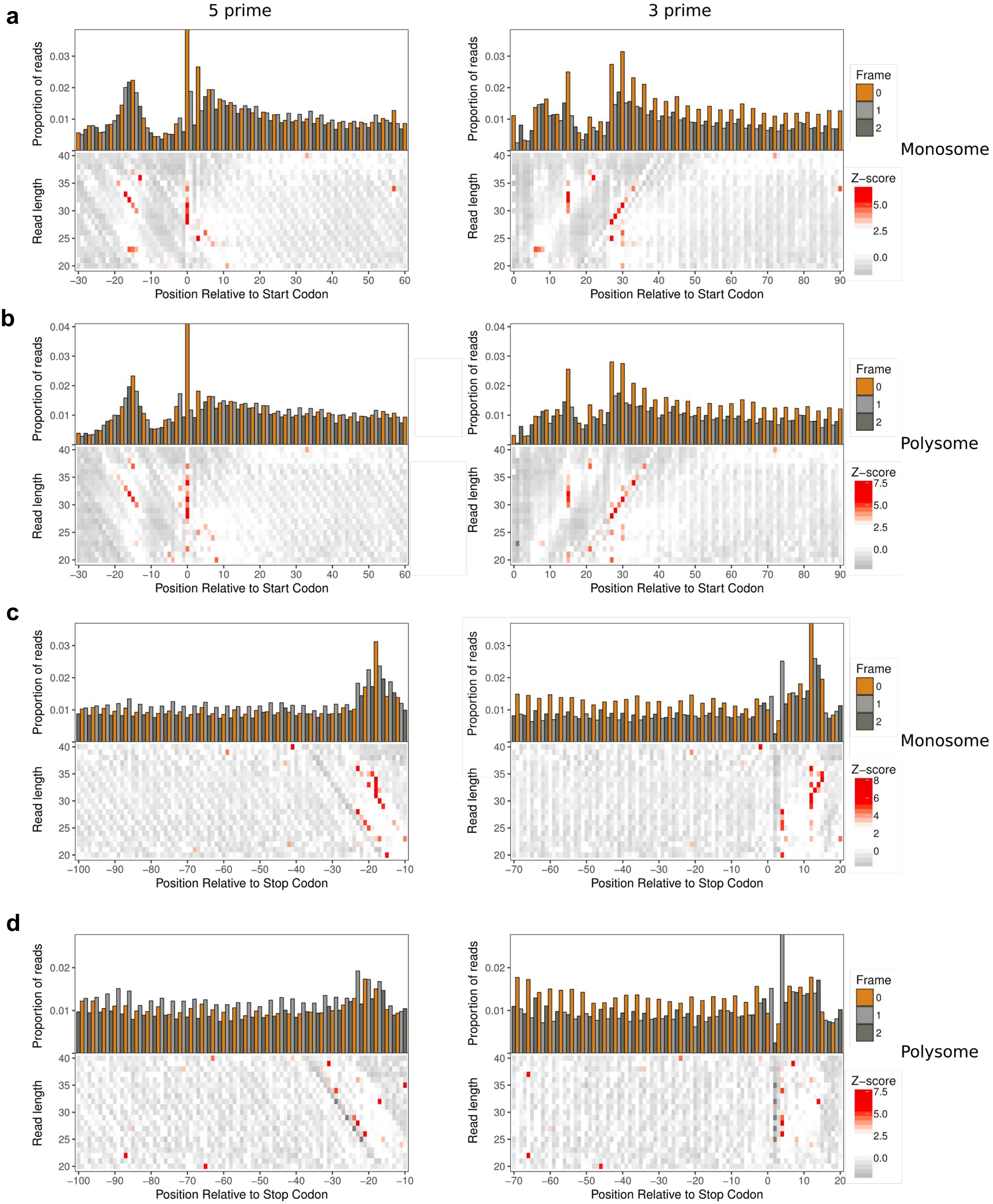
Received: 9 June 2017 Accepted: 9 August 2017
Published online: 30 August 2017

References

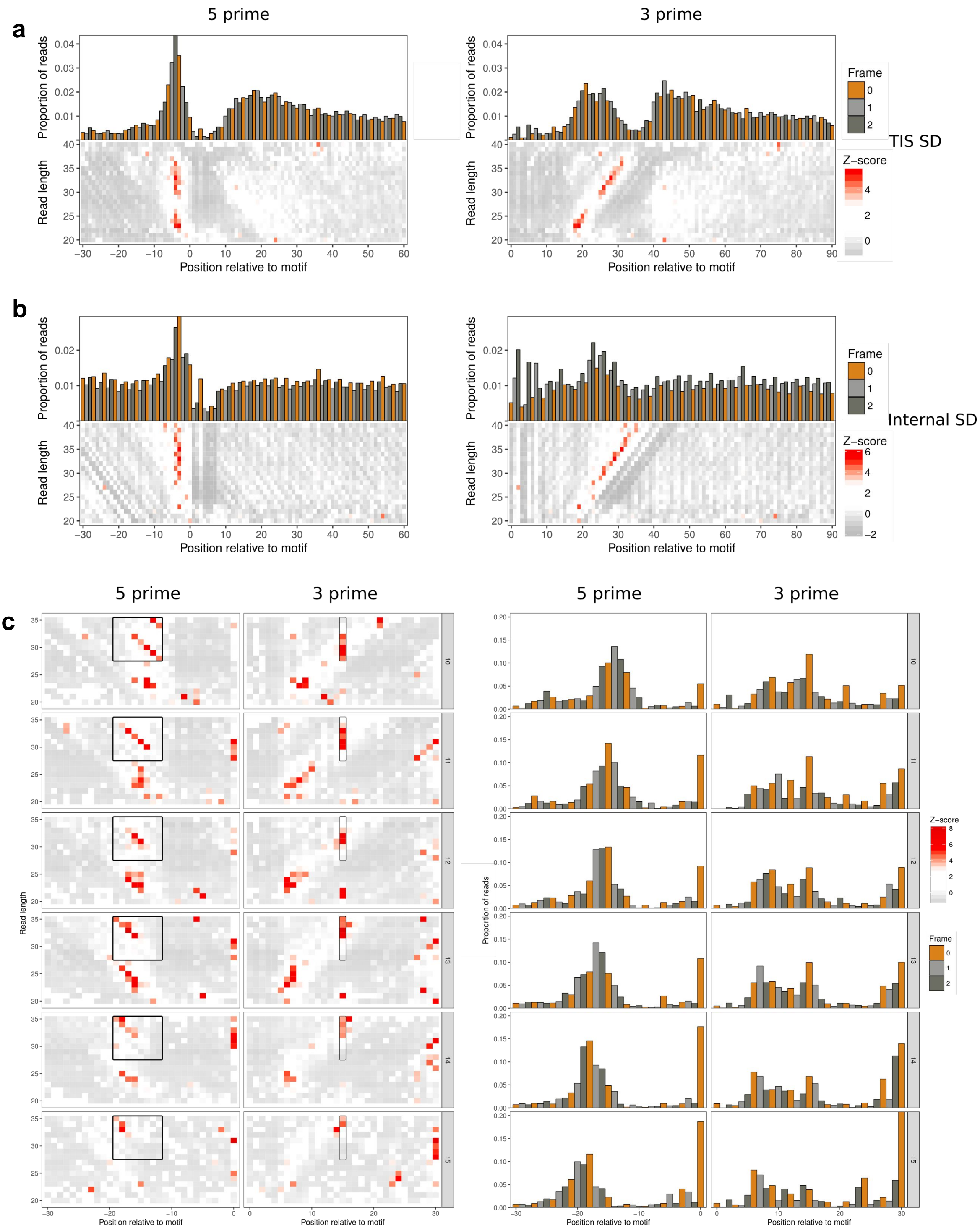
1. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1999;27:4636–41.
2. Brocchieri L, Kledal TN, Karlin S, Mocarski ES. Predicting coding potential from genome sequence: application to betaherpesviruses infecting rats and mice. J Virol. 2005;79:7570–96.
3. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
4. Hall J, Hazlewood GP, Surani MA, Hirst BH, Gilbert HJ. Eukaryotic and prokaryotic signal peptides direct secretion of a bacterial endoglucanase by mammalian cells. J Biol Chem. 1990;265:19996–9.
5. Kozak M. Initiation of translation in prokaryotes and eukaryotes. Gene. 1999; 234:187–208.
6. Suzek BE, Ermolaeva MD, Schreiber M, Salzberg SL. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics. 2001;17: 1123–30.
7. Zhu H-Q, Hu G-Q, Ouyang Z-Q, Wang J, She Z-S. Accuracy improvement for identifying translation initiation sites in microbial genomes. Bioinformatics. 2004;20:3308–17.

8. Ou H-Y, Guo F-B, Zhang C-T. GS-Finder: a program to find bacterial gene start sites with a self-training method. Int J Biochem Cell Biol. 2004;36:535–44.
9. Tech M, Morgenstern B, Meinicke P. TICO: a tool for postprocessing the predictions of prokaryotic translation initiation sites. Nucleic Acids Res. 2006; 34:W588–90.
10. Hartmann EM, Armengaud J. N-terminomics and proteogenomics, getting off to a good start. Proteomics. 2014;14:2637–46.
11. Berry IJ, Steele JR, Padula MP, Djordjevic SP. The application of terminomics for the identification of protein start sites and proteoforms in bacteria. Proteomics. 2016;16:257–72.
12. Nakahigashi K, Takai Y, Kimura M, Abe N, Nakayashiki T, Shiwa Y, et al. Comprehensive identification of translation start sites by tetracycline-inhibited ribosome profiling. DNA Res. 2016;23:193–201.
13. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science. 2009;324:218–23.
14. Ingolia NT, Lareau LF, Weissman JS. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell. 2011;147:789–802.
15. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, et al. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. 2014;8:1365–79.
16. O'Connor PBF, Li G-W, Weissman JS, Atkins JF, Baranov PV. rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. Bioinformatics. 2013;29:1488–91.
17. Mohammad F, Woolstenhulme CJ, Green R, Buskirk AR. Clarifying the translational pausing landscape in bacteria by ribosome profiling. Cell Rep. 2016;14:686–94.
18. Bazzini AA, Johnstone TG, Christiano R, Mackowiak SD, Obermayer B, Fleming ES, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. 2014;33:981–93.
19. Han Y, Gao X, Liu B, Wan J, Zhang X, Qian S-B. Ribosome profiling reveals sequence-independent post-initiation pausing as a signature of translation. Cell Res. 2014;24:842–51.
20. Woolstenhulme CJ, Guydosh NR, Green R, Buskirk AR. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. Cell Rep. 2015;11:13–21.
21. Chew G-L, Pauli A, Rinn JL, Regev A, Schier AF, Valen E. Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. Development. 2013;140:2828–34.
22. Calviello L, Mukherjee N, Wyler E, Zauber H, Hirsekorn A, Selbach M, et al. Detecting actively translated open reading frames in ribosome profiling data. Nat Methods. 2016;13:165–70.
23. Brar GA, Yassour M, Friedman N, Regev A, Ingolia NT, Weissman JS. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science. 2012;335:552–7.
24. Michel AM, Choudhury KR, Firth AE, Ingolia NT, Atkins JF, Baranov PV. Observation of dually decoded regions of the human genome using ribosome profiling data. Genome Res. 2012;22:2219–29.
25. Crappé J, Van Criekinge W, Trooskens G, Hayakawa E, Luyten W, Baggerman G, et al. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. BMC Genomics. 2013;14:648.
26. Pauli A, Norris ML, Valen E, Chew G-L, Gagnon JA, Zimmerman S, et al. Toddler: an embryonic signal that promotes cell movement via Apelin receptors. Science. 2014;343:1248636.
27. Duncan CDS, Mata J. The translational landscape of fission-yeast meiosis and sporulation. Nat Struct Mol Biol. 2014;21:641–7.
28. Fritsch C, Herrmann A, Nothnagel M, Szafranski K, Huse K, Schumann F, et al. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. Genome Res. 2012;22:2208–18.
29. Lee S, Liu B, Lee S, Huang S-X, Shen B, Qian S-B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2012;109:E2424–32.
30. Nakahigashi K, Takai Y, Shiwa Y, Wada M, Honma M, Yoshikawa H, et al. Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. BMC Genomics. 2014;15:1115.
31. Heyer EE, Moore MJ. Redefining the translational status of 80S monosomes. Cell. 2016;164:757–69.
32. Martens AT, Taylor J, Hilser VJ. Ribosome A and P sites revealed by length analysis of ribosome profiling data. Nucleic Acids Res. 2015;43:3680–7.
33. Hwang J-Y, Buskirk AR. A ribosome profiling study of mRNA cleavage by the endonuclease RelE. Nucleic Acids Res. 2017;45:327–36.
34. Shine J, Dalgarno L. Determinant of cistron specificity in bacterial ribosomes. Nature. 1975;254:34–8.
35. Nakagawa S, Niimura Y, Miura K-I, Gojobori T. Dynamic evolution of translation initiation mechanisms in prokaryotes. Proc Natl Acad Sci U S A. 2010;107:6382–7.
36. Muto A, Osawa S. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci. 1987;84:166–9.
37. Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol. 2013;9:675.
38. Goodman DB, Church GM, Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes. Science. 2013;342:475–9.
39. Del Campo C, Bartholomäus A, Fedyunin I, Ignatova Z. Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. PLoS Genet. 2015;11:e1005613.
40. Li G-W, Burkhardt D, Gross C, Weissman JS. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 2014;157:624–35.
41. Zhou J, Rudd KE. EcoGene 3.0. Nucleic Acids Res. 2012;41:D613–24.
42. Archer SK, Shirokikh NE, Beilharz TH, Preiss T. Dynamics of ribosome scanning and recycling revealed by translation complex profiling. Nature. 2016;535:570–4.
43. Gawron D, Ndah E, Gevaert K, Van Damme P. Positional proteomics reveals differences in N-terminal proteoform stability. Mol Syst Biol. 2016;12:858.
44. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10.
45. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
46. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
47. Balakrishnan R, Oman K, Shoji S, Bundschuh R, Fredrick K. The conserved GTPase LepA contributes mainly to translation initiation in Escherichia coli. Nucleic Acids Res. 2014;42:13370–83.
48. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.
49. H2O.ai. http://h2o.ai/resources. Accessed 10 May 2017.
50. Staes A, Impens F, Van Damme P, Ruttens B, Goethals M, Demol H, et al. Selecting protein N-terminal peptides by combined fractional diagonal chromatography. Nat Protoc. 2011;6:1130–41.
51. Van Damme P, Van Damme J, Demol H, Staes A, Vandekerckhove J, Gevaert K. A review of COFRADIC techniques targeting protein N-terminal acetylation. BMC Proc. 2009;3 Suppl 6:S6.
52. Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, et al. The proteomics of N-terminal methionine cleavage. Mol Cell Proteomics. 2006;5: 2336–49.
53. Bienvenut WV, Giglione C, Meinnel T. Proteome-wide analysis of the amino terminal status of Escherichia coli proteins at the steady-state and upon deformylation inhibition. Proteomics. 2015;15:2503–18.
54. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14:1188–90.
55. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
56. Edgar R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
57. Vizcaíno JA, Csordas A, Del-Toro N, Dianes JA, Griss J, Lavidas I, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 2016; 44:11033.

Supplemental information

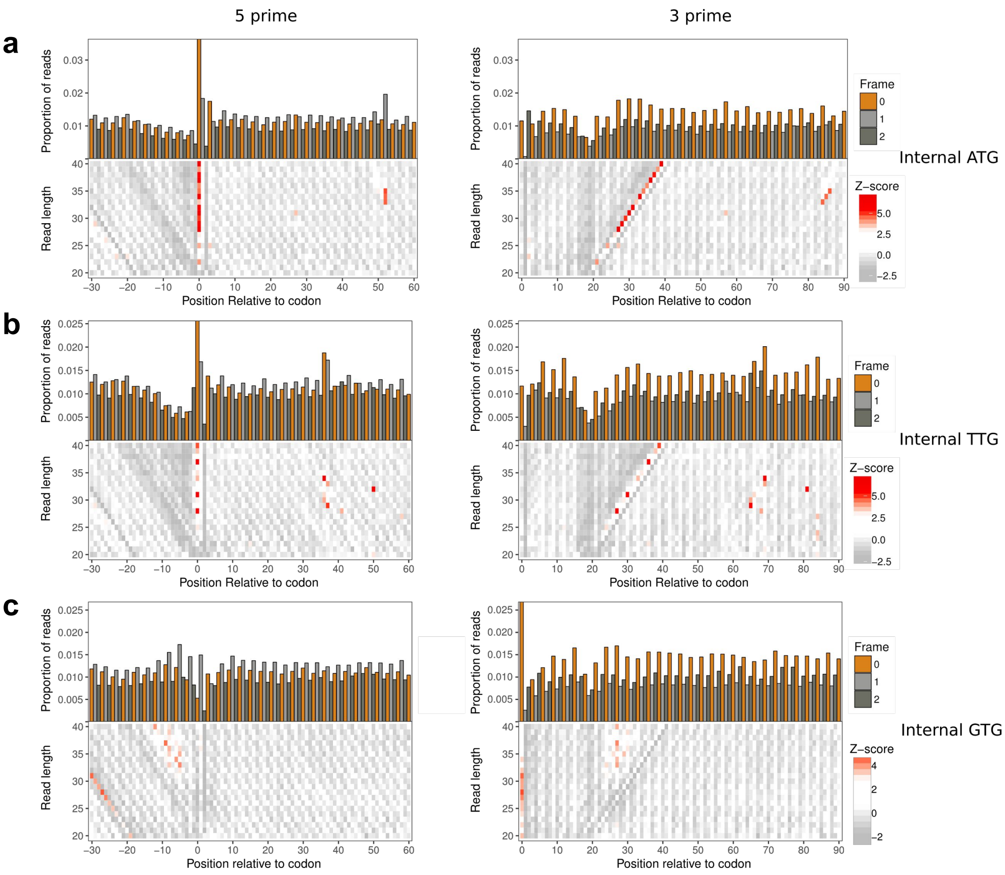


Supplementary_Fig_S1: Ribo-seq meta profiles at start and stop codons - *S.* Typhimurium
Ribo-seq meta-profiles in windows around start codons for annotated genes (n=4205) in the *S.* Typhimurium genome, contributions from each gene are scaled to a sum of one. (**upper**) Proportion of 5 or 3' ribo-seq read counts per nucleotide position, coloured by codon position. (**lower**) Heatmaps of 5' or 3' ribo-seq read counts per length, coloured by z-score per protected read length.



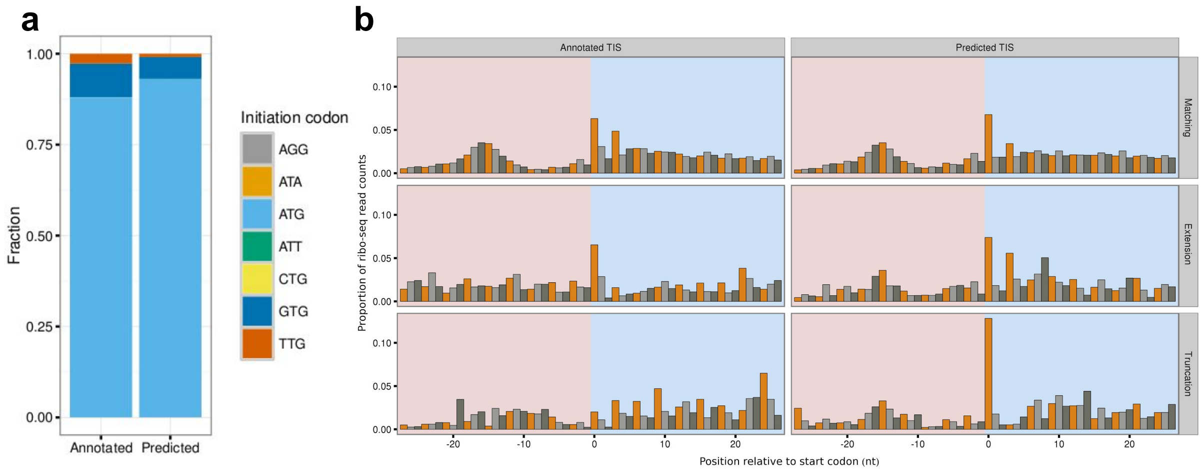
Supplementary_Fig_S2: Read length distributions at Shine Dalgarno motifs

Ribo-seq meta-profiles in windows around shine dalgarno sequences immediately upstream of a TIS (n=736) or at internal CDS positions (n=8564) in the *S.* Typhimurium genome, contributions from each motif window are scaled to a sum of one. Barcharts show the proportion of 5 or 3' ribo-seq read counts per nucleotide position, coloured by codon position. Heatmaps show 5' or 3' ribo-seq read counts per length, coloured by z-score per read length. (**a**) In relation to SD motifs upstream of initiation codons. (**b**) In relation to internal SD motifs within CDS regions. (**c**) In relation to SD motifs upstream of initiation codons, faceted by distance between the SD motif and initiation codon (nt).



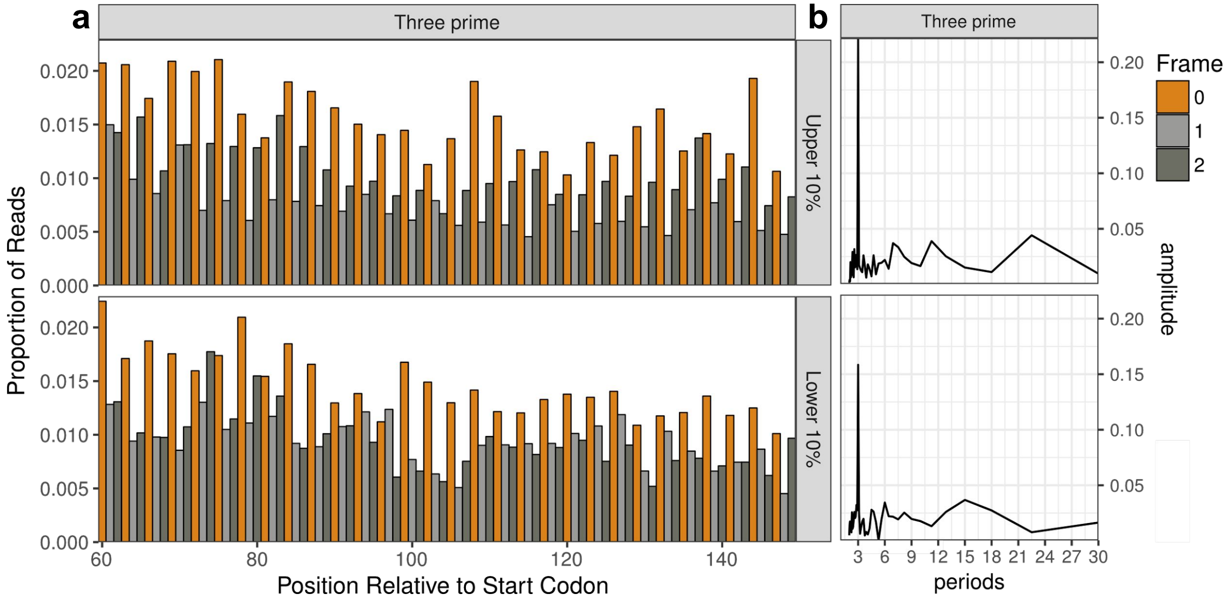
Supplementary_Fig_S3: Codon specific read length distributions.

S. typhimurium ribo-seq meta-profiles in windows around the most commonly used translation initiation codons at internal, in-frame CDS positions. Contributions from each codon window are scaled to a sum of one. (**upper**) Proportion of 5 or 3' ribo-seq read counts per nucleotide position, coloured by codon position. (**lower**) Heatmaps of 5' or 3' ribo-seq read counts per fragment length, coloured by z-score per read length. (**a**) ATG codons (n=35531). (**b**) GTG codons (n=35117). (**c**) TTG codons (n=17647).



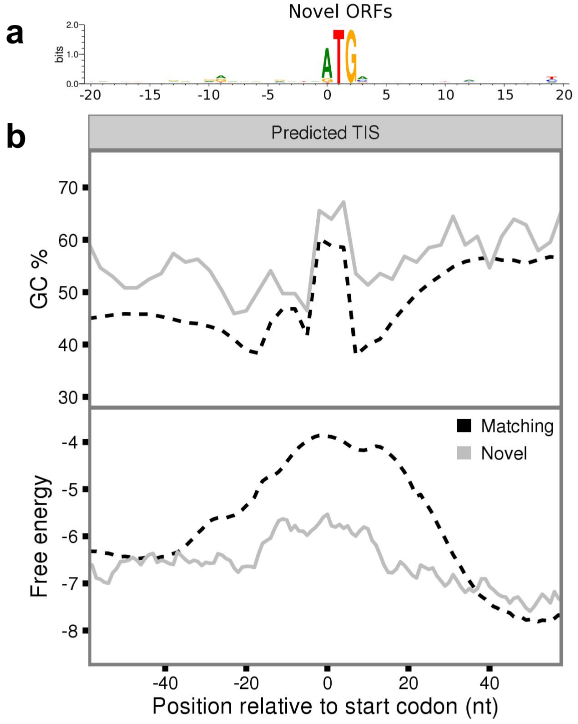
Supplementary_Fig_S4: Additional prediction support

(**a**) Showing the similarity in usage of different start codons between annotated (n=4653) and predicted (n=4334) *S.* Typhimurium ORFs. (**b**). Meta plots showing the proportion of scaled 5' ribo-seq read counts in relation to annotated or predicted translation initiation sites, for ORFs matching annotated genes (n=3853), predicted as extensions (n=214) or predicted as truncations (n=205), in the *S.* Typhimurium dataset. Contributions from each gene are scaled to a sum of one. Nucleotide positions are coloured by codon position. Upstream regions are highlighted in pink, downstream regions are highlighted in light blue.



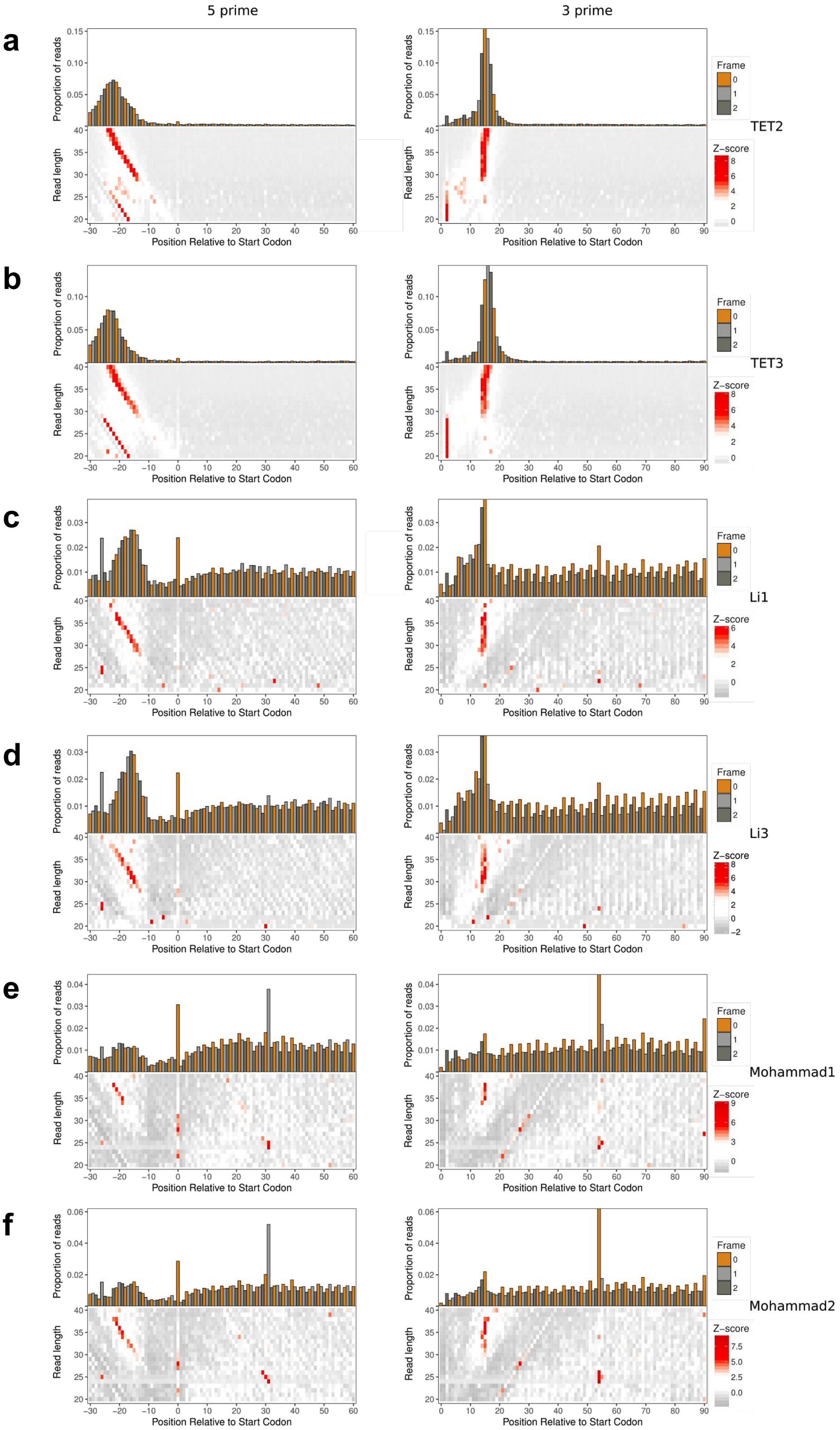
Supplementary_Fig_S5: Third codon periodicity and GC content.

(**a**) Proportion of 3' ribo-seq read counts per nucleotide position, from positions 60 - 150nt downstream of the annotated start codon in the *S.* Typhimurium genome, coloured by codon position. Contributions from each gene are scaled to a sum of one. (**upper**) the highest 10% of regions by third codon GC content (n=467). (**lower**) the bottom 10% of regions by third codon GC content (n=468). (**b**) fourier transform showing the periodicity in the distributions of (**a**).

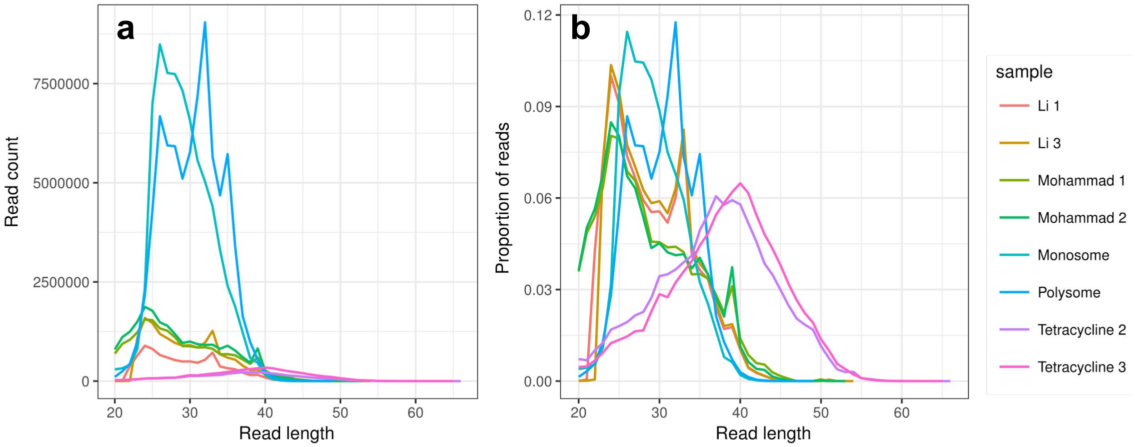


Supplementary_Fig_S6: Evidence for predicted novel translation initiation sites.

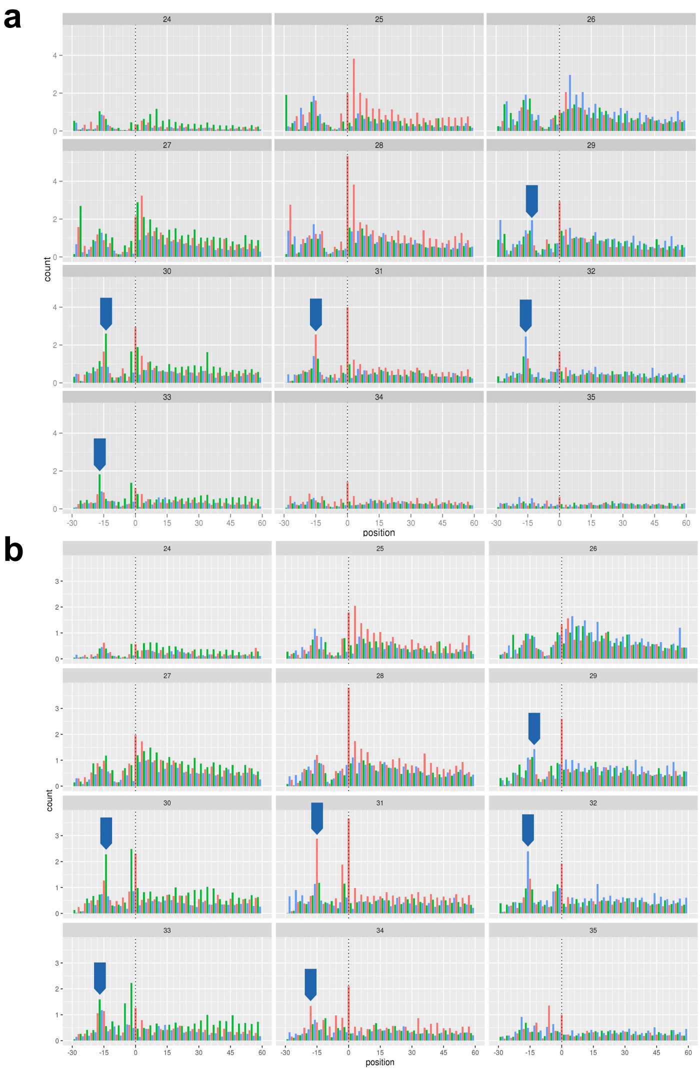
(**a**) Sequence motifs relation to predicted translation initiation sites (n=61), for sequences in all novel predicted ORFs in the *S.* Typhimurium dataset. (**B**) Meta-profiles in relation to annotated (n=3853) or predicted translation initiation sites. Black dotted lines representing ORFs matching annotated genes, grey lines represent novel predictions. (**upper**) Meta-profiles showing the percentage of GC content averaged in 9nt sliding windows, higher values downstream of the codon region are indicative of coding potential. (**lower**) Meta-profiles of free energy averaged in 39nt sliding windows, higher values represent a lower potential for secondary structure formation.



Supplementary_Fig_S7: Ribo-seq meta profiles at start codons - *E. Coli*
Ribo-seq meta-profiles in windows around start codons for annotated genes (n=3726) in the *E. coli* genome, contributions from each gene are scaled to a sum of one. (**upper**) Proportion of 5 or 3' ribo-seq read counts per nucleotide position, coloured by codon position. (**lower**) Heatmaps of 5' or 3' ribo-seq read length counts, coloured by z-score per read length.



Supplementary_Fig_S8: Library read length distributions
Read length distributions of absolute (**a**) or proportional (**b**) counts of aligned ribo-seq reads per library.



Supplementary_Fig_S9: Read lengths adjustments
The sum of scaled 5' ribo-seq counts from the (**a**) monosome or (**b**) polysome replicate, in -30 to +60nt windows around annotated start codons (n=4205) in *S.* Typhimurium, per read length, coloured by codon position. Contributions from each gene are scaled to a sum of one. Blue arrows indicate the peak corresponding to ribo-seq footprints translating the start codon.

Bibliography

[1] FastQC. URL: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. Accessed on 26/11/2018. 1.5.2

[2] FastX-Toolkit. URL: `http://hannonlab.cshl.edu/fastx_toolkit`. Accessed on 26/11/2018. 1.5.2

[3] ACEVEDO, J. M., HOERMANN, B., SCHLIMBACH, T., AND TELEMAN, A. A. Changes in global translation elongation or initiation rates shape the proteome via the Kozak sequence. *Scientific Reports 8*, 1 (Mar. 2018), 337. 5

[4] ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K., AND WALTER, P. *Molecular Biology of the Cell*. Garland Science, 2002. 1.1

[5] ALLAN DRUMMOND, D., AND WILKE, C. O. The evolutionary consequences of erroneous protein synthesis. *Nature Reviews Genetics 10*, 10 (Oct. 2009), 715–724. 1.2.2

[6] AMRANI, N., GANESAN, R., KERVESTIN, S., MANGUS, D. A., GHOSH, S., AND JACOBSON, A. A faux 3′-UTR promotes aberrant termination and triggers nonsense- mediated mRNA decay. *Nature 432*, 7013 (Nov. 2004), 112–118. 1.4.2

[7] ANDERSON, D. M., ANDERSON, K. M., CHANG, C.-L., MAKAREWICH, C. A., NELSON, B. R., MCANALLY, J. R., KASARAGOD, P., SHELTON, J. M., LIOU, J., BASSEL-DUBY, R., AND OLSON, E. N. A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell 160*, 4 (Feb. 2015), 595–606. 1.5.3

[8] ARTHUR, L. L., PAVLOVIC-DJURANOVIC, S., KOUTMOU, K. S., GREEN, R., SZCZESNY, P., AND DJURANOVIC, S. Translational control by lysine-encoding A-rich sequences. *Science Advances 1*, 6 (July 2015), e1500154. 1.4.1, 5

[9] ARTIERI, C. G., AND FRASER, H. B. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Research 24*, 12 (Oct. 2014), gr.175893.114–2021. 1.4.1, 1.5.3, 5

[10] ASPDEN, J. L., EYRE-WALKER, Y. C., PHILLIPS, R. J., AMIN, U., MUMTAZ, M. A. S., BROCARD, M., AND COUSO, J. P. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife 3* (Aug. 2014), 193. 1.5.3

[11] BARTHOLOMÄUS, A., DEL CAMPO, C., AND IGNATOVA, Z. Mapping the non-standardized biases of ribosome profiling. *Biological Chemistry 397*, 1 (2016), 23–35. 5

[12] BAZZINI, A. A., JOHNSTONE, T. G., CHRISTIANO, R., MACKOWIAK, S. D., OBERMAYER, B., FLEMING, E. S., VEJNAR, C. E., LEE, M. T., RAJEWSKY, N., WALTHER, T. C., AND GIRALDEZ, A. J. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO Journal 33*, 9 (May 2014), 981–993. 1.5.2, 1.5.3

[13] BEAUDOIN, J.-D., NOVOA, E. M., VEJNAR, C. E., YARTSEVA, V., TAKACS, C. M., KELLIS, M., AND GIRALDEZ, A. J. Analyses of mRNA structure dynamics identify embryonic gene regulatory programs. *Nature Structural & Molecular Biology 25*, 8 (Aug. 2018), 677–686. 1.4.1

[14] BENGTSON, M. H., AND JOAZEIRO, C. A. P. Role of a ribosome-associated E3 ubiquitin ligase in protein quality control. *Nature 467*, 7314 (Sept. 2010), 470–473. 1.4.2

[15] BHAT, M., ROBICHAUD, N., HULEA, L., SONENBERG, N., PELLETIER, J., AND TOPISIROVIC, I. Targeting the translation machinery in cancer. *Nature Reviews Drug Discovery 14*, 4 (Mar. 2015), 261–278. 1.5.3

[16] BISWAS, J., LIU, Y., SINGER, R. H., AND WU, B. Fluorescence Imaging Methods to Investigate Translation in Single Cells. *Cold Spring Harbor Perspectives in Biology* (Aug. 2018), a032722. 1.3

[17] BOLGER, A. M., LOHSE, M., AND USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics 30*, 15 (Apr. 2014), 2114–2120. 1.5.2

[18] BRAAT, A. K., YAN, N., ARN, E., HARRISON, D., AND MACDONALD, P. M. Localization-Dependent Oskar Protein Accumulation. Control after the Initiation of Translation. *Developmental Cell 7*, 1 (July 2004), 125–131. 1.4.3

[19] BRANDMAN, O., AND HEGDE, R. S. Ribosome-associated protein quality control. *Nature Structural &#38; Molecular Biology 23*, 1 (Jan. 2016), 7–15. 1.4.2

[20] BREAKER, R. R. Riboswitches and Translation Control. *Cold Spring Harbor Perspectives in Biology 10*, 11 (Nov. 2018), a032797. 1.3

[21] BRULE, C. E., AND GRAYHACK, E. J. Synonymous Codons: Choose Wisely for Expression. *Trends in Genetics 33*, 4 (Apr. 2017), 283–297. 1.4.1

[22] BUCKLEY, P. T., KHALADKAR, M., KIM, J., AND EBERWINE, J. Cytoplasmic intron retention, function, splicing, and the sentinel RNA hypothesis. *Wiley interdisciplinary reviews. RNA 5*, 2 (Mar. 2014), 223–230. 5

[23] BUSKIRK, A. R., AND GREEN, R. Ribosome pausing, arrest and rescue in bacteria and eukaryotes. *Phil. Trans. R. Soc. B 372*, 1716 (Jan. 2017), 20160183. 1.2.2

[24] Caliskan, N., Katunin, V. I., Belardinelli, R., Peske, F., and Rodnina, M. V. Programmed –1 Frameshifting by Kinetic Partitioning during Impeded Translocation. *Cell 157*, 7 (June 2014), 1619–1631. 1.4.1

[25] Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. Detecting actively translated open reading frames in ribosome profiling data. *Nature Methods 13*, 2 (Feb. 2016), 165–170. 1.5.3

[26] Chaney, J. L., and Clark, P. L. Roles for Synonymous Codon Usage in Protein Biogenesis. *Annual Review of Biophysics 44*, 1 (June 2015), 143–166. 1.4.1, 1.4.2

[27] Chaney, J. L., Steele, A., Carmichael, R., Rodriguez, A., Specht, A. T., Ngo, K., Li, J., Emrich, S., and Clark, P. L. Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLoS Computational Biology 13*, 5 (May 2017), e1005531. 1.4.2

[28] Chang, B., Halgamuge, S., and Tang, S.-L. Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes. *Gene 373* (May 2006), 90–99. 1.2.1

[29] Chang, Y.-F., Imam, J. S., and Wilkinson, M. F. The Nonsense-Mediated Decay RNA Surveillance Pathway. *dx.doi.org 76*, 1 (June 2007), 51–74. 1.4.2

[30] Charneski, C. A., and Hurst, L. D. Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLoS Biology 11*, 3 (Mar. 2013), e1001508. 1.4.1, 1.5.3, 5

[31] Chartron, J. W., Hunt, K. C. L., and Frydman, J. Cotranslational signal-independent SRP preloading during membrane targeting. *Nature 536*, 7615 (Aug. 2016), 224–228. 1.4.2

[32] Chen, J., Petrov, A., Johansson, M., Tsai, A., O'Leary, S. E., and Puglisi, J. D. Dynamic pathways of −1 translational frameshifting. *Nature 512*, 7514 (Aug. 2014), 328–332. 1.4.1

[33] Cheng, Z., Otto, G. M., Powers, E. N., Keskin, A., Mertins, P., Carr, S. A., Jovanovic, M., and Brar, G. A. Pervasive, Coordinated Protein-Level Changes Driven by Transcript Isoform Switching during Meiosis. *Cell 172*, 5 (Feb. 2018), 910–923.e16. 1.5.3

[34] Choe, Y.-J., Park, S.-H., Hassemer, T., Körner, R., Vincenz-Donnelly, L., Hayer-Hartl, M., and Hartl, F. U. Failure of RQC machinery causes protein aggregation and proteotoxic stress. *Nature 531*, 7593 (Mar. 2016), 191–195. 1.4.3

[35] Chu, J., Hong, N. A., Masuda, C. A., Jenkins, B. V., Nelms, K. A., Goodnow, C. C., Glynne, R. J., Wu, H., Masliah, E., Joazeiro, C. A. P., and Kay, S. A. A mouse forward genetics screen identifies LISTERIN

as an E3 ubiquitin ligase involved in neurodegeneration. *Proceedings of the National Academy of Sciences 106*, 7 (Feb. 2009), 2097–2103. 1.4.3

[36] CHU, J., AND PELLETIER, J. Therapeutic Opportunities in Eukaryotic Translation. *Cold Spring Harbor Perspectives in Biology 10*, 6 (June 2018), a032995. 1.5.3

[37] CLARK, I. E., WYCKOFF, D., AND GAVIS, E. R. Synthesis of the posterior determinant Nanos is spatially restricted by a novel cotranslational regulatory mechanism. *Current biology : CB 10*, 20 (Oct. 2000), 1311–1314. 1.4.3

[38] CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M. W., GAFFNEY, D. J., ELO, L. L., ZHANG, X., AND MORTAZAVI, A. A survey of best practices for RNA-seq data analysis. *Genome Biology 17*, 1 (Dec. 2016), 13. 1.5.2

[39] CRICK, F. H. C. On protein synthesis. *Cold Spring Harbor Laboratory Archives SB/11/5/4* (1957). 1

[40] CYMER, F., AND VON HEIJNE, G. Cotranslational folding of membrane proteins probed by arrest-peptide-mediated force measurements. *Proceedings of the National Academy of Sciences 110*, 36 (Sept. 2013), 14640–14645. 1.4.2

[41] DANA, A., AND TULLER, T. Determinants of Translation Elongation Speed and Ribosomal Profiling Biases in Mouse Embryonic Stem Cells. *PLoS Computational Biology 8*, 11 (Nov. 2012), e1002755. 1.5.3

[42] DANA, A., AND TULLER, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Research 42*, 14 (Aug. 2014), 9171–9181. 1.5.2

[43] DANA, A., AND TULLER, T. Mean of the Typical Decoding Rates: A New Translation Efficiency Index Based on the Analysis of Ribosome Profiling Data. *G3: Genes|Genomes|Genetics 5*, 1 (Jan. 2015), 73–80. 1.5.2

[44] DEVER, T. E., DINMAN, J. D., AND GREEN, R. Translation Elongation and Recoding in Eukaryotes. *Cold Spring Harbor Perspectives in Biology 10*, 8 (Aug. 2018), a032649. 1.2.2, 1.2.2, 1.3

[45] D'LIMA, N. G., MA, J., WINKLER, L., CHU, Q., LOH, K. H., CORPUZ, E. O., BUDNIK, B. A., LYKKE-ANDERSEN, J., SAGHATELIAN, A., AND SLAVOFF, S. A. A human microprotein that interacts with the mRNA decapping complex. *Nature Publishing Group 13*, 2 (Feb. 2017), 174. 1.5.3

[46] DOBIN, A., DAVIS, C. A., SCHLESINGER, F., DRENKOW, J., ZALESKI, C., JHA, S., BATUT, P., CHAISSON, M., AND GINGERAS, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics 29*, 1 (Oct. 2012), 15–21. 1.5.2

[47] DÖRING, K., AHMED, N., RIEMER, T., SURESH, H. G., VAINSHTEIN, Y., HABICH, M., RIEMER, J., MAYER, M. P., O'BRIEN, E. P., KRAMER, G., AND BUKAU, B. Profiling Ssb-Nascent Chain Interactions Reveals Principles of Hsp70-Assisted Folding. *Cell 170*, 2 (July 2017), 298–311.e20. 1.5.3

[48] Duchaine, T. F., and Fabian, M. R. Mechanistic Insights into MicroRNA-Mediated Gene Silencing. *Cold Spring Harbor Perspectives in Biology* (June 2018), a032771. 1.3, 1.5.3

[49] Duncan, C. D. S., and Mata, J. Effects of cycloheximide on the interpretation of ribosome profiling experiments in Schizosaccharomyces pombe. *Scientific Reports 7*, 1 (Sept. 2017), 10331. 5

[50] Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R., Weissman, J. S., and Sonenberg, N. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife 2* (Dec. 2013), e01179. 1.5.2, 1.5.3

[51] Duval, M., Korepanov, A., Fuchsbauer, O., Fechter, P., Haller, A., Fabbretti, A., Choulier, L., Micura, R., Klaholz, B. P., Romby, P., Springer, M., and Marzi, S. Escherichia coli Ribosomal Protein S1 Unfolds Structured mRNAs Onto the Ribosome for Active Translation Initiation. *PLoS Biology 11*, 12 (Dec. 2013), e1001731. 1.2.1

[52] Eichhorn, S. W., Guo, H., McGeary, S. E., Rodriguez-Mias, R. A., Shin, C., Baek, D., Hsu, S.-h., Ghoshal, K., Villén, J., and Bartel, D. P. mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Molecular cell 56*, 1 (Oct. 2014), 104–115. 1.5.3

[53] Fields, A. P., Rodriguez, E. H., Jovanovic, M., Stern-Ginossar, N., Haas, B. J., Mertins, P., Raychowdhury, R., Hacohen, N., Carr, S. A., Ingolia, N., Regev, A., and Weissman, J. S. A Regression-Based Analysis of Ribosome-Profiling Data Reveals a Conserved Complexity to Mammalian Translation. *Molecular cell 60*, 5 (Mar. 2015), 816–827. 1.5.3

[54] Fresno, M., Jiménez, A., and Vázquez, D. Inhibition of translation in eukaryotic systems by harringtonine. *European journal of biochemistry 72*, 2 (Jan. 1977), 323–330. 1.5.3

[55] Frischmeyer, P. A., and Dietz, H. C. Nonsense-Mediated mRNA Decay in Health and Disease. *Human Molecular Genetics 8*, 10 (Jan. 1999), 1893–1900. 1.4.2

[56] Fritsch, C., Herrmann, A., Nothnagel, M., Szafranski, K., Huse, K., Schumann, F., Schreiber, S., Platzer, M., Krawczak, M., Hampe, J., and Brosch, M. Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome Research 22*, 11 (Nov. 2012), 2208–2218. 1.5.3

[57] Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife 3* (Oct. 2014), 198. 1.4.1

[58] GAWROŃSKI, P., JENSEN, P. E., KARPIŃSKI, S., LEISTER, D., AND SCHARFF, L. B. Pausing of Chloroplast Ribosomes Is Induced by Multiple Features and Is Linked to the Assembly of Photosynthetic Complexes. *Plant Physiology 176*, 3 (Mar. 2018), 2557–2569. 5

[59] GELLER, R., PECHMANN, S., ACEVEDO, A., ANDINO, R., AND FRYDMAN, J. Hsp90 shapes protein and RNA evolution to balance trade-offs between protein stability and aggregation. *Nature Communications 9*, 1 (May 2018), 1781. 1.4.2

[60] GERASHCHENKO, M. V., AND GLADYSHEV, V. N. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Research 42*, 17 (July 2014), gku671–e134. 5

[61] GERASHCHENKO, M. V., LOBANOV, A. V., AND GLADYSHEV, V. N. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proceedings of the National Academy of Sciences 109*, 43 (Oct. 2012), 17394–17399. 1.5.3

[62] GINGOLD, H., TEHLER, D., CHRISTOFFERSEN, N. R., NIELSEN, M. M., ASMAR, F., KOOISTRA, S. M., CHRISTOPHERSEN, N. S., CHRISTENSEN, L. L., BORRE, M., SØRENSEN, K. D., ANDERSEN, L. D., ANDERSEN, C. L., HULLEMAN, E., WURDINGER, T., RALFKIÆR, E., HELIN, K., GRØNBÆK, K., ØRNTOFT, T., WASZAK, S. M., DAHAN, O., PEDERSEN, J. S., LUND, A. H., AND PILPEL, Y. A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. *Cell 158*, 6 (Sept. 2014), 1281–1292. 1.4.1

[63] GODCHAUX, W., ADAMSON, S. D., AND HERBERT, E. Effects of cycloheximide on polyribosome function in reticulocytes. *Journal of Molecular Biology 27*, 1 (July 1967), 57–72. 1.5.3

[64] GOLDMAN, D. H., KAISER, C. M., MILIN, A., RIGHINI, M., TINOCO, I., AND BUSTAMANTE, C. Mechanical force releases nascent chain-mediated ribosome arrest in vitro and in vivo. *Science 348*, 6233 (Apr. 2015), 457–460. 1.4.2

[65] GOODARZI, H., NGUYEN, H. C. B., ZHANG, S., DILL, B. D., MOLINA, H., AND TAVAZOIE, S. F. Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression. *Cell 165*, 6 (June 2016), 1416–1427. 1.4.1

[66] GOROCHOWSKI, T. E., IGNATOVA, Z., BOVENBERG, R. A. L., AND ROUBOS, J. A. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Research 43*, 6 (Mar. 2015), gkv199–3032. 1.4.1

[67] GRABER, T. E., HÉBERT-SEROPIAN, S., KHOUTORSKY, A., DAVID, A., YEWDELL, J. W., LACAILLE, J.-C., AND SOSSIN, W. S. Reactivation of stalled polyribosomes in synaptic plasticity. *Proceedings of the National Academy of Sciences 110*, 40 (Oct. 2013), 16205–16210. 1.4.3

[68] GUALERZI, C. O., AND PON, C. L. Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cellular and Molecular Life Sciences 72*, 22 (Aug. 2015), 4341–4367. 1.2.1

[69] GUO, H., INGOLIA, N., WEISSMAN, J. S., AND BARTEL, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature 466*, 7308 (Aug. 2010), 835–840. 1.5.3

[70] GUTIERREZ, E., SHIN, B.-S., WOOLSTENHULME, C. J., KIM, J.-R., SAINI, P., BUSKIRK, A. R., AND DEVER, T. E. eIF5A Promotes Translation of Polyproline Motifs. *Molecular cell 51*, 1 (July 2013), 35–45. 1.4.1

[71] GUYDOSH, N. R., AND GREEN, R. Dom34 Rescues Ribosomes in 3´ Untranslated Regions. *Cell 156*, 5 (Feb. 2014), 950–962. 1.5.2, 5

[72] HAN, Y., GAO, X., LIU, B., WAN, J., ZHANG, X., AND QIAN, S.-B. Ribosome profiling reveals sequence-independent post-initiation pausing as a signature of translation. *Cell Research 24*, 7 (July 2014), 842–851. 1.5.3

[73] HE, F., LI, X., SPATRICK, P., CASILLO, R., DONG, S., AND JACOBSON, A. Genome-Wide Analysis of mRNAs Regulated by the Nonsense-Mediated and 5´ to 3´ mRNA Decay Pathways in Yeast. *Molecular cell 12*, 6 (Dec. 2003), 1439–1452. 1.4.2

[74] HECK, A. M., AND WILUSZ, J. The Interplay between the RNA Decay and Translation Machinery in Eukaryotes. *Cold Spring Harbor Perspectives in Biology 10*, 5 (May 2018), a032839. 1.3, 1.4, 1.3, 1.4.2

[75] HELLEN, C. U. T. Translation Termination and Ribosome Recycling in Eukaryotes. *Cold Spring Harbor Perspectives in Biology 10*, 10 (May 2018), a032656. 1.5.3

[76] HERSCH, S. J., WANG, M., ZOU, S. B., MOON, K.-M., FOSTER, L. J., IBBA, M., NAVARRE, W. W., AND GOTTESMAN, S. Divergent Protein Motifs Direct Elongation Factor P-Mediated Translational Regulation in Salmonella enterica and Escherichia coli. *mBio 4*, 2 (May 2013), e00180–13–13. 1.2.2

[77] HERSHEY, J. W. B., SONENBERG, N., AND MATHEWS, M. B. Principles of translational control. *Cold Spring Harbor Perspectives in Biology* (June 2018), a032607. 1, 1.3, 1.3

[78] HINNEBUSCH, A. G. Structural Insights into the Mechanism of Scanning and Start Codon Recognition in Eukaryotic Translation Initiation. *Trends in Biochemical Sciences 42*, 8 (Aug. 2017), 589–611. 1.2.1

[79] HSIEH, A. C., LIU, Y., EDLIND, M. P., INGOLIA, N., JANES, M. R., SHER, A., SHI, E. Y., STUMPF, C. R., CHRISTENSEN, C., BONHAM, M. J., WANG, S., REN, P., MARTIN, M., JESSEN, K., FELDMAN, M. E., WEISSMAN, J. S., SHOKAT, K. M., ROMMEL, C., AND RUGGERO, D. The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature 485*, 7396 (May 2012), 55–61. 1.5.3

[80] HUSSMANN, J. A., PATCHETT, S., JOHNSON, A., SAWYER, S., AND PRESS, W. H. Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genetics 11*, 12 (Dec. 2015), e1005732. 5

[81] IKEMURA, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular Biology and Evolution 2*, 1 (Jan. 1985), 13–34. 1.4.1

[82] INGOLIA, N., BRAR, G. A., ROUSKIN, S., MCGEACHY, A. M., AND WEISSMAN, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature Protocols 7*, 8 (July 2012), 1534–1550. 1.5.1, 1.5.2

[83] INGOLIA, N., BRAR, G. A., STERN-GINOSSAR, N., HARRIS, M. S., TALHOUARNE, G. J. S., JACKSON, S. E., WILLS, M. R., AND WEISSMAN, J. S. Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *CellReports 8*, 5 (Sept. 2014), 1365–1379. 1.5.2

[84] INGOLIA, N., GHAEMMAGHAMI, S., NEWMAN, J. R., AND WEISSMAN, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* (Apr. 2009), 1–6. 1, 1.5, 1.5.2, 1.5.2, 1.5.3, 1.5.3

[85] INGOLIA, N., HUSSMANN, J. A., AND WEISSMAN, J. S. Ribosome Profiling: Global Views of Translation. *Cold Spring Harbor Perspectives in Biology* (July 2018), a032698. 1.3, 1.5, 1.6, 1.7, 1.8, 1.9

[86] INGOLIA, N., LAREAU, L. F., AND WEISSMAN, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell 147*, 4 (Nov. 2011), 1–14. 1.3, 1.5.2, 1.5.3, 1.5.3

[87] INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. Initial sequencing and analysis of the human genome. *Nature 409*, 6822 (Feb. 2001), 860–921. 1

[88] ISHIGAME, H., MOSAHEB, M. M., SANJABI, S., AND FLAVELL, R. A. Truncated Form of TGF-betaRII, But Not Its Absence, Induces Memory CD8+ T Cell Expansion and Lymphoproliferative Disorder in Mice. *The Journal of Immunology 190*, 12 (June 2013), 6340–6350. 1.4.2

[89] ISHIKAWA, K., MAKANAE, K., IWASAKI, S., INGOLIA, N., AND MORIYA, H. Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes. *PLoS Genetics 13*, 1 (Jan. 2017), e1006554. 1.5.3

[90] ISHIMURA, R., NAGY, G., DOTU, I., ZHOU, H., YANG, X. L., SCHIMMEL, P., SENJU, S., NISHIMURA, Y., CHUANG, J. H., AND ACKERMAN, S. L. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science 345*, 6195 (July 2014), 455–459. 1.4.3

[91] ITO-HARASHIMA, S., KUROHA, K., TATEMATSU, T., AND INADA, T. Translation of the poly(A) tail plays crucial roles in nonstop mRNA surveillance via translation repression and protein destabilization by proteasome in yeast. *Genes & Development 21*, 5 (Mar. 2007), 519–524. 1.4.2

[92] IVANOV, P., KEDERSHA, N., AND ANDERSON, P. Stress Granules and Processing Bodies in Translational Control. *Cold Spring Harbor Perspectives in Biology* (Aug. 2018), a032813. 1.3

[93] IWAKAWA, H.-O., AND TOMARI, Y. The Functions of MicroRNAs: mRNA Decay and Translational Repression. *Trends in Cell Biology 25*, 11 (Nov. 2015), 651–665. 1.3

[94] JOAZEIRO, C. A. P. Ribosomal Stalling During Translation: Providing Substrates for Ribosome-Associated Protein Quality Control. *doi.org 33*, 1 (Oct. 2017), 343–368. 1.4.2

[95] JUTZI, D., AKINYI, M. V., MECHTERSHEIMER, J., FRILANDER, M. J., AND RUEPP, M.-D. The emerging role of minor intron splicing in neurological disorders. *Cell Stress 2*, 3 (Mar. 2018), 40–54. 5

[96] KAROUSIS, E. D., AND MÜHLEMANN, O. Nonsense-Mediated mRNA Decay Begins Where Translation Ends. *Cold Spring Harbor Perspectives in Biology* (June 2018), a032862. 1.3, 1.3

[97] KEARSE, M. G., AND WILUSZ, J. E. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes & Development 31*, 17 (Oct. 2017), 1717–1731. 1.2.1

[98] KIM, D., PERTEA, G., TRAPNELL, C., PIMENTEL, H., KELLEY, R., AND SALZBERG, S. L. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology 14*, 4 (Apr. 2013), R36. 1.5.2

[99] KIM, H.-K., LIU, F., FEI, J., BUSTAMANTE, C., GONZALEZ, R. L., AND TINOCO, I. A frameshifting stimulatory stem loop destabilizes the hybrid state and impedes ribosomal translocation. *Proceedings of the National Academy of Sciences 111*, 15 (Apr. 2014), 5538–5543. 1.4.1

[100] KOUTMOU, K. S., SCHULLER, A. P., BRUNELLE, J. L., RADHAKRISHNAN, A., DJURANOVIC, S., AND GREEN, R. Ribosomes slide on lysine-encoding homopolymeric A stretches. *eLife 4* (Feb. 2015), 446. 1.4.1, 5

[101] LANGMEAD, B., AND SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods 9*, 4 (Apr. 2012), 357–359. 1.5.2

[102] LAREAU, L. F., HITE, D. H., HOGAN, G. J., AND BROWN, P. O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife 3* (2014), e01257. 1.5.2, 1.5.3, 5

[103] LEE, S., LIU, B., LEE, S., HUANG, S.-X., SHEN, B., AND QIAN, S.-B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences 109*, 37 (Sept. 2012), 14728–14729. 1.5.3

[104] LEE, T. I., AND YOUNG, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell 152*, 6 (Mar. 2013), 1237–1251. 1

[105] LEE, Y., ZHOU, T., TARTAGLIA, G. G., VENDRUSCOLO, M., AND WILKE, C. O. Translationally optimal codons associate with aggregation-prone sites in proteins. *PROTEOMICS 10*, 23 (Dec. 2010), 4163–4171. 1.4.2

[106] LI, G.-W., BURKHARDT, D., GROSS, C., AND WEISSMAN, J. S. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. *Cell 157*, 3 (Apr. 2014), 624–635. 1.5.3

[107] LI, G.-W., OH, E., AND WEISSMAN, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature 484*, 7395 (Apr. 2012), 538–541. 1.5.2

[108] LI, H., AND DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics 25*, 14 (July 2009), 1754–1760. 1.5.2

[109] LIU, B., HAN, Y., AND QIAN, S.-B. Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. *Molecular cell 49*, 3 (Feb. 2013), 453–463. 1.5.3

[110] LIU, T.-Y., HUANG, H. H., WHEELER, D., XU, Y., WELLS, J. A., SONG, Y. S., AND WIITA, A. P. Time-Resolved Proteomics Extends Ribosome Profiling-Based Measurements of Protein Synthesis Dynamics. *Cell Systems 4*, 6 (June 2017), 636–644.e9. 1.5.3

[111] LU, J., AND DEUTSCH, C. Electrostatics in the Ribosomal Tunnel Modulate Chain Elongation Rates. *Journal of Molecular Biology 384*, 1 (Dec. 2008), 73–86. 1.4.1

[112] MANDAL, A., MANDAL, S., AND PARK, M. H. Genome-Wide Analyses and Functional Classification of Proline Repeat-Rich Proteins: Potential Role of eIF5A in Eukaryotic Evolution. *PLoS ONE 9*, 11 (Nov. 2014), e111800. 1.4.1, 5

[113] MARINO, J., VON HEIJNE, G., AND BECKMANN, R. Small protein domains fold inside the ribosome exit tunnel. *FEBS Letters 590*, 5 (Mar. 2016), 655–660. 1.4.2

[114] MAYR, C. Evolution and Biological Roles of Alternative 3´UTRs. *Trends in Cell Biology 26*, 3 (Mar. 2016), 227–237. 1.3

[115] MENDELL, J. T., SHARIFI, N. A., MEYERS, J. L., MARTINEZ-MURILLO, F., AND DIETZ, H. C. Nonsense surveillance regulates expression of diverse classes of mammalian transcripts and mutes genomic noise. *Nature Genetics 36*, 10 (Oct. 2004), 1073–1078. 1.4.2

[116] Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappé, J., Gevaert, K., and Van Damme, P. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & Cellular Proteomics 12*, 7 (July 2013), 1780–1790. 1.5.3

[117] Michel, A. M., Choudhury, K. R., Firth, A. E., Ingolia, N., Atkins, J. F., and Baranov, P. V. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Research 22*, 11 (Nov. 2012), 2219–2229. 1.5.2, 1.5.2, 1.5.3

[118] Miettinen, T. P., and Björklund, M. Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions. *Nucleic Acids Research* (Dec. 2014), gku1310. 1.5.2

[119] Mitchell, P., and Tollervey, D. An NMD Pathway in Yeast Involving Accelerated Deadenylation and Exosome-Mediated 3′-5′ Degradation. *Molecular cell 11*, 5 (May 2003), 1405–1413. 1.4.2

[120] Mohammad, F., Woolstenhulme, C. J., Green, R., and Buskirk, A. R. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Reports 14*, 4 (Feb. 2016), 686–694. 1.5.3, 5, 5, 5

[121] Muhlrad, D., and Parker, R. Premature translational termination triggers mRNA decapping. *Nature 370*, 6490 (Aug. 1994), 578–581. 1.4.2

[122] Muhlrad, D., and Parker, R. Aberrant mRNAs with extended 3' UTRs are substrates for rapid degradation by mRNA surveillance. *RNA 5*, 10 (Oct. 1999), 1299–1307. 1.4.2

[123] Muto, H., and Ito, K. Peptidyl-prolyl-tRNA at the ribosomal P-site reacts poorly with puromycin. *Biochemical and Biophysical Research Communications 366*, 4 (Feb. 2008), 1043–1047. 1.4.1

[124] Nilsson, O. B., Hedman, R., Marino, J., Wickles, S., Bischoff, L., Johansson, M., Müller-Lucks, A., Trovato, F., Puglisi, J. D., O'Brien, E. P., Beckmann, R., and von Heijne, G. Cotranslational Protein Folding inside the Ribosome Exit Tunnel. *Cell Reports 12*, 10 (Sept. 2015), 1533–1540. 1.4.2

[125] Oh, E., Becker, A. H., Sandikci, A., Huber, D., Chaba, R., Gloge, F., Nichols, R. J., Typas, A., Gross, C. A., Kramer, G., Weissman, J. S., and Bukau, B. Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor In Vivo. *Cell 147*, 6 (Dec. 2011), 1295–1308. 1.5.3, 1.5.3

[126] Pauli, A., Norris, M. L., Valen, E., Chew, G. L., Gagnon, J. A., Zimmerman, S., Mitchell, A., Ma, J., Dubrulle, J., Reyon, D., Tsai, S. Q., Joung, J. K., Saghatelian, A., and Schier, A. F. Toddler: An

Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science 343*, 6172 (Feb. 2014), 1248636–1248636. 1.5.3

[127] PAVLOV, M. Y., WATTS, R. E., TAN, Z., CORNISH, V. W., EHRENBERG, M., AND FORSTER, A. C. Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proceedings of the National Academy of Sciences 106*, 1 (Jan. 2009), 50–54. 1.4.1

[128] PECHMANN, S., CHARTRON, J. W., AND FRYDMAN, J. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP *in vivo*. *Nature Structural & Molecular Biology 21*, 12 (Dec. 2014), 1100–1105. 1.4.2

[129] PECHMANN, S., AND FRYDMAN, J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology 20*, 2 (Feb. 2013), 237–243. 1.4.2

[130] PECHMANN, S., WILLMUND, F., AND FRYDMAN, J. The Ribosome as a Hub for Protein Quality Control. *Molecular cell 49*, 3 (Feb. 2013), 411–421. 1.4.2

[131] PEER, E., MOSHITCH-MOSHKOVITZ, S., RECHAVI, G., AND DOMINISSINI, D. The Epitranscriptome in Translation Regulation. *Cold Spring Harbor Perspectives in Biology* (July 2018), a032623. 1.5.3

[132] PEIL, L., STAROSTA, A. L., LASSAK, J., ATKINSON, G. C., VIRUMÄE, K., SPITZER, M., TENSON, T., JUNG, K., REMME, J., AND WILSON, D. N. Distinct XPPX sequence motifs induce ribosome stalling, which is rescued by the translation elongation factor EF-P. *Proceedings of the National Academy of Sciences 110*, 38 (Sept. 2013), 15265–15270. 1.2.2, 5

[133] PLOTKIN, J. B., AND KUDLA, G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Publishing Group 12*, 1 (Nov. 2010), 32–42. 1.2.2, 1.3

[134] POLJŠAK, B., AND MILISAV, I. Clinical implications of cellular stress responses. *Bosnian Journal of Basic Medical Sciences 12*, 2 (Sept. 2017), 122. 1.5.3

[135] POP, C., ROUSKIN, S., INGOLIA, N., HAN, L., PHIZICKY, E. M., WEISSMAN, J. S., AND KOLLER, D. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Molecular Systems Biology 10*, 12 (Dec. 2014), 770–770. 1.2.2, 1.3, 1.4.1, 1.5.3

[136] PROUD, C. G. Phosphorylation and Signal Transduction Pathways in Translational Control. *Cold Spring Harbor Perspectives in Biology* (June 2018), a033050. 1.3

[137] QIAN, W., YANG, J.-R., PEARSON, N. M., MACLEAN, C., AND ZHANG, J. Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. *PLoS Genetics 8*, 3 (Mar. 2012), e1002603. 1.4.1, 1.5.3

[138] REQUIÃO, R. D., DE SOUZA, H. J. A., ROSSETTO, S., DOMITROVIC, T., AND PALHANO, F. L. Increased ribosome density associated to positively charged residues is evident in ribosome profiling experiments performed in the absence of translation inhibitors. *RNA Biology 13*, 6 (May 2016), 561–568. 1.4.1, 5

[139] RICHTER, J. D., AND COLLER, J. Pausing on Polyribosomes: Make Way for Elongation in Translational Control. *Cell 163*, 2 (Aug. 2015), 292–300. 1.4, 1.4.3

[140] ROBICHAUD, N., SONENBERG, N., RUGGERO, D., AND SCHNEIDER, R. J. Translational Control in Cancer. *Cold Spring Harbor Perspectives in Biology* (June 2018), a032896. 1.3, 1.3.1, 1.5.3

[141] RODNINA, M. V. The ribosome in action: Tuning of translational efficiency and protein folding. *Protein Science 25*, 8 (June 2016), 1390–1406. 1.4, 1.4.1, 1.4.1, 1.4.1, 1.4.2, 1.4.2

[142] RODNINA, M. V. Translation in Prokaryotes. *Cold Spring Harbor Perspectives in Biology 10*, 9 (Apr. 2018), a032664. 1.1, 1.2.1, 1.2.2, 1.2.2, 1.2.3, 1.2.4

[143] ROUSKIN, S., ZUBRADT, M., WASHIETL, S., KELLIS, M., AND WEISSMAN, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature 505*, 7485 (Dec. 2013), 701–705. 1.4.1, 5

[144] SABI, R., AND TULLER, T. A comparative genomics study on the effect of individual amino acids on ribosome stalling. *BMC Genomics 16*, 10 (Dec. 2015), S5. 1.4.1

[145] SAGI, D., RAK, R., GINGOLD, H., ADIR, I., MAAYAN, G., DAHAN, O., BRODAY, L., PILPEL, Y., AND RECHAVI, O. Tissue- and Time-Specific Expression of Otherwise Identical tRNA Genes. *PLoS Genetics 12*, 8 (Aug. 2016), e1006264. 1.4.1

[146] SANTOS, D. A., SHI, L., TU, B. P., AND WEISSMAN, J. S. Cycloheximide can distort measurements of mRNA levels and translation efficiency. *Nucleic Acids Research 47*, 10 (Mar. 2019), 4974–4985. 5

[147] SCHULLER, A. P., AND GREEN, R. Roadblocks and resolutions in eukaryotic translation. *Nature Reviews Molecular Cell Biology 19*, 8 (Aug. 2018), 526–541. 1.2.3, 1.2.4

[148] SCHULLER, A. P., WU, C. C.-C., DEVER, T. E., BUSKIRK, A. R., AND GREEN, R. eIF5A Functions Globally in Translation Elongation and Termination. *Molecular cell 66*, 2 (Apr. 2017), 194–205.e5. 1.4.1

[149] SENDOEL, A., DUNN, J. G., RODRIGUEZ, E. H., NAIK, S., GOMEZ, N. C., HURWITZ, B., LEVORSE, J., DILL, B. D., SCHRAMEK, D., MOLINA, H., WEISSMAN, J. S., AND FUCHS, E. Translation from unconventional 5′ start sites drives tumour initiation. *Nature 541*, 7638 (Jan. 2017), 494–499. 1.5.3

[150] Shalgi, R., Hurt, J. A., Krykbaeva, I., Taipale, M., Lindquist, S., and Burge, C. B. Widespread regulation of translation by elongation pausing in heat shock. *Molecular cell 49*, 3 (Feb. 2013), 439–452. 1.5.3

[151] Shieh, Y.-W., Minguez, P., Bork, P., Auburger, J. J., Guilbride, D. L., Kramer, G., and Bukau, B. Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science 350*, 6261 (Nov. 2015), 678–680. 1.5.3

[152] Shoemaker, C. J., and Green, R. Translation drives mRNA quality control. *Nature Structural &#38; Molecular Biology 19*, 6 (June 2012), 594–601. 1.4.2

[153] Sonenberg, N., and Hinnebusch, A. G. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell 136*, 4 (Feb. 2009), 731–745. 1.5.3

[154] Stadler, M., and Fire, A. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA 17*, 12 (Dec. 2011), 2063–2073. 1.5.3

[155] Stefani, G., Fraser, C. E., Darnell, J. C., and Darnell, R. B. Fragile X Mental Retardation Protein Is Associated with Translating Polyribosomes in Neuronal Cells. *The Journal of neuroscience : the official journal of the Society for Neuroscience 24*, 33 (Aug. 2004), 7272–7276. 1.4.3

[156] Stein, K. C., and Frydman, J. The stop-and-go traffic regulating protein biogenesis: How translation kinetics controls proteostasis. *Journal of Biological Chemistry 294*, 6 (Feb. 2019), 2076–2084. 1.4.1, 1.5, 1.4.1, 1.4.2, 1.4.2, 1.4.2

[157] Steitz, J. A. Nucleotide sequences of the ribosomal binding sites of bacteriophage R17 RNA. *Cold Spring Harbor Symposia on Quantitative Biology 34* (1969), 621–630. 1.5

[158] Stern-Ginossar, N., Weisburd, B., Michalski, A., Le, V. T. K., Hein, M. Y., Huang, S.-X., Ma, M., Shen, B., Qian, S.-B., Hengel, H., Mann, M., Ingolia, N., and Weissman, J. S. Decoding Human Cytomegalovirus. *Science 338*, 6110 (Nov. 2012), 1088–1093. 1.5.3

[159] Sutton, M. A., Taylor, A. M., Ito, H. T., Pham, A., and Schuman, E. M. Postsynaptic Decoding of Neural Activity. eEF2 as a Biochemical Sensor Coupling Miniature Synaptic Transmission to Local Protein Synthesis. *Neuron 55*, 4 (Aug. 2007), 648–661. 1.4.3

[160] Tahmasebi, S., Khoutorsky, A., Mathews, M. B., and Sonenberg, N. Translation deregulation in human disease. *Nature Reviews Molecular Cell Biology 61* (July 2018), 1. 1.3, 1.3.1

[161] Thommen, M., Holtkamp, W., and Rodnina, M. V. Co-translational protein folding: progress and methods. *Current Opinion in Structural Biology 42* (Feb. 2017), 83–89. 1.1

[162] Thoreen, C. C., Chantranupong, L., Keys, H. R., Wang, T., Gray, N. S., and Sabatini, D. M. A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature 485*, 7396 (May 2012), 109–113. 1.5.3

[163] Tsai, C.-J., Sauna, Z. E., Kimchi-Sarfaty, C., Ambudkar, S. V., Gottesman, M. M., and Nussinov, R. Synonymous Mutations and Ribosome Stalling Can Lead to Altered Folding Pathways and Distinct Minima. *Journal of Molecular Biology 383*, 2 (Nov. 2008), 281–291. 1.4.2

[164] Warnecke, T., and Hurst, L. D. GroEL dependency affects codon usage—support for a critical role of misfolding in gene evolution. *Molecular Systems Biology 6*, 1 (Jan. 2010), 340. 1.4.2

[165] Weinberg, D. E., Shah, P., Eichhorn, S. W., Hussmann, J. A., Plotkin, J. B., and Bartel, D. P. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Reports 14*, 7 (Feb. 2016), 1787–1799. 1.5.3

[166] Wek, R. C. Role of eIF2α Kinases in Translational Control and Adaptation to Cellular Stress. *Cold Spring Harbor Perspectives in Biology 10*, 7 (July 2018), a032870. 1.3

[167] Wilson, D. N., Arenz, S., and Beckmann, R. Translation regulation via nascent polypeptide-mediated ribosome stalling. *Current Opinion in Structural Biology 37* (Apr. 2016), 123–133. 1.4.1

[168] Wohlgemuth, I., Brenner, S., Beringer, M., and Rodnina, M. V. Modulation of the rate of peptidyl transfer on the ribosome by the nature of substrates. *Journal of Biological Chemistry 283*, 47 (Nov. 2008), 32229–32235. 1.2.2, 1.4.1

[169] Wohlgemuth, I., Pohl, C., and Rodnina, M. V. Optimization of speed and accuracy of decoding in translation. *The EMBO Journal 29*, 21 (Nov. 2010), 3701–3709. 1.4.1

[170] Wolin, S. L., and Walter, P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *The EMBO Journal 7*, 11 (Nov. 1988), 3559–3569. 1.5

[171] Woolstenhulme, C. J., Guydosh, N. R., Green, R., and Buskirk, A. R. High-Precision Analysis of Translational Pausing by Ribosome Profiling in Bacteria Lacking EFP. *Cell Reports 11* (Apr. 2015), 1–9. 1.5.2

[172] Xie, J., Wang, X., and Proud, C. G. mTOR inhibitors in cancer therapy. *F1000Research 5* (2016), 2078. 1.5.3

[173] Xue, S., and Barna, M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nature Reviews Molecular Cell Biology 13*, 6 (June 2012), 355–369. 1.1

[174] Young, D. J., Guydosh, N. R., Zhang, F., Hinnebusch, A. G., and Green, R. Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3'UTRs In Vivo. *Cell 162*, 4 (Aug. 2015), 872–884. 1.5.3

[175] Zhou, T., Weems, M., and Wilke, C. O. Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. *Molecular Biology and Evolution 26*, 7 (July 2009), 1571–1580. 1.4.2

Graphic design: Communication Division, UIB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230860106 (print)
9788230848869 (PDF)