New insights into the regulatory role of mRNA secondary structure in *Escherichia coli* through next-generation sequencing

Dissertation
zur Erlangung des akademischen Grades
Doctor rerum naturalium
Dr. rer. nat.

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität Hamburg

vorgelegt von
Cristian Del Campo, M. Sc.

Hamburg, 2016

Die vorgelegte Arbeit wurde von September 2012 bis Mai 2015 am Institut für Biochemie und Biologie der mathematisch-naturwissenschaftlichen Fakultät der Universität Potsdam und von Juni 2015 bis Februar 2016 am Institut für Biochemie und Molekularbiologie am Fachbereich Chemie der Fakultät für Mathematik, Informatik und Naturwissenschaften an der Universität Hamburg unter Anleitung von Frau Prof. Dr. Zoya Ignatova angefertigt.

Gutachter:

Frau Pro f. Dr. Zoya Ignatova

Herr Prof. Dr. Ulrich Hahn

Tag der Disputation: 13 April 2016

Das vorliegende Exemplar der Dissertation war am 18.04.2016 zur Veröffentlichung freigegeben.

"The most beautiful thing we can experience is the mysterious.
It is the source of all true art and science. He to whom the emotion is a stranger, who can no longer pause to wonder and stand wrapped in awe, is as good as dead, his eyes are closed".

(Albert Einstein)

"No one lights a lamp and hides it in a clay jar or puts it under a bed.
Instead, they put it on a stand, so that those who come in can see the light.
For there is nothing hidden that will not be disclosed,
and nothing concealed that will not be known or brought out into the open".

(Luke, 8, 16-17)

Parts of this work were published in the following scientific articles:

"Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function"

Cristian Del Campo, Alexander Bartholomäus, Ivan Fedyunin, Zoya Ignatova

PLoS Genetics, 2015

"Probing dimensionality beyond the linear sequence of mRNA"

Cristian Del Campo, Zoya Ignatova

Current Genetics, 2015

Mapping the non-standardized biases of ribosome profiling

Alexander Bartholomäus, Cristian Del Campo and Zoya Ignatova

Biochemistry, 2016

Table of contents

List of abbreviations

rRNA	ribosomal RNA
aa-tRNA	aminoacyl-transfer RNA
IF	Initiation Factor
SD	Shine-Dalgarno sequence
aSD	anti-Shine-Dalgarno sequence
EF	Elongation Factor
RF	Release Factor
RBS	Ribosome Binding Site
30SIC	30S Initiation Complex
70SIC	70S Initiation Complex
ssRNA	single-stranded RNA
dsRNA	double-stranded RNA
sRNA	small RNA
SRP	Signal Recognition Particle
NGS	Next-Generation Sequencing
CDS	Coding DNA Sequence
UTR	Untranslated Region
RBP	RNA-Binding Protein
RNase	Ribonuclease
DMS	dimethyl sulfate
CMCT	N-cyclohexyl-N'-(2-morpholinoethyl)carbodiimide metho-p-toluenesulfonate
1M7	1-methyl-7-nitroisatoic anhydride
NAI	2-methylnicotinic acid imidazolide
PARS	Parallel Analysis of RNA Secondary structure
RPF	Ribosome Protected Fragment
MHE	Minimum Hybridization Energy
uORF	upstream Open Reading Frame
MNase	Micrococcal Nuclease
nt	nucleotide
BWT	Burrows-Wheeler Transform
DE	Differential Expression

Zusammenfassung

Messenger RNA fungiert als Informationsmolekül zwischen DNA und translatierenden Ribosomen. Immer mehr Studien messen mRNA eine zentralere Rolle in verschiedenen zellulären Prozessen bei. Auch wenn einzelne Beispiele zeigen, dass spezifische strukturelle Eigenschaften der mRNA die Stabilität von Transkripten sowie die Translation regulieren, so sind die Rolle und Funktion für das gesamte bakterielle Transkriptom noch unerforscht.

Next-Generation Sequencing hat sich als bedeutende Methode herausgestellt, um Einblicke in die Regulation von zellulären Prozessen zu gewinnen. Auch wenn einige Schritte besonders im Hinblick auf die Analyse der Daten sorgfältig geprüft werden müssen, liefert die NGS-Technik neue Erkenntnisse bezüglich der transkriptionellen und translationellen Regulation der Genexpression sowie über das Interaktom von Proteinen und Nukleinsäuren auf globaler Ebene.

Hier wurden drei Ansätze der *Deep-Sequencing*-Methode vereint und angewendet, um eine hochauflösende Sicht auf die mRNA-Sekundärstruktur, Translationseffizienz und mRNA-Häufigkeit auf globalem Level zu gewinnen. Wir konnten zuvor unbekannte strukturelle Eigenschaften in der mRNA von *E. coli* entdecken, die Auswirkungen auf die Translation und Degradation von mRNA haben. Ein Sequenzbereich, der kaum Sekundärstrukturen aufweist und vor der eigentlichen gencodierenden Sequenz vorkommt fungiert als zusätzliche unspezifische Bindestelle von Ribosomen und erleichtert so die Initiation der Translation. Trotz der intrinsischen Neigung sekundäre und tertiäre Interaktionen einzugehen, sind Sekundärstrukturen innerhalb von codierenden Sequenzen hochdynamisch und beeinflussen die Translation lediglich nur an wenigen Positionen. Eine Sekundärstruktur vor dem Stopcodon ist angereichert in Genen, die ein UAA als Stopcodon verwenden und spielt demnach wahrscheinlich für die Termination der Translation eine Rolle. Die Analyse auf globaler Ebene hat weiterhin eine allgemeine Erkennungssequenz der RNase E aufgedeckt, welche die endonukleolytische Spaltung initiiert. Somit wird in der vorliegenden Arbeit zum ersten Mal das „RNA-Strukturom“ von *E. coli* bestimmt, was den Einfluss der mRNA Struktur als direkten Effektor an einer Vielzahl von Prozessen wie Translation und mRNA-Degradation hervorhebt.

Zusätzlich haben wir die Vor- und Nachteile neuer Technologien kritisch begutachtet, die auf NGS-Methoden basieren, um zum einen die RNA-Struktur aufzuklären und zum anderen translatierende Ribosomen zu detektieren, um einen nützlichen Leitfaden für die korrekte Wahl der entsprechenden Methode bezüglich der Anwendungsbedürfnisses und angesichts der Auflösung der Datenanalyse zu geben.

Abstract

Messenger RNA acts as an information molecule between DNA and translating ribosomes. Emerging evidence places mRNA more centrally in various cellular processes. Although individual examples show that specific structural features of mRNA regulate translation and transcript stability, the role and function for the whole bacterial transcriptome remains unknown.

Next-generation sequencing emerged as a powerful tool to gain insights in regulation of cellular processes. Although with some pitfalls that need to be carefully assessed in data analysis, NGS-based techniques provided new insights in transcriptional and translational regulation of gene expression and protein-nucleic acid interactome on a global level.

Combining three deep-sequencing approaches to provide a high resolution view of global mRNA secondary structure, translation efficiency and mRNA abundance, we unraveled unseen structural features in *E. coli* mRNA with implications in translation and mRNA degradation. A poorly structured site upstream of the coding sequence serves as an additional unspecific binding site of the ribosomes and facilitates initiation of translation. Despite intrinsically prone to establish secondary and tertiary interactions, secondary structures within coding sequences are highly dynamic and influence translation only within a very small subset of positions. A secondary structure upstream of the stop codon is enriched in genes terminated by UAA codon with likely implications in translation termination. The global analysis further substantiates a common recognition signature of RNase E to initiate endonucleolytic cleavage. This work determines for the first time the *E. coli* RNA structurome, highlighting the contribution of mRNA secondary structure as a direct effector of a variety of processes, including translation and mRNA degradation.

Additionally, we critically review pros and cons of emerging new technologies, the NGS-based approaches to assess RNA structure and to profile translating ribosomes, in order to provide a useful guide for a correct choice of relative corresponding technique, in regards of the application needs and considering the resolution of each data analysis.

1. Introduction

1.1 The genetic information flow: from gene to physiological function

According to the central dogma of biology, the synthesis of a protein starts from the transcription of the genetic information written in the DNA into an intermediate molecule, named messenger RNA (mRNA), which is later translated into a protein by complex molecular machine, the ribosome.

Proteins are the molecular "tool" of life, which the cell uses to grow, reproduce, interact with the environment, and respond to environmental changes. Their synthesis is energetically expensive process and it is fascinating to discover how many processes the cell evolved in order to tightly regulate each single step of protein production.

In bacteria, the RNA polymerase transcribes protein-encoding genes into mRNA, a molecule generally considered as a mere carrier of the genetic information that however regulates its cellular localization (Martin & Ephrussi, 2009, Buxbaum *et al.*, 2015), gene expression (Mortimer *et al.*, 2014) and stress response (Winkler & Breaker, 2005, Kortmann & Narberhaus, 2012), through its folding in tridimensional structures. Transcription is assisted by transcription factors and regulatory proteins able to interact both with the RNA polymerase and with sequences upstream of the transcriptional start, named promoter. Thus, a first layer of regulating gene expression is established through the control of the expression of certain transcription factors or modulating their interactions with the RNA (Jacob & Monod, 1961).

As soon as it is released from the RNA polymerase, the nascent RNA folds into tridimensional, secondary structures, due to the Watson-Crick interactions between ribonucleotides. Generally, these structures are quite dynamic and fluctuate between open and close conformation (Mahen *et al.*, 2010) and their role is essential for the cell. For example, rho-independent terminated transcripts contain a sequence named "terminator" that folds into a stem-loop structure right after is transcribed, destabilizing the interaction with the RNA polymerase that immediately detaches from the RNA (Wilson & von Hippel, 1995). This is just one example of the important regulatory role of RNA secondary structure (more details will be provided in paragraph 1.3).

In prokaryotes, the transcript is released directly into the cytosol to be translated. In reality, translation of the messenger is assumed to be co-transcriptional (Miller *et al.*, 1970) and the ribosome preventing the backtracking of the RNA polymerase (Proshkin *et al.*, 2010).

However, recent single-molecule microscopy shows that in *E. coli* most of the translation is not coupled to transcription, but rather takes place on mRNA that has already diffused away from the nucleoid region to ribosome-rich cytoplasmic regions (Bakshi *et al.*, 2012).

Translation is executed by the ribosome, a large macromolecular machine which in prokaryotes consists of three ribosomal RNA (rRNA) and 52 proteins (Schuwirth *et al.*, 2005), that are assembled in two subunits. The 5S and 23S rRNA are part of the large 50S subunit while the 16S rRNA is part of the small 30S subunit. The ribosome contains three distinct sites: A-, P- and E-site (Melnikov *et al.*, 2012). The A-site is the entrance point for the aminoacyl-transfer RNA (aa-tRNA), except for the first aminoacyl-tRNA, which enters directly at the P site (Laursen *et al.*, 2005). In the peptidyl-transferase center, between A- and P-site, the peptide bond is formed between the aa-tRNA and the nascent polypeptide chain. The E site represents the exit site of the deacylated tRNA (Burkhardt *et al.*, 1998).

Translation can be divided in three main subprocesses: initiation, elongation and termination (Fig. 1.1). In the initiation, the small ribosomal subunit (30S) forms an initiation complex with the initiator tRNA (fMet-tRNA) bound to initiation factor 2 (IF2), two additional factors (IF1 and IF3) and the mRNA (Laursen *et al.*, 2005, Simonetti *et al.*, 2009). The fMet-tRNA enters directly in the P-site, differently from all other tRNAs entering from the A-site. A specific sequence upstream of the start codon, named Shine-Dalgarno sequence (SD), drives the interaction with a complementary anti-SD sequence on the 16S rRNA of the small subunit and aligns the first (start) codon of the protein coding sequence in the P-site of the ribosome (Shine & Dalgarno, 1975, Kaminishi *et al.*, 2007). The length and nucleotide composition of this ribosome-binding sequence along with its availability to interact with the cellular environment determine the efficiency of translation initiation (Kozak, 1999, Osterman *et al.*, 2013), which the cell exploits to regulate the translational process (see paragraph 1.2.1). After the formation of the initiation complex, the big subunit joins the small subunit and the elongation phase starts.

In the elongation, the assembled ribosome moves along the mRNA and “translates” the nucleotide information encoded in the RNA into the amino-acid language (Rodnina *et al.*, 1999, Noeske & Cate, 2012). The ‘interpreting bilingual molecule’ of this process is the tRNA, an RNA molecule that links specifically the coding nucleotide triplets, termed codon, to the corresponding amino acid (Zamecnik, 2005, Rodnina & Wintermeyer, 2011). The genetic code is composed of 61 sense codons encoding the standard 20 amino acids,

additionally to the 3 coding for the translational stop signal. Because 18 of 20 amino acids are encoded by multiple synonymous codons, the genetic code is termed “degenerate” (Reichmann *et al.*, 1962).

Aminoacylated-tRNA assembled in ternary complex with the elongation factor Tu (EF-Tu) and GTP brings the amino acids to the ribosome. The corresponding amino acid is selected through the interaction of the codon with the anticodon sequence of the tRNA (Labuda *et al.*, 1984). After GTP hydrolysis, the 3′ end of aminoacyl-tRNA accommodates in the peptidyl transferase center and immediately enters the peptidyl transfer reaction. Formation of the peptide bond results in deacylated tRNA in the P site and peptidyl-tRNA in the A site (Rodnina *et al.*, 1999). Hydrolysis of GTP by EF-G triggers displacing of the peptidyl-tRNA from the A site to the P site, while the deacylated tRNA is transferred from the P site to the E site from where it dissociates (Rodnina & Wintermeyer, 2011). The cycle then restarts with a new aa-tRNA entering the A-site and the simultaneous exit of the deacylated-tRNA from the E site.

Once the ribosome encounters one of the three stop codons (i.e., UAG, UAA, UGA), translation is terminated and the polypeptide chain is released from the ribosome (Fig. 1.1) (Korkmaz *et al.*, 2014). The stop codons are not recognized by any tRNAs. Instead, they interact with two different proteins: release factor 1 (RF1) recognizes UAA and UAG stop codon, while release factor 2 (RF2) associates with UAA and UGA (Scolnick *et al.*, 1968). These factors bind to any ribosome with a stop codon positioned in the A site, ‘forcing’ the peptidyl transferase in the ribosome to catalyze the hydrolytic cleavage of the nascent chain from the peptidyl-tRNA (Youngman *et al.*, 2007). This reaction releases the polypeptide chain from the tRNA. The ribosome dissociates into two subunit through the hydrolysis of GTP, mediated by a third release factor (RF3) (Freistroffer *et al.*, 1997).

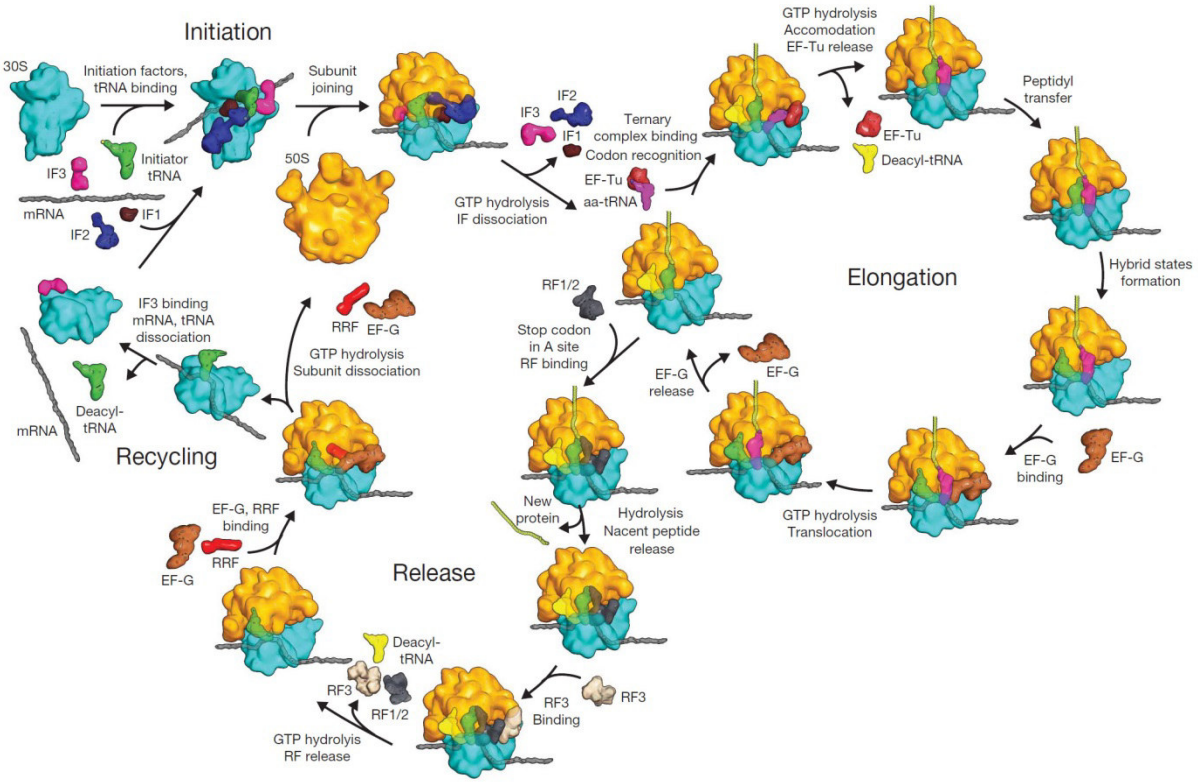


Figure 1.1 | Schematic of the prokaryotic translation cycle. Translation can be subdivided into: initiation, elongation, termination (or release) and ribosome recycling. IF, initiation factor; EF, elongation factor; RF, release factor. (Adopted from (Schmeing & Ramakrishnan, 2009))

1.2 Translational regulation in prokaryotes

Bacteria are versatile organisms able to live in a wide range of conditions and to adapt to environmental changes. To adjust fast to new growth condition and to respond to external stress stimuli, bacteria fine-tune their proteome via multi-step regulations at both transcriptional and translational level. Particularly, the second one is a much faster process, enabling a quick reshaping of the proteome and activating stress response mechanisms (Dahan *et al.*, 2011, Gingold & Pilpel, 2011, Picard *et al.*, 2012, Starosta *et al.*, 2014). Indeed, mRNA concentration correlates only partially with the protein abundance on a single cell level ($R^2 \approx 47–77\%$ in *E. coli* (Corbin *et al.*, 2003, Lu *et al.*, 2007, Taniguchi *et al.*, 2010) and even less in other bacteria (Dressaire *et al.*, 2010).

Up to date, many aspects of translational regulation have been described, yet mechanistic insights are missing or not completely understood. In the next paragraphs, we will review some examples of translational control in prokaryotes, focusing on the regulation of translation initiation and elongation.

1.2.1 Starting from the beginning: regulating initiation to control the whole process

In bacteria, translation regulation is mostly executed by targeting the translation initiation process (Spirin, 2002, Duval *et al.*, 2015). As described earlier, in this phase the small subunit of the ribosome together with the initiator tRNA (fMet-tRNA) and the three initiation factors (IF1, IF2, IF3) binds the mRNA at its ribosome binding site (RBS), forming the 30S initiation complex (30SIC). The RBS includes different elements: the SD sequence complementary to the anti-SD (aSD) sequence at the 3' end of the 16S rRNA (Shine & Dalgarno, 1974), the start codon and some additional sequences (enhancers) that improve the ribosome-mRNA interaction (Stormo *et al.*, 1982). For many bacterial mRNAs, the selection of the appropriate initiation codon (the canonical AUG, GUG or UUG) depends largely on the formation of this short SD-aSD double helix.

The three initiation factors cooperate to correctly position the fMet-tRNA at the start codon, ensuring that no other aa-tRNA will cover that position (Simonetti *et al.*, 2008, Julian *et al.*, 2011). IF1 associates with the 30S ribosomal subunit in the A site, preventing an aminoacyl-tRNA from entering and favoring the binding of IF3 and IF2 (Milon *et al.*, 2012). IF2 is a GTPase which maintains the initiator tRNA in the correct P/I position (Simonetti *et al.*, 2013). IF3 discriminates against non-canonical initiation codon (AUU and AUC) (Sussman *et al.*, 1996, Petrelli *et al.*, 2003) exerting this proofreading activity during the joining of the subunits.

Once the 30SIC complex is assembled, the large ribosomal subunit (50S) joins, forming the 70S initiation complex (70SIC), with the parallel release of all initiation factors (Allen *et al.*, 2005, Julian *et al.*, 2011).

The binding of the mRNA to the 30S subunit is likely the slowest and thus rate-limiting step of the initiation. A free 30S ribosomal subunit can bind non-specifically to any single-stranded RNA (ssRNA) region, through the ssRNA-specific S1 ribosomal protein (Draper & von Hippel, 1978, Hajnsdorf & Boni, 2012). If this region is located close to a translational initiation site, the mRNA-30S interaction is stabilized by the SD-aSD and the codon–anticodon annealing (Gualerzi & Pon, 2015). However, if the SD sequence is not accessible and, for example, occluded in a stem-loop structure, the binding of the small subunit is delayed until the unfolding of this hairpin. This led to the hypothesis, mostly triggered by

mathematic modelling, of a “stand-by site” at which the 30S subunits dwells until the SD becomes available (de Smit & van Duin, 2003, Studer & Joseph, 2006, Espah Borujeni *et al.*, 2014). This site is suggested to increase the local concentration of 30S, facilitating the assembling of the whole 30SIC (de Smit & van Duin, 2003, Marzi *et al.*, 2007, Vimberg *et al.*, 2007).

The SD sequence is thought to be the most effective determinant of translation initiation efficiency. Indeed, different SD features, including its length, distance from the ATG (termed spacing) and accessibility, were shown to modulate protein expression (Vimberg *et al.*, 2007). Its sequence consists of three to nine contiguous bases in the mRNA complementary to some or all of bases 1534 to 1542 (ACCUCCUUA) at the 3’ end of 16S rRNA. The strength of the SD-aSD interaction depends on the extent of base pairing interactions and on the mRNA SD sequence. In this context, it is surprising that SD sequences in some mRNAs consist of only three or four bases, despite the conserved length of nine bases of the anti-SD in 16S rRNA (Chang *et al.*, 2006). As proposed by Kozak (1983) and found to be consistent in 30 prokaryotes (Ma *et al.*, 2002), the efficiency of translation is usually higher when the SD interaction involves the core of anti-SD sequence in the 16S rRNA (i.e. CCUC or CUCC) than the off-center region. A strong SD is more efficient in counteracting the mRNA secondary structures that may hinder ribosome access to RBS (de Smit & van Duin, 1994a). However, an overly extended SD may have an inhibitory effect on the translation initiation (Komarova *et al.*, 2002). The spacing between the SD sequence and the initiation codon is also an important determinant of protein synthesis yield (Chen *et al.*, 1994). A genome-wide analysis evidenced that this distance ranges from 4 to 18 nt among 4122 genes of *Escherichia coli* (Shultzaberger *et al.*, 2001), with an optimal length around 9 nt (Ringquist *et al.*, 1992, Chen *et al.*, 1994).

Finally, the accessibility of translation initiation signals by the 30S subunit can be restricted by mRNA secondary structure (Gold, 1988, de Smit & van Duin, 1994b). In general, stem-loop hairpins involving the RBS hide the SD sequence or the start codon, preventing the interaction with the complementary region of the 16S rRNA. Indeed, even when a SD is missing, local absence of secondary structure permits translation of mRNAs (Scharff *et al.*, 2011).

Also, more complex mRNA structures, termed translational operators, can directly sense the environmental cues, and/or can be recognized by trans-acting factors, which range from

metabolites to trans-acting small non coding RNA (sRNA) and proteins (Spirin, 2002, Winkler & Breaker, 2005). Conformational rearrangements of the structured RBS induced by environmental stimuli represent an evolutionary strategy for the cell to regulate translation in a fast and direct way.

Mechanisms of control of translation initiation can be grouped in two categories: 1) mediated by trans-acting element (protein or small RNA) which prevents binding of the ribosome to the mRNA acting either at the stand-by site or at the RBS, and 2) mediated by cis-acting mRNA elements acting as sensors (reviewed in (Spirin, 2002).

To the first group belong RNA binding proteins able to regulate their own mRNA, since the RBS of the latter have similar features to the RNA targets of the binding protein. For example, the translation initiation factor IF3 negatively controls its own synthesis (Butler *et al.*, 1987). IF3 binds to the 30S subunit and inhibits translation initiation at codons other than AUG, GUG, or UUG (Sussman *et al.*, 1996). The initiation codon of IF3 is a non-canonical AUU. When the IF3 concentration increases, the number of 30S subunits bound to IF3 increases, causing a decrease in the translation of its own mRNA (Butler *et al.*, 1987). In an analogous fashion, some ribosomal proteins are able to bind the RBS of the cistron encoding several ribosomal protein (ribo-protein genes are generally structured in operons). The 5' UTR of the polycistrionic mRNA folds in a similar way to the region of the rRNA, with which the ribosomal protein is interacting. This mechanism is valid for the ribosomal protein S8 on the *spc* operon (Cerretti *et al.*, 1988) and, in a similar way, for the protein S1 on its own, monocistrionic mRNA (Boni *et al.*, 2001).

Translation attenuation is a general mechanism where translation is prevented by sequestering the SD sequence in a hairpin-loop structure, which opens only when the ribosome stalls on an upstream sequence that prevents the refolding of the structure. The *secA* operon constitutes a clear example of attenuation (Nakatogawa *et al.*, 2004). The SD sequence of *secA* is generally folded with the stop codon region of the upstream located *secM*. Translation of *secA* occurs only when the ribosome translating *secM* moves till a stalling site located five codons upstream of the *secM* stop codon, inducing the opening of the stem-loop structure and the exposition of the *secA* SD to a new ribosome. (Nakatogawa *et al.*, 2004). Similar mechanisms were also found in other organism, e.g. stalling at the end of MifM ORF to allow translation of *yidC2* in *Bacillus subtilis* (Chiba *et al.*, 2009).

Alternatively to protein, the RBS can be also masked by small RNAs. Through this interaction, the RyhB sRNA affects the expression of at least 18 operons in response to iron limitation (Masse *et al.*, 2005). The *sodB* mRNA carries an unstructured RBS with both a strong SD sequence and AUG codon, making it a very efficient system for initiation. To repress its expression, the sRNA RyhB together with the global regulative factor Hfq targets the RBS, preventing ribosomal binding (Geissmann & Touati, 2004).

For other mRNAs, the sensing activity and the resulting translational activation or inactivation is performed directly by the mRNA itself, through the action of elements present in cis. mRNA structures present in 5' UTR are able to sense physicochemical signals, like pH, metabolite concentration (Winkler & Breaker, 2005), and temperature (Kortmann & Narberhaus, 2012).

A riboswitch is characterized by a complex RNA structure, composed by two functional domains: an "aptamer" domain, which senses the environment and binds the target molecule, and an expression platform, which modulate expression of the structural genes (Winkler & Breaker, 2005).

As a general mechanism, the binding of the target metabolite to this aptamer induces conformational changes in the expression platform, which activates or inhibits gene expression through folding or unfolding of the RBS. A well-characterized example is the thi box aptamer of the *E. coli thiM/C* genes. The TPP (thiamine pyrophosphate) binding to the thi box causes occlusion of the downstream RBS (Winkler *et al.*, 2002).

RNA thermometers are structures able to sense temperature shifts. In this case, the heat induces unfolding of the regulatory region, which covers also the RBS (Kortmann & Narberhaus, 2012). Some heat shock genes (like the *prfA* gene of *Listeria monocytogenes*) carry a motif named ROSE (repressor of heat-shock gene expression), characterized by non-canonical base pairs (G-U), which is highly sensitive to heat, and gradually melts proportionally to temperature rise, exposing the RBS which is occluded in structure at ambient temperature (Narberhaus *et al.*, 2006).

1.2.2 Control of translation elongation: the ribosomal path through the codon usage bias

After initiation, the ribosome proceeds into elongation, catalyzing the formation of peptide bonds between amino acids added in a series determined by the codon sequence of the mRNA. In *E. coli*, the ribosome elongates the nascent chain with an average elongation rate around 14 amino acids/second (Young & Bremer, 1976, Varenne *et al.*, 1984, Proshkin *et al.*, 2010). The speed of elongation is not constant and additional information embedded in the coding sequence determines the local ribosome speed. The nucleotide sequence of a messenger RNA does not only encode the amino acid but the selection of one specific codon within a set of synonymous codons (i.e., codons codifying for the same amino acid) is evolutionary forced to modulate ribosome speed towards optimization of protein expression (Plotkin & Kudla, 2011). The different usage of the synonymous codons results in a bias of the frequency of occurrence of a codon (Shabalina *et al.*, 2013, Quax *et al.*, 2015).

Since its discovery, codon bias was suggested to positively correlate with gene expression level in both prokaryotes (Ikemura, 1985, Bulmer, 1987, Kanaya *et al.*, 1999) and eukaryotes (Ikemura, 1985, Akashi, 1994, Duret, 2000). At least in prokaryotes, the concentration of the cognate tRNAs correlates with the frequency of occurrence of a codon (Kanaya *et al.*, 1999). Optimizing the overall codon sequence to more frequent codons does, at least in some cases, result in increased heterologous gene expression (Gustafsson *et al.*, 2004). This is only true for single-domain protein, while solubility of multi-domain improves when synonymous substitutions of slow translated codons are inserted at the border of protein structural domains (Hess *et al.*, 2015). Highly expressed genes are enriched in codons usually read by most abundant tRNAs (Dong *et al.*, 1996). Since tRNAs reach ribosome only driven by molecular diffusion (Fluitt *et al.*, 2007), translational rate depends mainly on the tRNA concentration, with codons pairing to highly abundant tRNA translated at higher rates then codons read by lowly abundant tRNAs (Berg & Kurland, 1997, Zhang & Ignatova, 2009). Stretches in the mRNA enriched in non-optimal codons induce transient ribosomal pausing (Zhang *et al.*, 2009). The slow-translating regions are located at the domain boundaries of multidomain proteins and actively coordinate the co-translational folding (Komar, 2009, Zhang *et al.*, 2009, Yu *et al.*, 2015).

Ribosome profiling, a recently developed technology that allows determining the position of the translating ribosomes on a transcriptome-wide level (Ingolia *et al.*, 2009) (see paragraph

1.4), emerged as promising approach to give deeper insights into the analysis of the codon bias on translation elongation. However, it raised more questions, than answers. Many studies found no correlations between speed of elongation and frequency of the codon usage (Ingolia *et al.*, 2011, Qian *et al.*, 2012, Ingolia *et al.*, 2014, Pop *et al.*, 2014). In contrast, translational slow-down was correlated with wobble base-pairing (Stadler & Fire, 2011), or attributed to sequences encoding specific amino acid stretches, such as consecutive proline residues (Woolstenhulme *et al.*, 2013) or positively charged amino acids (Charneski & Hurst, 2013). In bacteria, Shine-Dalgarno like sequences was suggested to be the main determinant of ribosomal pausing (Li *et al.*, 2012c). A recent study, however, noticed experimental biases in the preparation of these datasets which most likely triggered this obviously wrong conclusions: a pre-treatment with elongation inhibitor to stabilize the ribosome-RNA interaction provokes the loss of the codon resolution, thus masking the translational rate dependence on fast and slow translated codon associated to tRNA abundance (Hussmann *et al.*, 2015). Also, two recent studies concluded that rare codons with less abundant cognate tRNAs are decoded slower, thus resulting in decreased translation elongation rates (Dana & Tuller, 2014, Gardin *et al.*, 2014). Additionally, refinements in the profiling method argued that SD-like motifs have no effect on elongation rates and that the previous observation derived from pitfalls in sample processing (Mohammad *et al.*, 2016).

The debate on the influence of codon usage on translational elongation is still on-going as well as the improvement of techniques able to detect it. Indeed, bioinformatic analyses of ribosome profiling datasets often resulted in conflicting conclusions (Tuller *et al.*, 2010a, Charneski & Hurst, 2013, Artieri & Fraser, 2014, Pop *et al.*, 2014, Hussmann *et al.*, 2015), because of the sensitivity of this technique, which can be influenced by growth conditions, depth of coverage, cloning or sequencing biases, methods of bioinformatics analysis, and experimental noise (Artieri & Fraser, 2014, Gardin *et al.*, 2014, Lareau *et al.*, 2014, Nakahigashi *et al.*, 2014, Hussmann *et al.*, 2015). A recent experimental work took advantage of a cell-free translation system from the fungus *Neurospora crassa*, which exhibits a strong codon usage bias, to demonstrate that codon usage has indeed a function in regulating protein synthesis by affecting co-translational protein folding (Yu *et al.*, 2015).

Additionally to the single codon, the context surrounding a specific position also affects translational elongation. Within a gene, synonymous codons recognized by the one tRNA

tend to cluster, generating a bias termed co-occurrence (Cannarozzi *et al.*, 2010). The effect of co-occurrence involves both frequent and rare codons and is most prominent in highly expressed genes that must be rapidly induced, such as those involved in stress response (Cannarozzi *et al.*, 2010). The hypothesis behind this is that the re-use of the same codon increases the probability to recycle the same tRNA, given a fast recharging by the corresponding amino-acyl-tRNA synthetase that co-localizes with the ribosome (Cannarozzi *et al.*, 2010, Godinic-Mikulcic *et al.*, 2014).

One more variable involved in codon context is selection for codon pairs. In *E. coli*, as well as in humans, codon pairs have been shown to be overrepresented (Gutman & Hatfield, 1989) or almost completely avoided (Coleman *et al.*, 2008). The reason of this phenomenon is still unknown, but it has been shown that modification of codon pairs in the poliovirus genome results in several fold reduction in protein yield and a reduction in viral infectivity of 1,000-fold in mammalian cells (Coleman *et al.*, 2008). However, codon pair bias is a direct consequence of dinucleotide bias and it is still discussed whether virus attenuation is an effect of codon pair or dinucleotide deoptimization (Kunec & Osterrieder, 2016).

In many species, the region immediately downstream of the start codon shows a preference for certain codons with a highly debated origin of the selective pressure that shapes this usage. Tuller *et al.* found a 'ramp' of codons corresponding to rare tRNAs (estimated from gene copy number) in the first 90–150 nucleotides of genes. The authors hypothesized that such ramp would slow down the ribosome entering the elongation phase, reducing the risk of ribosomal traffic jams towards the 3′ end (Tuller *et al.*, 2010a). Other analysis reveal an alternative explanation: reduction of mRNA secondary structures around the start codon is the selection force, rather than the rarity of a codon (Kudla *et al.*, 2009, Bentele *et al.*, 2013, Goodman *et al.*, 2013). Low propensity of the mRNA to fold facilitates initiation and start of protein synthesis. Analysis of large libraries of synonymous variants of reporter genes in *E. coli* and *S. cerevisiae* showed that variation in protein expression can be explained by differences in mRNA folding around the start, both for heterologous (Kudla *et al.*, 2009) and for endogenous gene expression (Bentele *et al.*, 2013, Goodman *et al.*, 2013). However, in a recent work, Tuller argued that the synthetic constructs used to test this possibility were selected for strong folding, masking an eventual presence of codon ramp (Tuller & Zur, 2015).

The primary nucleotide sequence encodes also structural information for mRNA folding and synonymous substitutions can induce conformational changes, causing formation of new stable hairpin loops and elements of higher-order folding (Shabalina *et al.*, 2013). Recently, a trade-off between tRNA abundance and mRNA secondary structure support was proposed to keep translation elongation rate constant, selecting for fast-translated codons in highly structured regions (Gorochowski *et al.*, 2015). This would produce a well-distributed coverage of ribosomes along the transcript that would prevent mRNA degradation. Additionally, this would combine the detrimental and beneficial effects on elongation rate intrinsic in their individual role (Gorochowski *et al.*, 2015).

Despite the extended evidences of its regulatory function in translational initiation (see paragraph 1.2.1), the role of mRNA secondary structure in global translational control remains unclear. In the next paragraph, we review in depth the most recent findings on its implications in translational regulation.

1.3 The multiple functions of a simple base-pairing: the regulatory role of mRNA secondary structure

The ribonucleic acid bases exhibit an intrinsic propensity to fold and form double-stranded helices linked by complementary Watson-Crick pairs separated by single-stranded regions in the shape of stem-loop hairpins (Brion & Westhof, 1997), energetically very stable ($\Delta G^{\circ} = -1$ to -3 kcal mol^{-1} per base pair) (Turner *et al.*, 1988). RNA structure forms already during transcription (Kramer & Mills, 1981, Lai *et al.*, 2013), on the same timescale as RNA synthesis (Brehm & Cech, 1983). The speed of transcription ranges from 20 to 80 nt/sec in bacteria (Pan & Sosnick, 2006), a longer time scale compared to the fast folding of RNA, which is known to occur on a range of 10–100 μsec (Al-Hashimi & Walter, 2008) and can persist for minutes or hours (Sosnick & Pan, 2003, Thirumalai & Hyeon, 2005, Al-Hashimi & Walter, 2008). RNA polymerase pausing while transcribing assists co-transcriptional folding (Toulme *et al.*, 2005, Wong *et al.*, 2007), which was shown to happen sequentially both *in vivo* and *in vitro* (Mahen *et al.*, 2005, Mahen *et al.*, 2010). *In vivo*, RNA is highly flexible and can rapidly exchange conformations (LeCuyer & Crothers, 1994, Mahen *et al.*, 2010). Indeed, folding often involves transient RNA structure elements, i.e., structural features that are only present for a specific time span (Kramer & Mills, 1981, Repsilber *et al.*, 1999).

Since in bacteria transcription is also associated with translation, one could assume that most of the mRNA is linear or already covered by ribosomes. However, since the RNA folding time is much faster than the ribosome association to the mRNA, structures can rapidly form, generating, for example, regulative domain like the one present in riboswitches (Yakhnin *et al.*, 2006). On the other side, in the kinetic model, there is a certain time lapse during which ribosomes can initiate translating the nascent transcript, before formation of the long-range interaction in the mRNA (Groeneveld *et al.*, 1995).

As already discussed, redundancy in the genetic code gives various levels of freedom for the optimization of translation, through the modulation of coexistence of different regulative factors, simply mediated by selection of the nucleotide sequence. Already in 1970s, it was suggested that redundancy of the genetic code allows preservation of both protein and mRNA structure (White *et al.*, 1972, Fitch, 1974). Indeed, a trinucleotide structural periodic pattern is an intrinsic property of the genetic code conserved in all genes of different species (Shabalina *et al.*, 2006). Additionally, the need to maintain intact mRNA structures imposes additional evolutionary constraints on bacterial genomes, which go beyond preservation of structure and function of the encoded proteins (Chursov *et al.*, 2013, Mao *et al.*, 2013).

Although stable secondary structures capable of interfering with translation tend to be avoided in mRNA coding regions, significant biases in favor of local RNA structures have been found in several bacterial species and yeast (Katz & Burge, 2003), with native mRNAs having a lower calculated folding free energy than random sequences (Seffens & Digby, 1999).

Computational predictions support the hypothesis of an existing positive relationship between mRNA folding energies in coding sequences and translational efficiency, that however does not directly depend on the susceptibility of RNA to degradation (Zur & Tuller, 2012). Furthermore, structures within the coding sequence seems to have a direct effect on the translating ribosome, slowing down the rate of elongation (Tuller *et al.*, 2010b), suggesting that mRNA secondary structures serve as elongation brakes to control the speed and hence the fidelity of protein translation and explaining why highly expressed genes tend to have strong mRNA folding, slow translational elongation, and conserved protein sequences (Yang *et al.*, 2014). An additional explanation for the selection of high structure within protein coding sequences is a compensatory effect for translation of fast codons, in order to preserve the

translation rate constant: indeed, highly or lowly structured regions are enriched in fast- or slow-translated codons, respectively (Gorochowski *et al.*, 2015).

Biophysical experiments further proved the ability of stem-loop to decelerate the ribosome: optical-trap studies, along with FRET experiments, have quantified the decrease of translational rate caused by very large mRNA duplexes, together with the ability of the ribosome to unfold the structure (Wen *et al.*, 2008, Qu *et al.*, 2011, Chen *et al.*, 2013). Thus, highly stable structure, like pseudoknot, are known to stall the ribosome and induce frameshift, a backtrack that brings it out of frame (Chen *et al.*, 2014, Kim *et al.*, 2014)

Despite the firmly established relationship between RNA folding and translation, the function of this *liason* is not determined yet.

Coupling the susceptibility of paired or unpaired nucleotides to chemical modification as well as enzymatic cleavage with new advances in next generation sequencing (NGS) (reviewed in the next paragraph) enables transcriptome-wide determination of mRNA structures (Wan *et al.*, 2011, Mortimer *et al.*, 2014, Kwok *et al.*, 2015). Meta-genome analysis revealed conserved structural features in specific regions or gene groups, which unravel the surprisingly extensive regulatory role of mRNA folding in many cellular processes.

The predicted triplet structural periodicity in the genetic code was confirmed by periodicity of probe reactivity (i.e., reactivity cycling regularly every three nucleotides) within coding sequences (CDS) but not untranslated regions (UTRs) in yeast, mouse, and human *in vitro* (Kertesz *et al.*, 2010, Li *et al.*, 2012a, Wan *et al.*, 2014), and in Arabidopsis and mouse *in vivo* (Ding *et al.*, 2014, Incarnato *et al.*, 2014, Spitale *et al.*, 2015).

The structural content of the 5´UTRs and 3´UTRs relative to the coding regions varies from organism to organism. The 5´UTRs and 3´UTRs were less structured than the coding regions on average for *S. cerevisiae* and *A. thaliana* (probably due to the high processing into small regulatory RNAs in plants (Zheng *et al.*, 2010, Li *et al.*, 2012b), whereas opposite results were obtained for *Drosophila melanogaster*, *Caenorhabditis elegans*, mouse and human mRNAs (Kertesz *et al.*, 2010, Li *et al.*, 2012a, Li *et al.*, 2012b, Incarnato *et al.*, 2014, Wan *et al.*, 2014). Differences were also highlighted in the structure at the start and stop codon: local minima were found in yeast, mouse, human and two metazoans (Kertesz *et al.*, 2010, Li *et al.*, 2012a, Wan *et al.*, 2012, Incarnato *et al.*, 2014, Wan *et al.*, 2014, Spitale *et al.*, 2015) (but not in plants (Li *et al.*, 2012b, Ding *et al.*, 2014)). As already discussed, low secondary structure

around the initiation site facilitate ribosomal access (Kudla *et al.*, 2009, Scharff *et al.*, 2011, Bentele *et al.*, 2013). Indeed, an anti-correlation was observed in *S. cerevisiae* between RNA structure of the region right upstream of the translation start site and ribosome density throughout the transcript, a proxy for translational efficiency (Kertesz *et al.*, 2010). These results provided the first experimental evidence that there is a selective pressure for low structure around the start codon, supporting previous studies, which came to the same conclusion analyzing a reporter protein expression in synthetic libraries (Kudla *et al.*, 2009, Scharff *et al.*, 2011, Bentele *et al.*, 2013). However, these libraries were all expressed in *E. coli* and the question whether the mRNA structural profile in bacteria reflects the one in yeast is still open.

A relationship between secondary structure on translation was also evidenced in plants, in which the more structured transcripts were more ribosome-associated than the less structured ones (Li *et al.*, 2012b), even though the reason of this correlation is still unknown.

In vivo, mRNAs tend to be more often unfolded or more dynamic (Ding *et al.*, 2014, Rouskin *et al.*, 2014); particularly stress-related transcripts in plants were found to have greater "single-strandedness" (Ding *et al.*, 2014), a behavior also observed in human cells (Wan *et al.*, 2014), while genes involved in basic biological functions such as gene expression, protein maturation and processing show a more conserved, stable structure (Ding *et al.*, 2014). Generally, transcripts containing RNA duplexes in the protein coding sequence were poorly translated, whereas those with 3'UTR duplexes were highly translated (Sugimoto *et al.*, 2015) The lower structure propensity *in vivo* was related to the unwinding activity of the ribosome (Li *et al.*, 2012a, Wan *et al.*, 2014, Sugimoto *et al.*, 2015),or ATP-dependent RNA helicase unwinding the RNA (Rouskin *et al.*, 2014), which are lacked in *in vitro* experiments explaining the global tendency of the mRNA to be more folded (Kertesz *et al.*, 2010). Additionally, structures detected *in vivo* have a strong propensity for high thermostability and match structures identified *in vitro* at high temperature in a thermal unfolding study (Wan *et al.*, 2012). The latter work characterized the mRNA structurome of *S. cerevisiae* at different folding energies showing that mRNAs thermodynamically more stable across the entire transcript were enriched during heat shock *in vivo* (Wan *et al.*, 2012). In particular, 3' UTR structures prevent degradation by the exosome complex, which is active at higher temperature and requires a 3′ ssRNA region of about 30 nucleotides in its targets (Wan *et al.*, 2012). These findings have been further validated *in vitro* and proved the existence of an RNA-

thermometer-like mechanism in eukaryotes, involved in this case in preventing degradation rather than translation activation (Wan *et al.*, 2012). Thus, the subset of stable mRNA regions characterized *in vivo* provides promising candidates for novel functional RNA structures, few of which were already tested and shown to effect protein synthesis (Rouskin *et al.*, 2014) while many others remain to be characterized.

In vivo studies on eukaryotic cells showed how stable RNA structures at the 5′ of splicing sites are generally avoided, since, if present, they inhibit the first step of splicing (Ding *et al.*, 2014, Wan *et al.*, 2014) while intersection of RNA folding dataset with iCLIP experiments, identifying target sites of RNA binding proteins (RBP), revealed structural features of consensus sequence specific for each RBP target site (Incarnato *et al.*, 2014, Wan *et al.*, 2014, Spitale *et al.*, 2015)

Finally, studies on a human parent–offspring trio characterized single-nucleotide polymorphism associated with changes in RNA structure (riboSNitches). About 1907 sites resulted in structural switch between the first and second generation, 211 of which were associated with changes in gene expression. Additionally, riboSNitches were absent at the level of specific sites (3'UTRs, predicted miRNAtarget sites and RBP binding sites), suggesting that they exhibit a detrimental effect, as well as other single-nucleotide polymorphism involved in disease-onset (Wan *et al.*, 2014).

In summary, all the latest findings on the mRNA structure *in vivo* evidence the extensive role of mRNA secondary structure in regulating various stages of cell life. However, it is surprising to notice that all these studies focused on eukaryotes, leaving a big gap of knowledge about bacterial organism, one of the most used models in molecular biology.

1.4 Assessing translation and mRNA structure of the whole transcriptome by means of Next-Generation Sequencing

The power of next-generation sequencing relies on the possibility to sequence in parallel millions of DNA or RNA fragments (Reuter *et al.*, 2015). This highly informative technology can be applied to any kind of traditional, nucleotide sequence-based assay, allowing extending its potential from single-molecule study to a more global, cell-wide approach with systemic view. Thus, NGS-based techniques are within the best tool to study complex network of cell regulation, among which translation is probably the most sensitive and flexible layer.

Particularly, the combination of different, complementary sequencing techniques can represent a powerful toolbox to correctly position the molecular pieces of the cell puzzle and define their interplay. In the last 10 years, many techniques have been developed to assess *in vivo* RNA abundance, ribosomal density, protein-nucleic acid interaction, and mRNA secondary structure. The invention and conjugation of these approaches contributed enormously to our knowledge of cell biology.

The relationship between RNA secondary structure and translation are within the focus of this work, thus here we want to compare different techniques to assess RNA folding (Wan *et al.*, 2011, Kwok *et al.*, 2015) and progression of the ribosomes along the transcript (Brar & Weissman, 2015), in the living cell.

RNA structure can be probed by means of chemicals or enzymes that modify the single- or double-stranded ribonucleotide bases, followed by detection of the modification (Ehresmann *et al.*, 1987). Ribonucleases (RNases) recognize specific ss-regions or ds-regions of RNA and cleave the RNA backbone at those sites. While the great advantage is to have a complementary ss-/ds-stranded information, the large physical size of the RNases prevents them from reaching all the bases and their membrane-impermeable nature limits their use to *in vitro* applications (Ehresmann *et al.*, 1987). In contrast, chemical probes are cell-permeable and their size is smaller, however they have a cytotoxic effect, they are unable to modify all four nucleotides and are restricted to unpaired bases (Ehresmann *et al.*, 1987). Up to date, NGS-coupled RNA structure probing techniques took advantage of enzymes ds-specific RNase V1 (Kertesz *et al.*, 2010, Li *et al.*, 2012a, Li *et al.*, 2012b) and ss-specific Nuclease S1 (Kertesz *et al.*, 2010), Nuclease P1 (Underwood *et al.*, 2010), RNase I (Zheng *et al.*, 2010, Li *et al.*, 2012a, Li *et al.*, 2012b) and of small chemicals (all ss-specific) dimethyl sulfate (DMS), targeting only A and C (Ding *et al.*, 2014, Incarnato *et al.*, 2014, Rouskin *et al.*, 2014) (Talkish *et al.*, 2014), N-cyclohexyl-N'-(2-morpholinoethyl)carbodiimide metho-p-toluenesulfonate (CMCT) targeting U and G (Incarnato *et al.*, 2014), 1-methyl-7-nitroisatoic anhydride (1M7) (Siegfried *et al.*, 2014) and 2-methylnicotinic acid imidazolide (NAI) or derivatives (Spitale *et al.*, 2015) interacting with all the four bases.

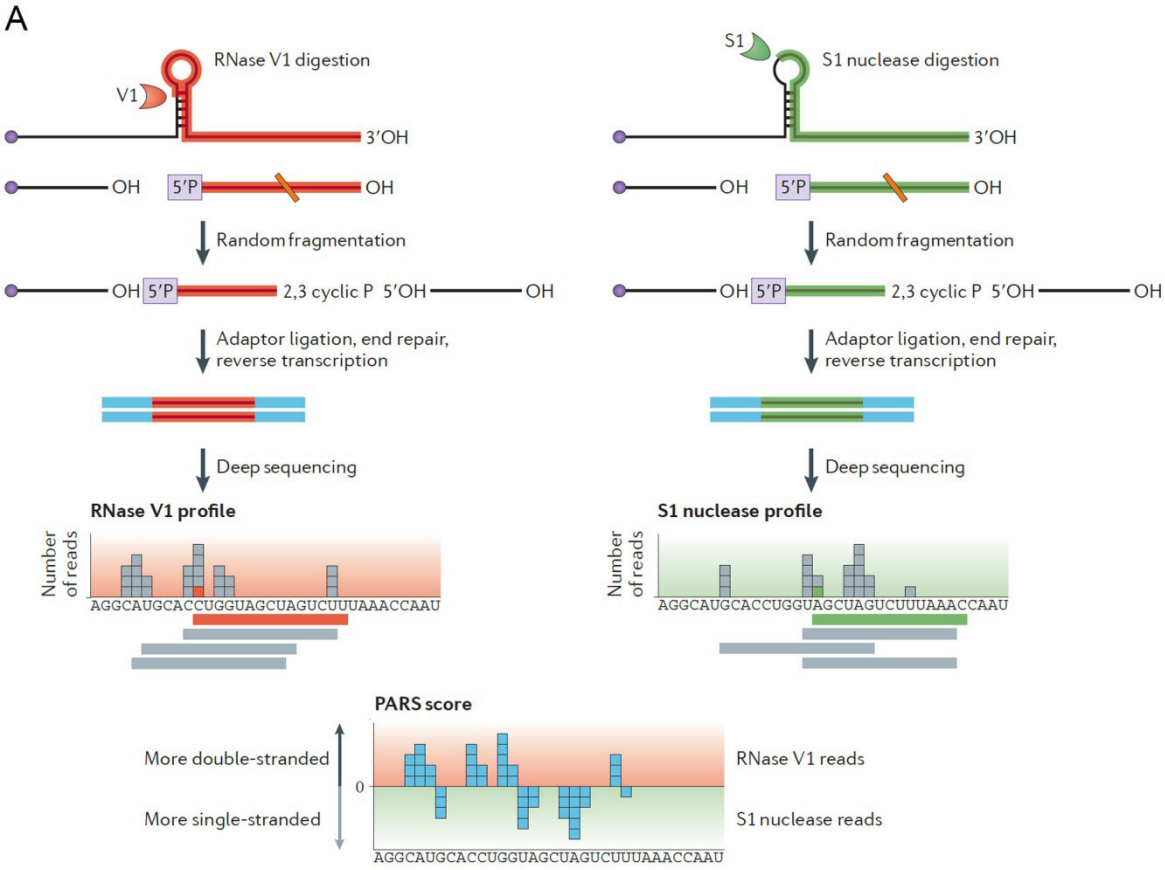
To exemplify the differences between the enzyme- and chemical-based experiments, we will describe the main feature of the first published method using enzymes (Parallel Analysis of RNA Secondary structure) and the most-used-chemical approach (DMS-Seq).

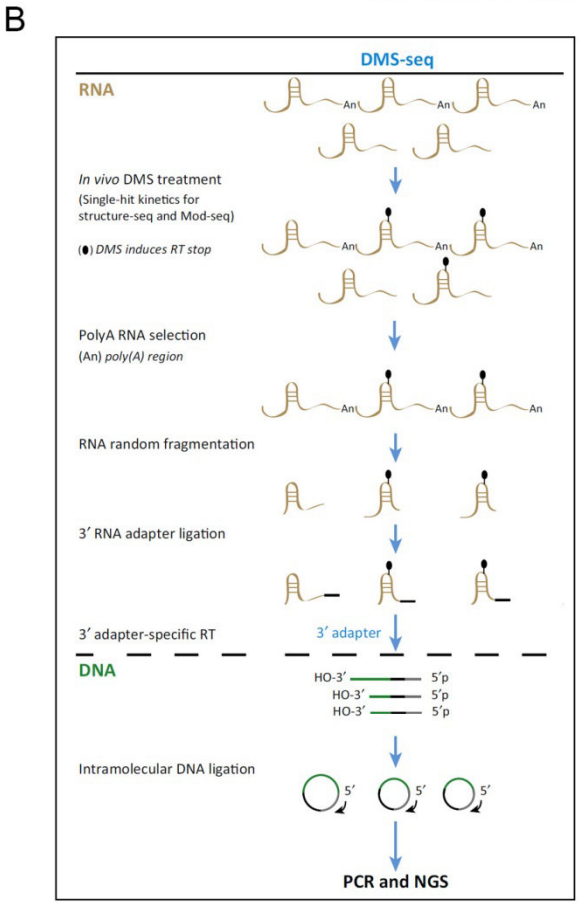
Parallel Analysis of RNA Structure (PARS), gives information on the intrinsic propensity of each RNA nucleotide to be involved in a double-stranded structure or in a linear single-strand (Kertesz *et al.*, 2010). Total RNA is isolated and enriched in protein(ribosome)-free mRNA, *in vitro*, which after refolding is digested either with a ss-specific nuclease S1 or with a ds-specific RNase V1, resulting in a 5′P leaving group. The enzymatically probed RNA is then fragmented. As enzymatic cleavage products contain 5′P, whereas fragmentation and degradation products would have 5′OH, only true structured sites can be ligated to the adaptors and reverse transcribed. The cDNA is then subjected to massively parallel sequencing, the first nucleotide of each mapped read bears the structural information (Fig. 1.2 A). The structural propensity of each nucleotide is then computed, calculating the logarithmic ratio between the ds- and ss-information, i.e. between the sequencing reads covering this nucleotide in the double- and single-stranded libraries; this value is termed "PARS score".

In DMS-seq (Rouskin *et al.*, 2014), and with small differences elsewhere (Ding *et al.*, 2014), RNA is treated with DMS *in vivo* and poly(A) RNA is selected. RNA is randomly fragmented to generate smaller sequences to which a 3' RNA adapter is ligated. At this point, 3' adapter-specific reverse transcription is performed to generate cDNA reads which stop at the DMS-modified sites. Intramolecular circular DNA ligation is then performed, followed by PCR and NGS (Fig. 1.2 B).

The two approaches contributed with equal power to the investigation of mRNA secondary structure role in yeast, plant, mouse and human cells. However, they both have pros and cons one should consider before data processing. These issues will be critically addressed in this work (see chapter 2.2).

To address the regulatory role of mRNA structure in *E. coli*, the mRNA structurome-assessing techniques can be coupled with other NGS-based techniques, which give information, for example, on mRNA abundance or ribosome position. RNA-Seq protocol involves the isolation of total RNA from cells, which is then processed for deep-sequencing, providing a snapshot of all RNAs in the cell; it can be used for quantifying RNA in the cell (Mortazavi *et al.*, 2008).


A
RNase V1 digestion
V1
3′OH
OH
5′P
OH
Random fragmentation
OH 5′P 2,3 cyclic P 5′OH OH
Adaptor ligation, end repair, reverse transcription
Deep sequencing
RNase V1 profile
Number of reads
AGGCAUGCACCUGGUAGCUAGUCUUUAAACCAAU
S1
S1 nuclease digestion
3′OH
OH
5′P
OH
Random fragmentation
OH 5′P 2,3 cyclic P 5′OH OH
Adaptor ligation, end repair, reverse transcription
Deep sequencing
S1 nuclease profile
Number of reads
AGGCAUGCACCUGGUAGCUAGUCUUUAAACCAAU
PARS score
More double-stranded
More single-stranded
0
RNase V1 reads
S1 nuclease reads
AGGCAUGCACCUGGUAGCUAGUCUUUAAACCAAU


B
DMS-seq
RNA
An
In vivo DMS treatment
(Single-hit kinetics for structure-seq and Mod-seq)
(●) DMS induces RT stop
PolyA RNA selection
(An) poly(A) region
RNA random fragmentation
3′ RNA adapter ligation
3′ adapter-specific RT
3′ adapter
DNA
HO-3′ 5′p
Intramolecular DNA ligation
5′
PCR and NGS

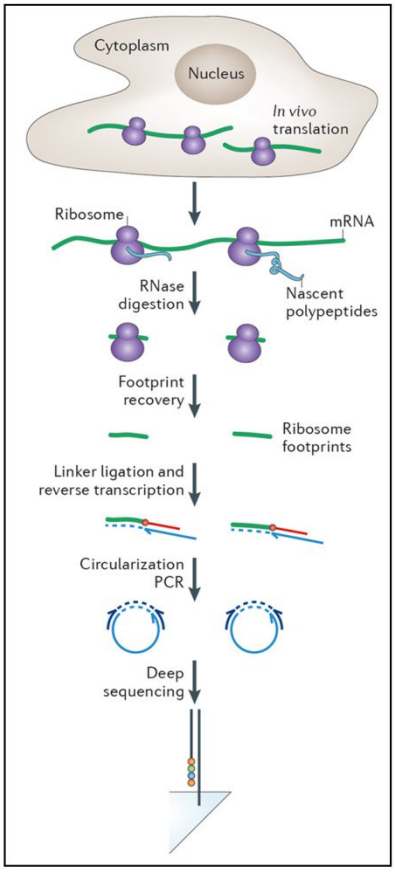

C
Cytoplasm
Nucleus
In vivo translation
Ribosome
mRNA
RNase digestion
Nascent polypeptides
Footprint recovery
Ribosome footprints
Linker ligation and reverse transcription
Circularization PCR
Deep sequencing

Figure 1.2 | Experimental workflow for PARS, DMS-Seq and Ribosome Profiling. **(A)** In PARS, poly(A) selected RNA is folded *in vitro* and incubated with either RNase V1 or S1 nuclease to probe for double- and single-stranded regions, respectively. RNases leaves a 5′P group that can be directly ligated to adaptors, after a random fragmentation step. The cDNA library is sequenced using high-throughput sequencing and the resulting reads are mapped to the genome. The logarithmic ratio of ds- versus ss-mapped counts is termed PARS score, whereby a positive or a negative PARS score indicates that a base is double-stranded or single-stranded, respectively (adopted from (Wan *et al.*, 2011)). **(B)** For DMS-seq, RNA is modified with DMS *in vivo*. poly(A) RNA is selectively subjected to random fragmentation to generate RNA fragments. 3' adapter is ligated to the RNA fragments and reverse transcription generates the cDNA. Intramolecular circular DNA ligation makes the library suitable for PCR and NGS. (Adopted from (Kwok *et al.*, 2015)). **(C)** RNase digestion of translating polysomes yields ribosome-protected mRNA fragments (RPFs), which are recovered and converted into a cDNA library through ligation of a linker followed by reverse transcription and circularization PCR. cDNA libraries are then analysed by deep sequencing. (Adopted from (Ingolia, 2014))

Ribosome profiling is based on the fact that during translation the ribosome protects certain mRNA fragment from nucleolytic digestion (Wolin & Walter, 1988), which on the contrary degrades "naked" mRNA separating the ribosome-protected fragments (RPFs) (Fig. 1.2C). The purified RPFs are ligated to adaptors, followed by a reverse transcription and subsequent deep-sequencing of the cDNA. The generated reads are aligned to the reference genome revealing the position of translating ribosomes with a nucleotide resolution (Ingolia *et al.*, 2009).

NGS-based techniques gained tremendous deep into system biology and the integration of multiple deep-sequencing derived databases, included the one here described, allowed to collect valuable knowledge about the regulation of cellular processes. Thus, it is attractive to use this combined approach to gain new insights on regulation of protein synthesis.

1.5 Aim of the thesis

Recently, many studies focused on determining mRNA secondary structure landscape in vivo, directly in the cell. All these studies focused on eukaryotic cells, leaving a large gap of knowledge on the bacterial world. The findings published up to date show the central role of mRNA folding in the cell life, suggesting that additional layer of yet-to-be-unraveled information is hidden in the nucleotide sequence, which determines the gene expression beyond the genetic code.

The aim of this work is to determine the impact of mRNA secondary structure on translation in the bacterial model organism *Escherichia coli*. Combining the power of NGS-based PARS analysis of mRNA structure with RNA-Seq, we aimed also to explore the role of RNA folding on RNA stability. In addition, coupling with ribosomal profiling will reveal the effect of duplexes in the coding sequence on ribosomal pausing and determine the role of unstructured or structured regions in regulating translational initiation efficiency and translation termination fidelity.

1.6 Structure of the thesis

The current dissertation is organized in three chapters.

The experimental work is presented in the first chapter of the results section. The work was performed in close collaboration with Alexander Bartholomäus and recently published.

"Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function"

Cristian Del Campo, Alexander Bartholomäus, Ivan Fedyunin, Zoya Ignatova

PLoS Genetics, 2015

Contribution: CDC conceived and designed the mRNA-structure experiments on single-gene level and on global cell-wide level. Figures 1, 2, 3, 4c-d, 5, S1, S2, S3, S5, S6, S7b, S8b arisen from this analysis. IF produced the ribosome profiling. Together with AF, who is by training bioinformatician and run the computational analysis, CDC analyzed the results of all deep sequencing experiments.

The following two chapters constitute a critical assessment of the reported literature about ribosome profiling and NGS-based, RNA structure determining techniques to highlight their biases and potentiality, with the aim to provide a useful tool for a correct choice of the relative technique, in regards of the needed application. These two works were also recently published.

"Probing dimensionality beyond the linear sequence of mRNA"

Cristian Del Campo, Zoya Ignatova

Current Genetics, 2015

Mapping the non-standardized biases of ribosome profiling

Alexander Bartholomäus, Cristian Del Campo and Zoya Ignatova

Biochemistry, 2016

Contribution: CDC wrote the experimental part of the manuscript relative to the polysome isolation. AB wrote the computational part of the manuscript on data analysis.

2. RESULTS

2.1 Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function

2.1.1 Adopting Parallel Analysis of RNA Secondary Structure (PARS) in *E. coli*

The PARS protocol is based on the partial, specific cleavage of structured or unstructured, cell-extracted and *in-vitro* refolded RNAs by RNases targeting double-stranded (RNase V1) or single-stranded (nuclease S1) regions, respectively. The digestion of both enzymes generates a 5'-phosphorylated (5'P) and 3'-hydroxyl (3'OH) fragment that can be ligated immediately to a 5' adapter, in order to select only products derived from the RNases cleavage, and not from a random fragmentation. Indeed, after the digestion, a further treatment with sodium hydroxide will randomly shorten the RNase-produced fragments, in order to enrich them in the length range of 50-200 nt, suitable for deep-sequencing. Since randomly fragmented products contain 5′-hydroxyl (5′OH) groups, they will not be ligated to the 5' adapter. Subsequent 3′ adapter ligation, RT and PCR steps result in the unique amplification of nuclease-cleaved fragments, since they contain both 5′ and 3′ adapters. The sequencing reads generated by NGS are mapped against the genome or transcriptome of the target organism and the propensity of each nucleotide to be structured, defined as "PARS score", is then calculated.

During RNA structure probing, the RNases should cleave with single-hit kinetics, allowing on average a single cut per molecule, so that potential conformational changes arising from the first enzymatic cleavage are not additionally processed by the RNases. To define the optimal enzyme concentration, we digested 2 µg of total RNA with different amount of RNase V1 and nuclease S1. Total RNA of *E. coli* cells grown till exponential phase was isolated and 2 µg were completely denatured, cooled on ice and *in-vitro* refolded, slowly increasing the temperature from 4 °C to 23 °C, in RNA structure buffer at pH 7. Although the pH could be adjusted to the optimal pH for the enzymes activity, this would not be a correct approach. Indeed, changes in the pH can induce a different RNA refolding, giving different structural

information depending on the used enzyme. Thus, RNA is refolded at a constant pH of 7 for both enzymes, which pH represents the closest value to the physiological cytoplasmic environment, where mRNA translation generally takes place. When RNA was digested using the enzyme amount and incubation time published elsewhere (Kertesz *et al.*, 2010, Wan *et al.*, 2013), the reaction resulted in a complete degradation of the sample (Fig. 2.1 A).

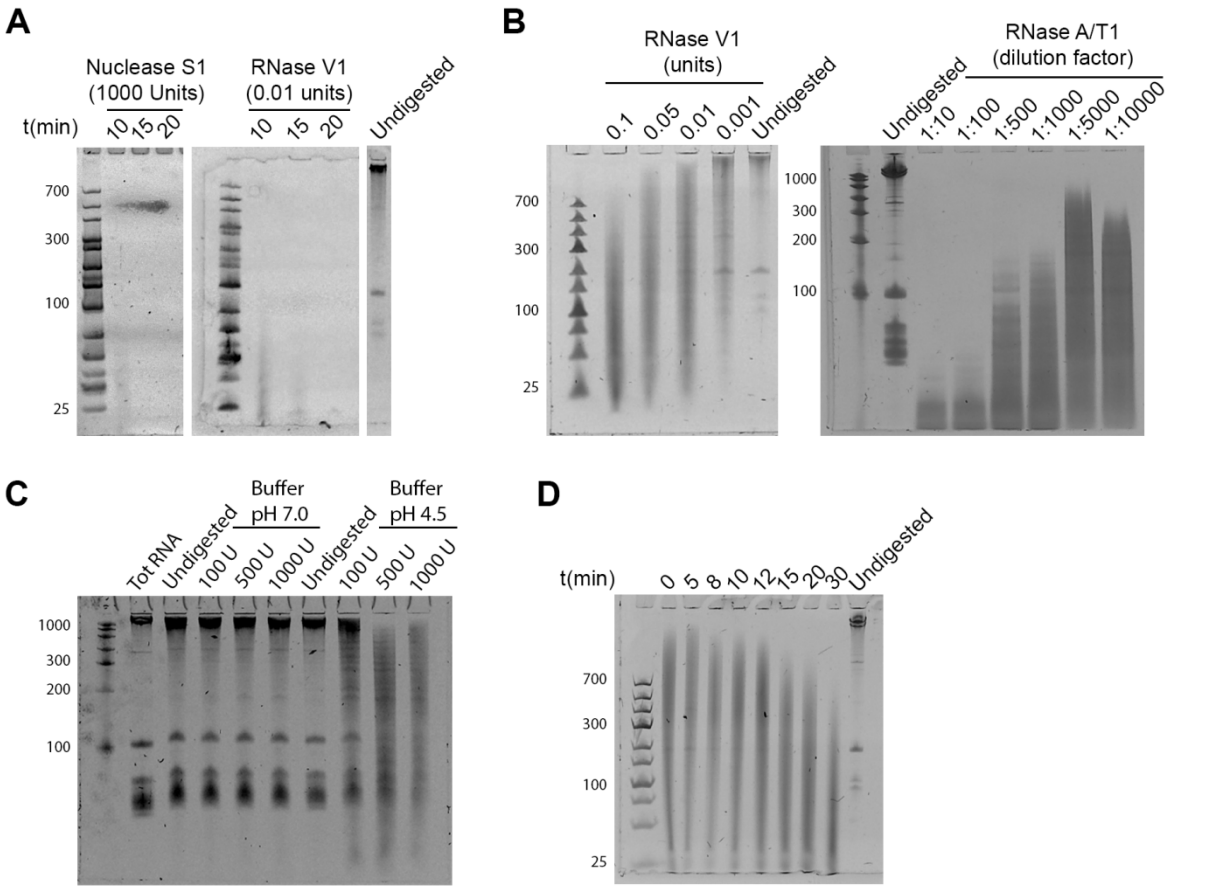


Figure 2.1 | Testing mRNA enzymatic digestion to identify best PARS conditions. (**A**) Denaturing PAGE of total RNA from *E. coli* digested with 1000 U of Nuclease S1 or 0.01 U of RNase V1 for 10, 15 or 20 min at 23 °C, as was done by Kertesz *et al.*, 2010. The RNA is barely detectable, ie. mostly degraded. (**B**) Denaturing PAGE of total RNA digested with 0.1, 0.05, 0.01 and 0.001 U of RNase V1 or with 1:10, 1:100, 1:500, 1:1000, 1:5000 and 1:10000 dilution of a mixture of RNase A and RNase T1 (Thermo Scientific), incubated for only 1 minute at 23°C. 0.05 U of RNase V1 and a 1:5000 dilution of mixed RNase A and T1 gave the best distributed fragmentation, without loss of small fragments. (**C**) Digestion of total RNA with 100, 500 and 1000 U of Nuclease S1 at pH 7.0 or pH 4.5, incubated for 1 min at 23 °C and loaded on a denaturing PAGE. (**D**) Random alkaline fragmentation of RNase A/T1 digested sample (in optimized conditions), incubated for 0, 5, 8, 10, 12, 15, 20 and 30 minutes at 95°C in alkaline fragmentation buffer.

In a following publication, testing the RNA folding energies at different temperatures with a similar approach (Wan *et al.*, 2012), the authors used a similar amount of RNase V1 but a much shorter incubation time. Thus, we tested this enzymes concentration, incubating the sample for only 1 minute and stopped the digestion by a phenol/chloroform extraction. In these conditions, the digestion generates a smear of shorter fragments along the whole gel, indicating that the cleavage took place but didn't completely degrade the RNA (Fig 1.1 B, left panel). A concentration of 0.05 units of RNase V1 gave the best digestion profile and was chosen for the following experiments.

While RNase V1 optimal pH is about 7, Nuclease S1 shows an optimal activity at pH 4.0 (Vogt, 1973). This probably explains the need of a very high amount of enzyme used in the original protocol (Kertesz *et al.*, 2010). Indeed, the RNA digestion by Nuclease S1 showed almost no activity of the enzyme at pH 7. Even the same units of enzyme as used in the original protocol (1000 units) were not able to cleave efficiently, unless the buffer with pH 4.5 was used for the reaction (Fig. 1.1 C).
Although the use of Nuclease S1 is advantageous because it generates 5'-phosphorylated fragments, we reasoned that an extremely high concentration of this nuclease and reaction at pH far from the optimal pH could create artifacts. In fact, it has been shown before that double-stranded DNA, double-stranded RNA, and DNA-RNA hybrids are relatively resistant to the enzyme; double-stranded fragments are completely digested by the nuclease S1 if they are exposed to large amounts of the enzyme (Green & Sambrook, 2012).

A first alternative enzyme to S1 nuclease is P1 nuclease, whose specificity towards single-stranded RNA was already exploited in Frag-SEQ approach, a deep-sequencing method to identify unfolded RNA regions of the whole transcriptome (Underwood *et al.*, 2010). As well as for S1 nuclease, P1 also leaves a 5'-phosphate, which in turn is useful to produce the sequencing library. However, similarly to S1, P1 has the activity optimum pH equal to 4.5 (Romier *et al.*, 1998).
A mixture of RNase T1 and RNAse A represents an additional alternative to S1 nuclease, since they are more active towards RNA in general (S1 cleaves DNA 5 times more efficiently than RNA), specific against single-stranded RNA and have an optimal pH of activity around 7 (RNase A pH=7 (Tripathy *et al.*, 2013); RNase T1 pH=7.5 (Sato & Egami, 1957)). The

ribonuclease T1 targets specifically the 3'-phosphate of guanosines, while RNase A the 3'-phosphate of pyrimidines (uracyl and cytosine) (Nichols & Yue, 2008). Furthermore, the fragments originated from these ribonucleases lack the 5' phosphorylation. Nevertheless, due to the high specificity to ss-RNA at physiological pH, we decided to use RNases A and T1 and complemented the missing 5'-P with an additional step of phosphorylation, following the cleavage.

To identify the RNase A/T1 mix concentration for single-hit kinetics digestion, a range of different enzymes dilutions was tested and a 1:5000 dilution was chosen to perform the following experiments (Fig. 1.1 B).
To enrich the digested mRNA in 50-200 nt long fragments, suitable for deep-sequencing, the sample was subjected to random alkaline fragmentation. The optimal incubation time was chosen by testing different time points and visualizing the enrichment in the specified nucleotide range on a denaturing gel (Fig. 1.1 D). An incubation time of 12 minutes resulted in the highest fragment enrichment and was consequently adopted.

2.1.2 PARS validation

To assess the intrinsic propensity of the *E. coli* transcriptome to partition in secondary structures, we isolated total RNA (i.e. in absence of proteins and ribosomes) from exponentially growing *E. coli* culture, enriched in mRNA through depletion of small and ribosomal RNA and subjected it to PARS (Fig. 2.2). The mRNA was completely denatured, cooled on ice and slowly refolded to 23 °C, which was the incubation temperature of the enzymatic cleavage, when the optimized concentration of RNase V1 or RNase A/T1 mix was added. The reaction time was set to 1 minute and the digestion was stopped with a phenol/chloroform extraction. The single-stranded digested sample was phosphorylated at the 5'-end and later both the RNase-cleaved samples were randomly fragmented by alkaline hydrolysis, in order to enrich the digestion products in a range between 50-200 nt, suitable for deep-sequencing.
RNA fragments in the mentioned range were selected from gel for preparation of a high-throughput sequencing library, following a slightly modified Illumina TruSeq Small RNA protocol (see Materials and Methods). Importantly and differently from the classic method of

library generation, the 5'-adapter was firstly ligated to the fragments, followed by ligation of the 3'-adapter, in order to preserve the RNA structure-dependent cleavage information.

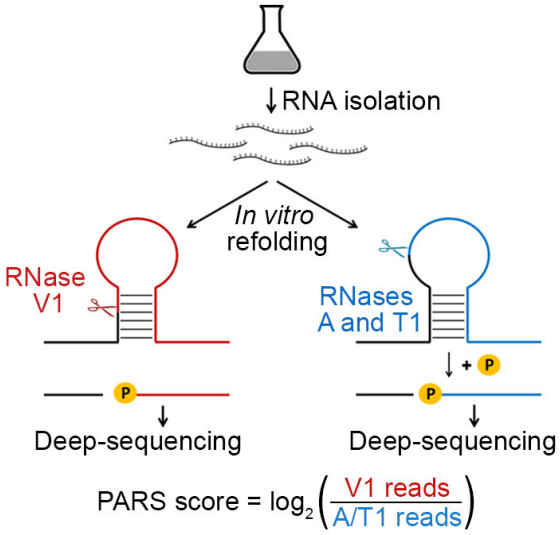


Figure 2.2 | Overview of modified PARS approach. RNase V1 cleaves double-stranded RNA and combination of RNases A/T1 the single stranded RNA with optimal activities at physiological pH (7.0). RNAse A/T1 usage requires an additional phosphorylation step prior to library generation.

The cDNA libraries were sequenced on the Illumina HiSeq2000 sequencing machine, resulting in ~50 million reads per sample (with a range of ~46.6 – ~53.0 million reads/sample), of which ~18 million reads/sample were uniquely mapped to the *E. coli* MG1655 genome (range: ~17.5 – ~19.4 million reads/sample). Two biological replicates for each sample were produced. The sequences obtained from the 2 independent experiments are highly reproducible across replicates, as results from the Pearson correlation between the log2 of read coverage for each transcript (Pearson coefficients R = 0.96 and R = 0.95 for V1 and A/T1 digestions, respectively, Fig. 2.3 A,B). Good reproducibility was also evident on single-gene level (Fig. 2.3 D).

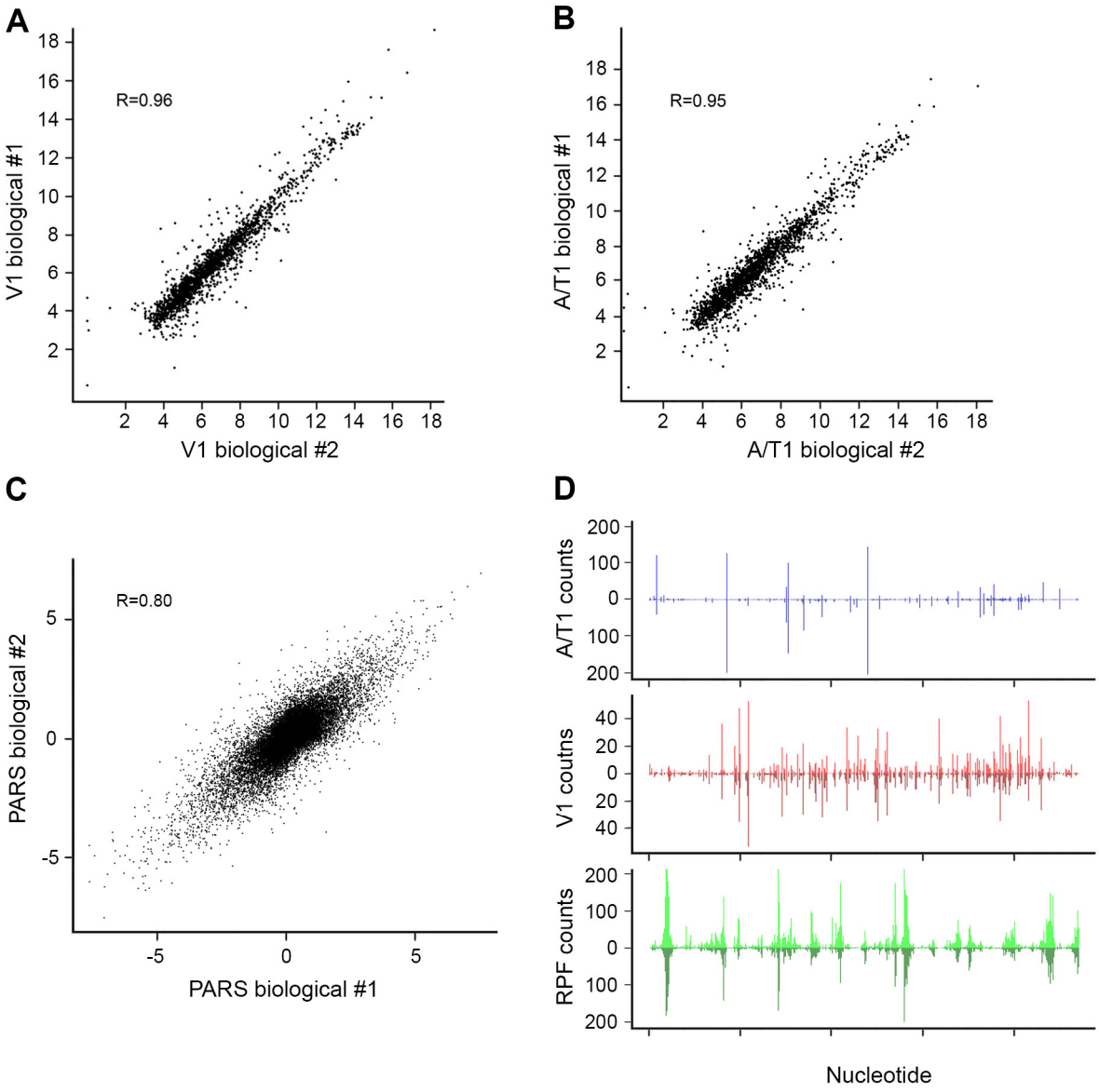


Figure 2.3 | Reproducibility of the PARS results. (**A,B**) Pearson correlation between the log2 of read coverage for each transcript with load >1 (see panel E) digested by RNase V1 (**A**) or RNases A/T1 (**B**) in the two biological replicates. (**C**) Reproducibility of the PARS score for each nucleotide. To reduce the crowding in the plot, only 200000 randomly selected nucleotides were plotted. (**D**) Single gene example on the reproducibility of the various sequencing data.

To assess the involvement into secondary structure of each nucleotide, a "structural score", named PARS score, is calculated for each position. The PARS score is defined as the logarithmic ratio of the number of RNase V1-derived reads divided by the number of RNase A/T1-derived reads, as also reported in the equation in Fig. 2.2.

Since the mRNA structure are folded with a different degrees of stability and they are able to rapidly reshape (Mahen *et al.*, 2010), switching from a close to an open conformation and

vice versa, the digestion with the 2 different RNases can potentially cleave the same nucleotide, giving apparent contradictory information. The PARS score takes in account this possibility and by calculating the ratio of the fragments number produced by the double-stranded versus single-stranded specific RNases for each position allow to define how probable is to find that position in a folded or unfolded status. Thus, positive PARS score indicates the propensity of a nucleotide to be double-stranded, while negative PARS score the propensity to be single-stranded. To check whether the PARS score was reproducible, implicating that the ratio of the RNases cleavages is conserved in the biological replicates, we calculated the Pearson correlation between the PARS score of each nucleotide in the replicates. The correlation coefficient R=0.80 indicates a good reproducibility on a transcriptome-wide level (Fig. 2.3 C).

In order to select transcripts having an average read coverage sufficient for reliable further analysis, we calculated the transcript load, defined as the sum of combined PARS readouts of the biological replicates per transcript divided by the effective transcript length (that is the annotated transcript length minus the number of unmappable nucleotides) (Kertesz *et al.*, 2010). At a selected threshold of 1.0, indicating an average coverage of one read per nucleotide (Kertesz *et al.*, 2010) (Fig. 2.4 A), we obtained structural information for ~900,000 nucleotides covering 2,536 *E. coli* genes. The results from PARS are in excellent agreement with known RNA structures and match four experimentally validated RNA structures (Fig 2.4 B; Fig. 7.1), including also the whole 16S rRNA. Furthermore, we performed additional independent experimental validation of the ppiC transcript; the PARS values recapitulate the results from orthogonal structural probing of ppiC (Fig. 2.4 C).

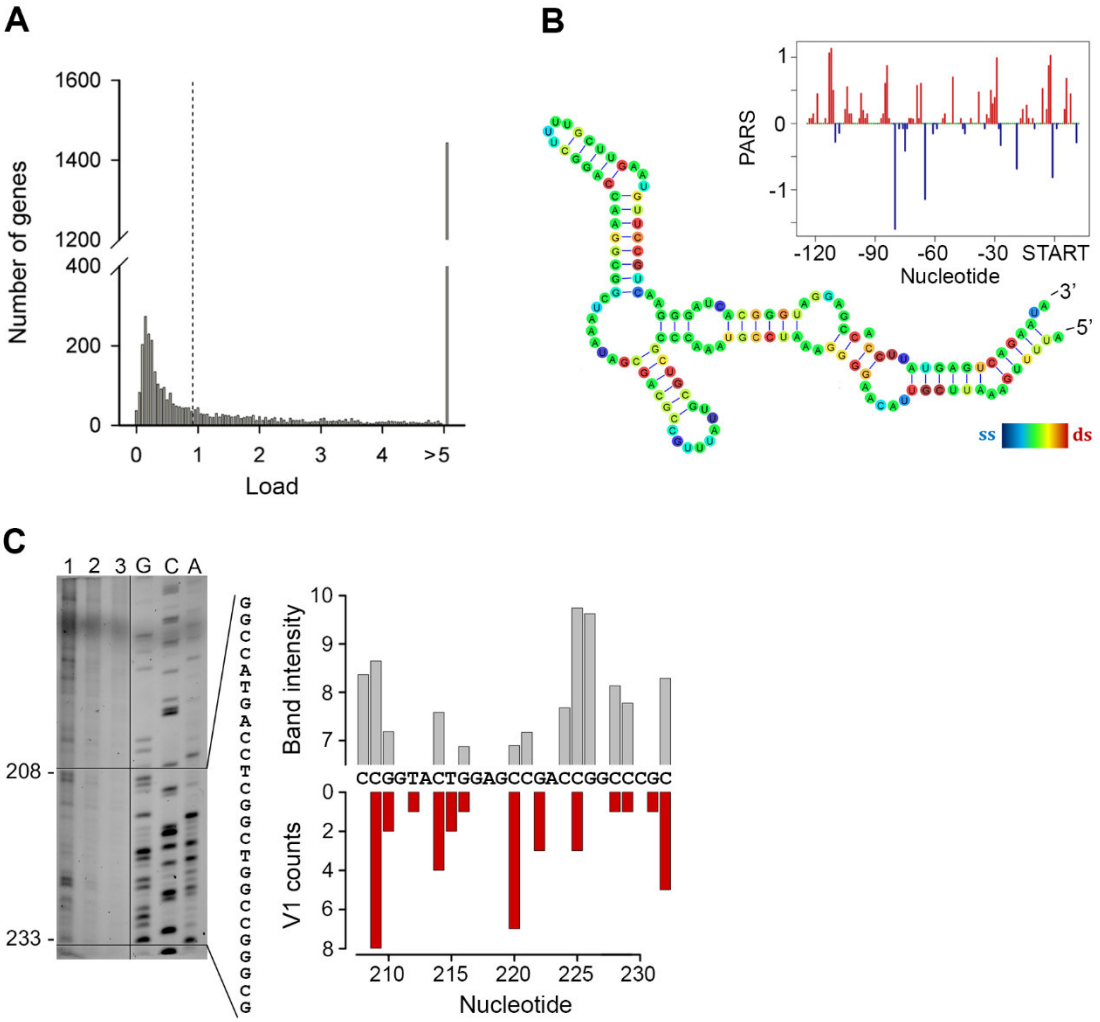


Figure 2.4 | Reproducibility of the PARS results. (**A**) Number of transcripts as a function of the transcript load (Schmidt *et al.*, 1995, Kertesz *et al.*, 2010), i.e. the PARS readouts from the merged biological replicates divided by the effective transcript length (that is the annotated transcript length minus the number of unmappable nucleotides). A threshold of 1 (vertical dashed line) was selected as also used previously for yeast PARS data (Kertesz *et al.*, 2010). (**B**) The PARS score of the *rpoS* leader sequence (inset) was overlaid with the experimentally determined structure (Soper & Woodson, 2008). Double-stranded nucleotides with positive PARS score are colored red, single-stranded nucleotides with negative PARS score–blue, nucleotides with missing PARS score or equal to zero– green. The color intensity of the *rpoS* nucleotides reflects the PARS scores (rainbow legend). (**C**) Footprint analysis of fluorescently-labeled *ppiC* mRNA digested with 0.05 U (lane 1) or 0.01 U (lane 2) RNase V1 compared to undigested mRNA (lane 3). The RNase V1-digestion pattern mirrors the V1 sequencing counts. The graphic insert represents an exemplary comparison between the intensity of the bands (gray bars) from designated area from the gel (horizontal lines between 207–234 nt) and the counts for the same gene obtained from the deep sequencing of the RNase V1 digested sample. The sequence derived from the Sanger sequencing (included next to the gel) is complementary to that in the plot.

2.1.3 PARS reveals globally conserved structural features among *E. coli* transcripts

To highlight the presence of conserved structural motives among the whole transcriptome, we performed a metagene analysis of all the transcripts aligned either at their start or stop codons. *E. coli* CDSs have a propensity to form double-stranded structure to a level that is similar to the structure propensity of the 5'- and 3'-untranslated regions (UTRs) (Fig. 2.5). This global trend is different from that in eukaryotic organisms. In yeast, UTRs are less structured than CDSs (Kertesz *et al.*, 2010). Conversely, in metazoans (Li *et al.*, 2012a) and humans (Wan *et al.*, 2014) UTRs are, on average, more structured than coding regions. A well-defined periodic pattern is present only in the CDSs but not in the 5' and 3'UTRs as detected by discrete Fourier transform (Fig. 7.2 A) with first nucleotide being the most structured (Fig. 7.2 B). Three nucleotide periodicity is also detected in yeast (Kertesz *et al.*, 2010), *A. thaliana* (Ding *et al.*, 2014), mouse (Incarnato *et al.*, 2014) and human (Wan *et al.*, 2014) and is intrinsic to the structure of the genetic code (see the periodic pattern of the GC content, Fig. 2.5), consistent with prior computational predictions for various genomes (Shabalina *et al.*, 2006). Although this periodicity can also come from a degradation pattern (Pelechano *et al.*, 2015), this would be responsible only for the 4-5% of the total transcripts, indicating that a trinucleotide structural frequency is anyway intrinsic in the codon code.

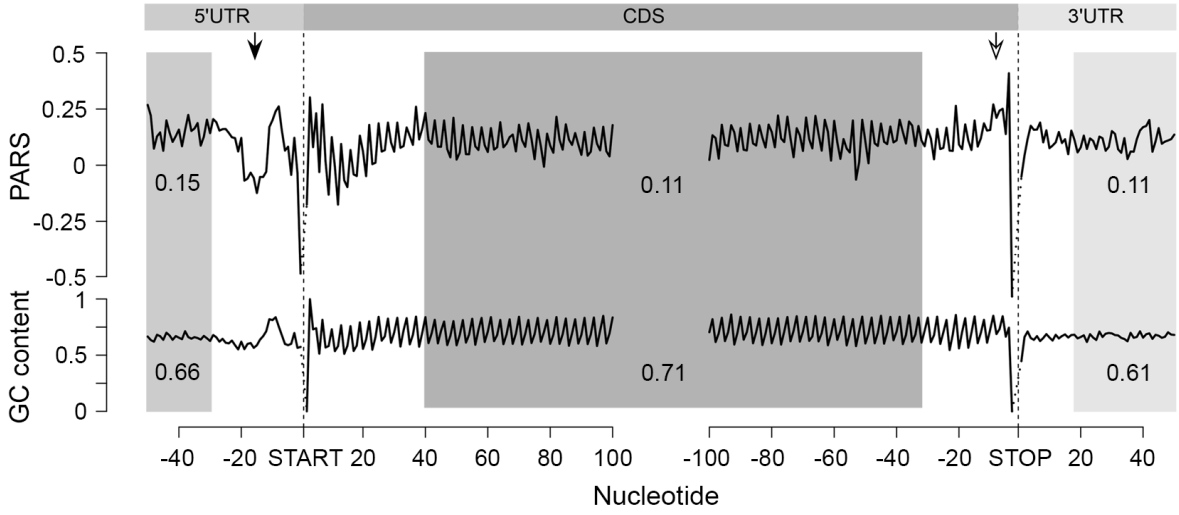


Figure 2.5 | Metagene analysis of protein-coding transcripts. Average PARS score for each nucleotide (top) and GC content (bottom) across the 5'UTRs, CDS and 3'UTRs of all protein-coding transcripts, aligned at the start or stop codon, respectively. For the shaded areas, the average PARS scores or GC content is calculated; thus note the deviations from the total GC content of 51% in *E. coli*. Unstructured region upstream of the start codon and structured sequence preceding the stop codon are marked by arrows with filled and open arrow heads, respectively.

We noticed, however, that in some regions the mRNA structure deviates from the nucleotide content, e.g. a uniform unstructured region around 20 nt upstream of the initiation start and more structured region upstream of the termination codon (Fig. 2.5). These positions may provide candidate sites for functional conformation of mRNA *in vivo* and we address their role below.

The region 10-30 nt downstream of the initiation was also less structured than the average PARS score of the CDS (Fig. 2.5). Less structured regions at the 5' start of the CDSs facilitate initiation and general gene expression (Bentele *et al.*, 2013, Goodman *et al.*, 2013), a trend which is also present in the human (Wan *et al.*, 2014) but not in the yeast (Kertesz *et al.*, 2010) transcriptome.

2.1.4 Intrinsic secondary structure propensity of the CDS influences elongation only locally in some genes

We next asked whether the intrinsic secondary structure propensity of the CDS influences the translation (elongation) efficiency and correlates with mRNA abundance in the cell. We complemented the PARS analysis with ribosome profiling which captures the positions of translating ribosomes with nucleotide resolution (Ingolia *et al.*, 2009) which showed high reproducibility between biological replicates on a global (Fig. 7.3) and single gene level (Fig. 2.3 D). We hypothesized that a persisting mRNA structure would induce ribosomal pausing which would be detected by enrichment of ribosome-protected fragments (RPFs) upstream of an mRNA structured stretch. A structured stretch was defined when 6 nt within a window of 10 nt show a positive PARS score (for details see Methods section and Fig. 7.4 A). In total, within the CDSs we extracted 908 stretches with high structure propensity *in vitro*. For the majority of the structured stretches we did not detect an accumulation of the RPF upstream of them (Fig. 2.6 A and 7.4 A) suggesting that the majority of these structures may not persist *in vivo* and do not influence the elongating ribosomes that is consistent with the observation in yeast and mammalian cells (Rouskin *et al.*, 2014). Nonetheless, a sizeable fraction of structured sites in the CDS (above the 80th percentile, 87 positions) caused ribosomal pausing, i.e. $L_1 > L_2$ (Eq. 2 and 3; Fig. 2.6 A). Along with the genes with previously validated structures (Fig. 2.6 B and Table 7.1), our analysis revealed some promising candidates for novel

functional RNA structures (Fig. 2.6 C and Table 7.1). One of the genes, *deaD*, encodes a DEAD-box RNA helicase that functions in large ribosomal subunit assembly (Iost & Dreyfus, 2006) and RNA degradation under cold shock (Resch *et al.*, 2010). Contrary to the prevailing views for DeaD function at only low temperature, recent evidence describes its expression over a broad temperature range but with large variation in expression level (Vakulskas *et al.*, 2014). It is tempting to speculate that the newly identified persistent structure in *deaD* (Fig. 2.6 C) may regulate its expression level at different temperatures through a structure-induced translational pausing.

Slow-translated regions, mostly formed by clustering of suboptimal codons, are enriched in *E. coli* membrane proteins at the beginning of their transmembrane domains (Fluman *et al.*, 2014). Similarly to yeast, these regions may promote interaction with the signal recognition particle (Pechmann *et al.*, 2014) and thus facilitate membrane targeting and translocation. Since a large fraction of the identified structural sites that correlated with accumulated RPF reads were in membrane proteins (Table 7.1), we analyzed the distance between the pausing positions and start of the transmembrane domains. The majority of the pausing sites were within 11 to 80 amino acids downstream of the membrane domains (Figure 2.6 D). Strikingly, this distance interval closely resembles the 30-72 amino acid span needed to exit the ribosomal tunnel (Woolhead *et al.*, 2004). Thus, secondary structure-induced ribosome stalling may play a role in membrane targeting in a manner similar to the transient pausing of translation by suboptimal codons (Fluman *et al.*, 2014, Pechmann *et al.*, 2014).

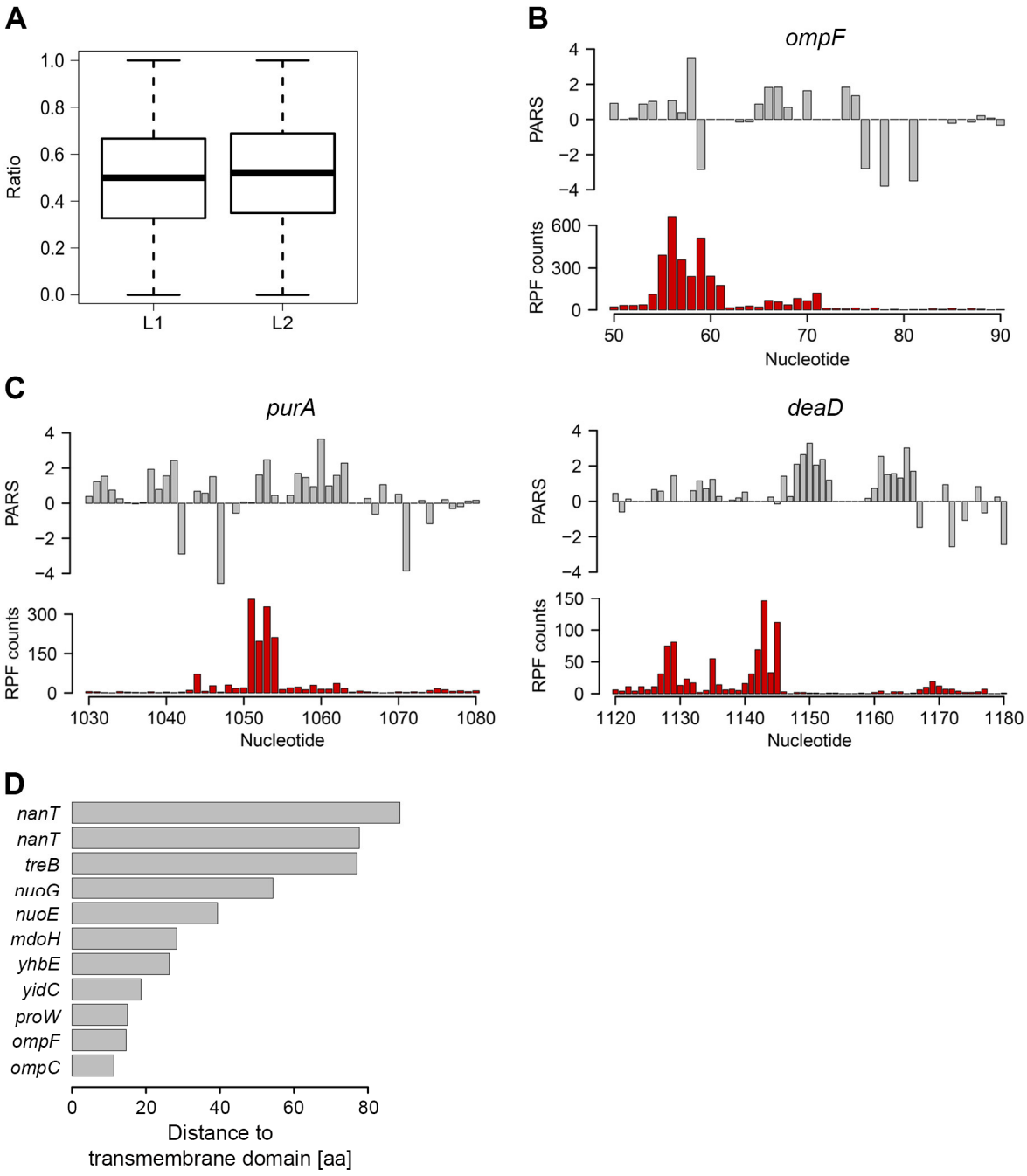


Figure 2.6 | Ribosomal pausing induced by secondary structure in CDS. (**A**) Globally, ribosomal pausing is not significantly affected by the presence of secondary structure in the CDS. Box plot analysis of the ratio of RPF upstream (L1) calculated from Eq 2 and downstream (L2) calculated from Eq 3 of detected secondary structures (P = 0.1209, Kolmogorov-Smirnov test). (**B,C**) Ribosomal pausing is observed within coding sequences above the 80th percentile (panel A). Examples of *ompF* transcript with previously validated secondary structure (Schmidt *et al.*, 1995) (**B**) as well as newly detected genes (**C**) for which a local secondary structure causes non-uniform ribosomal distribution. Aligned PARS score (upper panel, gray) with the RPF counts (bottom panel, red) at each nucleotide. (**D**) Distance of the last residue of a transmembrane helix and the first nucleotide of a detected secondary structure which causes ribosomal stalling. The transmembrane helices of membrane proteins with structure-induced ribosome accumulation were predicted with www.cbs.dtu.dk/services/TMHMM/. Note that for *nanT* two structured regions were detected; the upper one reports on the structured region detected at 1234 nt. aa, amino acid.

2.1.5 mRNA abundance correlates with the mean structural propensity of the coding sequence

Clearly, under physiological conditions, the secondary structure propensity of the majority of CDSs had no impact on the elongating ribosomes. However, mRNA structure is important for a variety of processes, including maintenance of stability and half-life (Carrier & Keasling, 1997). To quantify the transcriptome, we performed an RNA-Seq experiment (Mortazavi *et al.*, 2008) which exhibited high reproducibility between biological replicates (Fig. 7.3). Comparison of the mean PARS score over the CDS revealed a clear correlation with the mRNA abundance (Fig. 2.7 A,B): the 30% most abundant transcripts exhibited higher secondary structure than the 30% least abundant genes ($p = 2.2*10^{-16}$, Mann-Whitney test, Fig. 2.7 C). Thus, we next asked whether low abundance transcripts are more susceptible to degradation. In *E. coli*, RNase E is a key enzyme in RNA metabolism and has a major influence on the mRNA life cycle (Mackie, 2013). Recent RNA-Seq-based analysis identified ~1,800 RNase E target sites within *E. coli* mRNAs (Clarke *et al.*, 2015). Within the genes with a transcript load over the threshold of 1.0 (Fig. 2.4 A), we identified 64 RNase E cleavage positions (Fig. 2.8 A, Table 7.2) which score among the first 100 cleavage sites (Clarke *et al.*, 2015). However, those genes did not cluster within the gene group with the lowest abundance and lowest propensity to form secondary structure.

The cleavage site of RNase E is at an unpaired sequence (Clarke *et al.*, 2015) which lacks a specific sequence motif but is rather enriched in A and U (Fig. 2.8 A, inset). Single gene studies propose the importance of stem-loop structures 5' adjacent to the A/U rich target sites of RNase E (Ehretsmann *et al.*, 1992, McDowall *et al.*, 1994, Moll *et al.*, 2003). Strikingly, we observed this common signature for the 64 identified RNaseE target sites: the unpaired target region is preceded by a structured mRNA stretch (Fig. 2.8 A). Also, this structural signature is common for all additional ~1,800 RNase E target sites. Furthermore, we analyzed the structural features of additional endonucleases which have been identified under RNase E-depleted conditions (Clarke *et al.*, 2015). The target sites of other endonucleases bears no secondary structure upstream the cleavage site and thus significantly differ than that of RNase E (Fig. 2.8 B) implying that the structural signature of the RNase E target sites is of importance for its recognition. Notably, the target sites of all endonucleases lack a specific consensus sequence motif but are rather enriched in specific nucleotides (Fig. 2.8 C). This observation is consistent with mutational study of the unpaired RNase E cleavage site, which

suggests that RNase E cleavage is affected by the extent of A and U rather than their order (McDowall *et al.*, 1994).

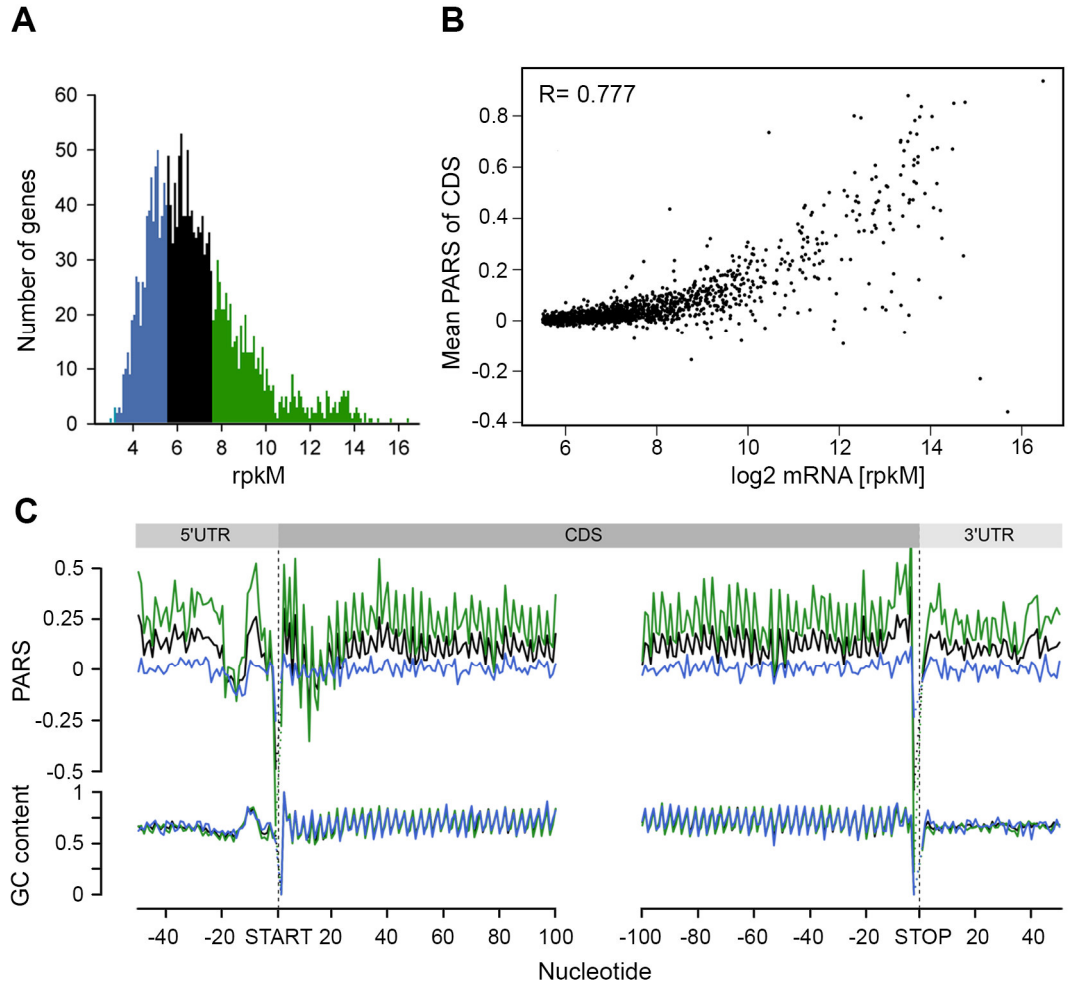


Figure 2.7 | mRNA structure correlates with mRNA abundance. (**A**) Distribution of transcript abundance, expressed in gene read counts normalized by the length of CDS per kilobase and the total mapped reads per million (rpkM). The 30% least (blue) and most (green) abundant genes from the reliably detected genes (Fig. 7.3) are highlighted. (**B**) Dependence of the mean PARS score on the mRNA abundance of the middle (black) and most (green) abundant transcripts as defined in panel A. R = 0.777, Pearson correlation coefficient. (**C**) Average PARS score (top) and GC content (bottom) for each position of all transcripts (black curve) as well as the 30% most (green) and least (blue) abundant..

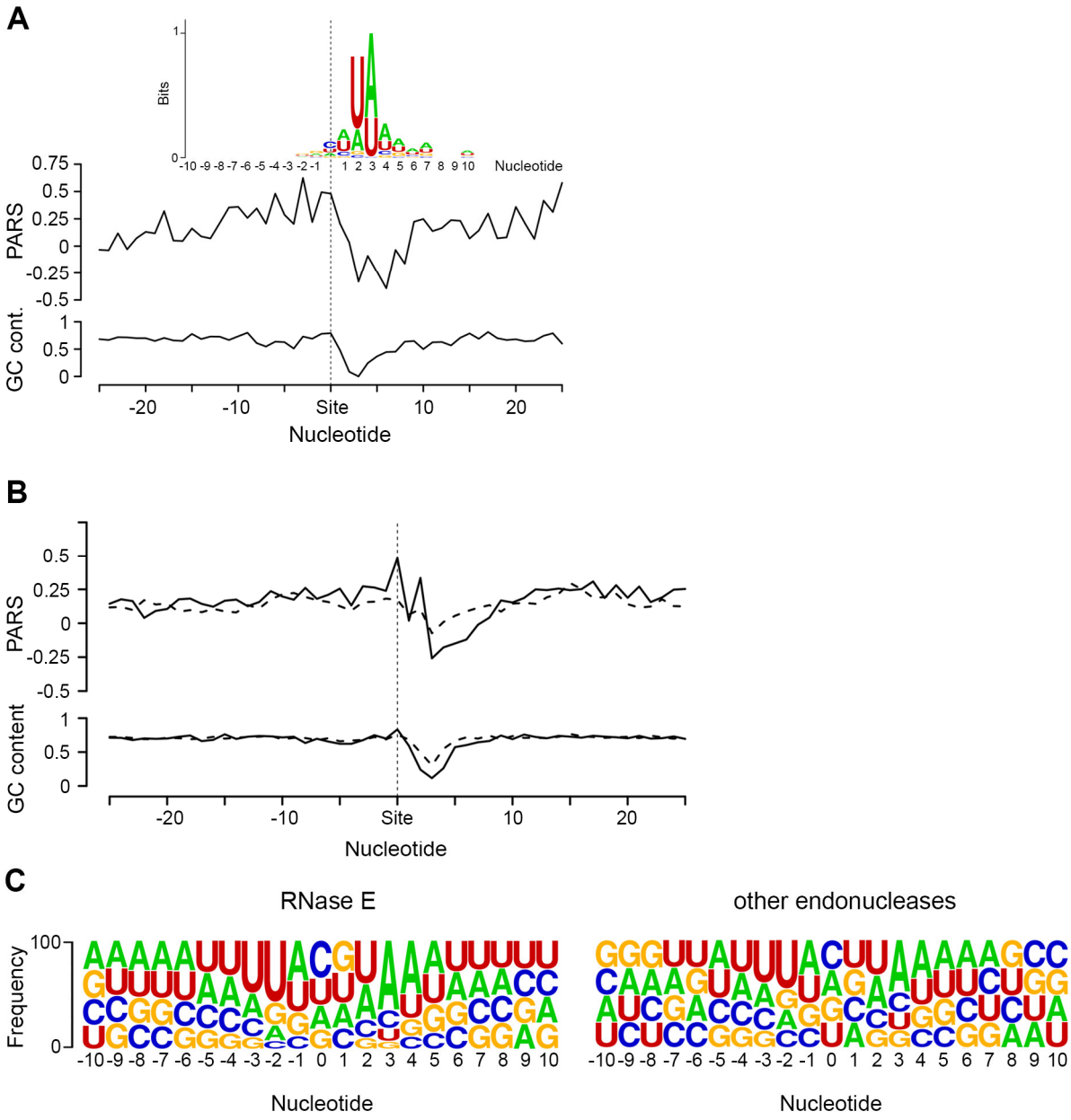


Figure 2.8 | Structural signature of RNase E target sites. (**A**) Average PARS score (top) and GC content (bottom) for each position around the top 64 RNase E cleavage sites (Table 7.2). Inset, the sequence logo of the aligned RNase E cleavage sites, spanning from -10 to +10 nt. (**B**) The structural signature of the RNase E target sites differ significantly from that of other endonucleases (-8 to +2 nt, p = 0.0066, Mann-Whitney test). Average PARS score (top) and GC content (bottom) for each position around all identified ~1,800 RNase E cleavage sites (solid line) and additional ~5000 endonucleolytic sites (dashed line) detected under RNase E-depleted conditions (Clarke *et al.*, 2015). (**C**) Frequency of the nucleotides around the RNase E cleavage site or other endonucleases whose PARS plot is shown in B.

2.1.6 Unstructured sequence upstream of the start codon is a general feature of *E. coli* genes

We detected a unique structural feature for the *E. coli* transcripts which is not present in yeast and human (Kertesz *et al.*, 2010, Wan *et al.*, 2014): the region 7-12 nt upstream of the start codon is significantly more structured (mean value 0.17) than the average CDS (mean value 0.11, Fig. 2.5 C, marked with an arrow). A large fraction of genes in *E. coli* is initiated by Shine-Dalgarno sequence upstream of the start codon and its hybridization strength to the anti-SD of the 16S rRNA (3'-UCCUCCAC-5') determines initiation fidelity. We computed the minimum hybridization free energy (MHE) between the anti-SD sequence and genes whose translation was initiated by SD which revealed four major groups (referred to as strong, medium, weak, and no SD groups, Fig. 2.9 A). [The complete list of all parameters plotted in Fig. 2.9 A is available on our webpage (http://www.chemie.uni-hamburg.de/bc/ignatova/tools-and-algorithms.html)]. A randomized sample of the same size displayed different MHE distribution (Fig. 7.5), implying the functional importance of different SD groups. Moreover, the four groups that are selected based on the strength of the SD:anti-SD pairing resemble previous definitions (which however use a threshold of *MHE* value of -4.4 kcal/mol to select for more stringent SD sequences) (Ma *et al.*, 2002). Note that we did not use any threshold and also included SDs with lower MHE (weak SD) that occur naturally, e.g., AAGG (Wood *et al.*, 1984) with MHE of −2.9 kcal/mol.

In general, the GC content of each SD group mirrored the SD strength. SD:anti-SD base pairing is crucial to align the P-site of the ribosome on the start codon, hence the optimal spacing between the SD and the start codon is 7-8 nt (Ringquist *et al.*, 1992, Osterman *et al.*, 2013) which we also detected independent of the strength of the SD (Fig. 2.9 A). To our surprise, we did not observe any correlation between SD strength and translation efficiency, which was determined by the density of ribosomes (RPF) per mRNA (Pearson correlation, $R = 0.03$, Fig. 2.9 A). Highly translated genes did not preferably cluster in any of the SD groups (Chi-square test: $p = 0.3539$, black symbols, Fig. 2.9 A). Notably, even some genes lacking an SD sequence were also highly translated (Fig. 2.9 A). We also noticed that for genes with strong and medium SD more RPFs accumulated in the SD vicinity (Fig. 2.9 B); these genes were slightly more structured in the SD vicinity than genes with weak SD or those lacking an SD, which is however mirrored in the GC content in this region (Fig. 2.9 C).

By analyzing the profiles of the gene groups with different SD strength, we noticed one striking feature: the region starting at ~20 nt upstream of the start codon is the most unstructured region within each gene (mean value of -0.06 for the region -22 to -13 nt, Fig. 2.9 C). Strikingly, this feature is not recapitulated by the GC content suggesting that it is not selected through A/U-rich sequences and may play active role in regulating translation initiation. Clearly, ribosomes attach to this unstructured site since we detected reads in the ribosome profiling data set at this location (Fig. 2.9 B). The ribosome binds in a biphasic-kinetics mode to some mRNAs and both phases have clear implications for the expression of the corresponding gene (Studer & Joseph, 2006). While the second transition in the kinetic curves represents the positioning of the anti-SD of 16S rRNA over the SD sequence, the role of first phase is unclear (Studer & Joseph, 2006). Usually multiphasic transitions suggest multiple binding events, thus we hypothesized that this unpaired region might represent an additional unspecific binding site of the 30S to facilitate its positioning over the SD. To examine the physiological importance of this unpaired site in expression of the encoded protein, we compared four different sites: AU-rich sequences with low (i.e. unstructured) and high (i.e. structured) PARS score and GC-rich sequences with low and high PARS score. Each site was fused to the first 50 nt of *adhE* (SD and first 42 nt of the CDS) upstream of the YFP. The resulting expression was quantified by flow cytometry (schematic in Fig. 2.9 D). Notably, constructs with less structured upstream regions resulted in higher expression than their more structured counterparts with similar sequence content (compare AU-rich with single- and double-stranded docking site — *adhE* vs *cspE*, or GC-rich with single- and double-stranded docking site — *ppiD* vs *accD*; Fig. 2.9 D). The variant with unpaired AU-rich region exhibited higher expression than the one with unpaired GC-rich sequence (compare *adhE* and *ppiD*, Fig. 2.9 D). In general, AU-rich single-stranded regions are less structured than GC-rich single-stranded regions, which correlates with the mean PARS score over this region (-30 – -12 nt upstream of the start codon); the mean PARS score of unpaired AU-rich *adhE* is -0.564 and of the GC-rich *ppiD* is -0.495 (Fig. 2.9 D). The *adhE* gene exhibited the highest expression, which might be argued that it due to using part of *adhE* as an invariable backbone in our constructs (schematic Fig. 2.9 D). To exclude this possibility, we replaced the invariable *adhE* part with a fragment of the same size originating from *ppiD* or from *accD* (SD and first 42 nt of the CDS), presenting the two least expressed docking sites (Fig. 7.5 B).

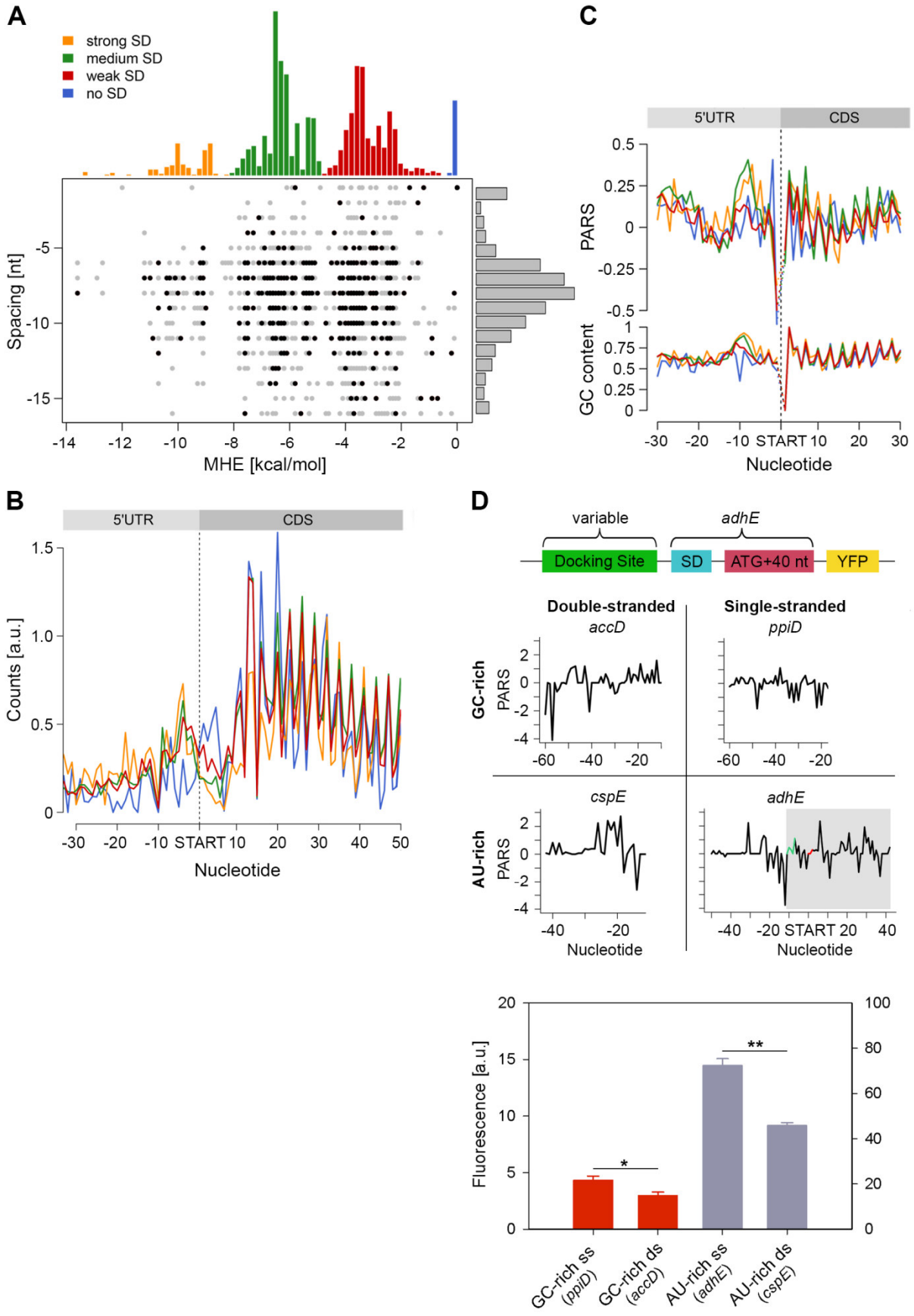

A
strong SD
medium SD
weak SD
no SD
Spacing [nt]
MHE [kcal/mol]
B
5'UTR
CDS
Counts [a.u.]
Nucleotide
C
5'UTR
CDS
PARS
GC content
Nucleotide
D
variable
adhE
Docking Site
SD
ATG+40 nt
YFP
Double-stranded
Single-stranded
accD
ppiD
GC-rich
cspE
adhE
AU-rich
PARS
Nucleotide
Fluorescence [a.u.]
GC-rich ss (ppiD)
GC-rich ds (accD)
AU-rich ss (adhE)
AU-rich ds (cspE)

Figure 2.9 | Stronger SD sequence has a higher propensity to form secondary structure which does not correlate with the translation efficiency. (**A**) SD strength does not correlate with translation efficiency (i.e. the total RPFs per coding mRNA) of a gene. SD hybridization energies fall into four major distributions: strong SD,MHE< -8.5 kcal/mol; medium SD, -8.5 <MHE< -4.4 kcal/mol; weak SD, -4.4 <MHE< -2 kcal/mol; no SD,MHE> -2.0 kcal/mol. For each gene (dot) theMHE(horizontal axis) of the SD sequence is plotted against SD spacing (vertical axis), defined as the distance between the second to last nucleotide of the SD and the start codon. Genes with the 30% highest ribosomal density are highlighted as black dots. (**B**) Cumulative plots of ribosomal density for all genes grouped by SD strength. Genes were aligned by the first nucleotide of the start codon. (**C**) Average PARS score smoothed over 3 nt (top) andGCcontent (bottom) for each position of the four SD strength classes, aligned by the start codon. The four different SD groups are color coded as in panel A. (**D**) FACS expression analysis of adhE whose original docking site was replaced by three other docking sites with clearly different sequence (AU-rich or GC-rich) and different PARS score. Only the sequence upstream of the SD (green on the PARS profiles) was replaced. The common part of adhE which is fused to YFP (schematic inset) is shadowed on the PARS profiles. The average PARS score over the docking site (12 to 30 nt upstream of the start codon, red on the PARS profiles are): adhE–-0.564, ppiD–-0.495, cspE–0.724, accD–0.665. Data are means (n = 3) ± standard error of the mean (s.e.m.). *, P <0.05; **, P <0.01.

Replacing the original region upstream of the SD with the most unstructured sequence of *adhE* enhanced the expression by twofold for *ppiD* and by sixfold for *accD* (Fig. 7.5 B).

In sum, our results feature the poorly structured region at ~20 nt upstream of the start codon as an additional binding site of the ribosome distinct from SD binding, and its secondary structure propensity correlates with the expression of the downstream CDS.

2.1.7 Higher secondary structure upstream of the stop codon has a likely role in termination

In the metagenome analysis we noticed that the region upstream of the stop codon is more structured than the average PARS score of the CDS and 3'-UTR, whereas a GC content of this region does not significantly differ from the average CG content of the CDS (Fig. 2.5). Genes terminated with the UAA codon exhibited the highest propensity to form secondary structures in the 3'-termini of the CDS ($p = 2.2*10^{-16}$, Mann-Whitney test, Fig. 2.10 B). Notably, we observed an enrichment of RPF reads ~10-30 nt upstream of the UAA-termination codon ($p = 6.94*10^{-6}$, Mann-Whitney test) suggesting a persistent secondary structure (Fig. 2.10 C).

In *E. coli*, a large fraction (53%) of protein-coding genes is organized as polycistronic mRNAs in operons to facilitate the association and physical interactions of functionally

related proteins. The SD sequence of an overlapping or a closely positioned downstream gene (Fig. 7.6 A) may influence our analysis, resulting in an apparent higher structure in the 3' vicinity of the upstream gene. Thus, we next separately analyzed the secondary structure upstream of the stop codon of protein-coding genes organized in operons from those in non-operons; the operon group is additionally divided in two groups: non-overlapping, with a distance of $\geq$ 30 nt from the downstream gene, and overlapping, with a downstream gene located < 30 nt to the upstream gene. Only UAA-terminated genes showed increased PARS score (p = 0.00023 for non-overlapping, $p = 3.2*10^{-10}$ for overlapping, $p = 4.07*10^{-5}$ for non-operon, Mann-Whitney test, Fig. 2.10 A) in the 3' vicinity of the coding sequence and this feature is not mirrored by the GC content. Also, the frequency of the three stop codons (UAA, UAG and UGA) is similar for all gene groups and resembles stop codon usage in the genome (Fig. 7.6 B).

We hypothesized that secondary structure upstream of the stop codon may influence the termination fidelity of the UAA-terminated genes. Additional in-frame stop codons may act as safeguards against leaky termination. We reasoned that if the structure in the vicinity of the UAA stop codon influences termination, those genes would show lower frequency of ribosomes in the 3'-UTR. We analyzed the ribosome occupancy downstream of the UAA- and UGA-terminated genes (considering it in general as a readthrough). Overlapping genes were excluded from this analysis as ribosomes terminating the upstream gene cannot be unambiguously distinguished from ribosomes initiating the downstream gene. Strikingly, we observed a low but significant fraction of RPF reads downstream of the UGA stop codon while RPF reads in the 3' UTR of the UAA-terminated genes were nearly not detectable (Fig. 2.10 C). This phenomenon occurred in the background of a similar distribution of additional in-frame stop codons downstream of all terminating codons: UAA – 10.7%, UGA – 8.7% and UAG – 7.4%. Together, this analysis suggests that structure upstream of the stop codon may enhance the termination fidelity of the UAA-codon terminated genes.

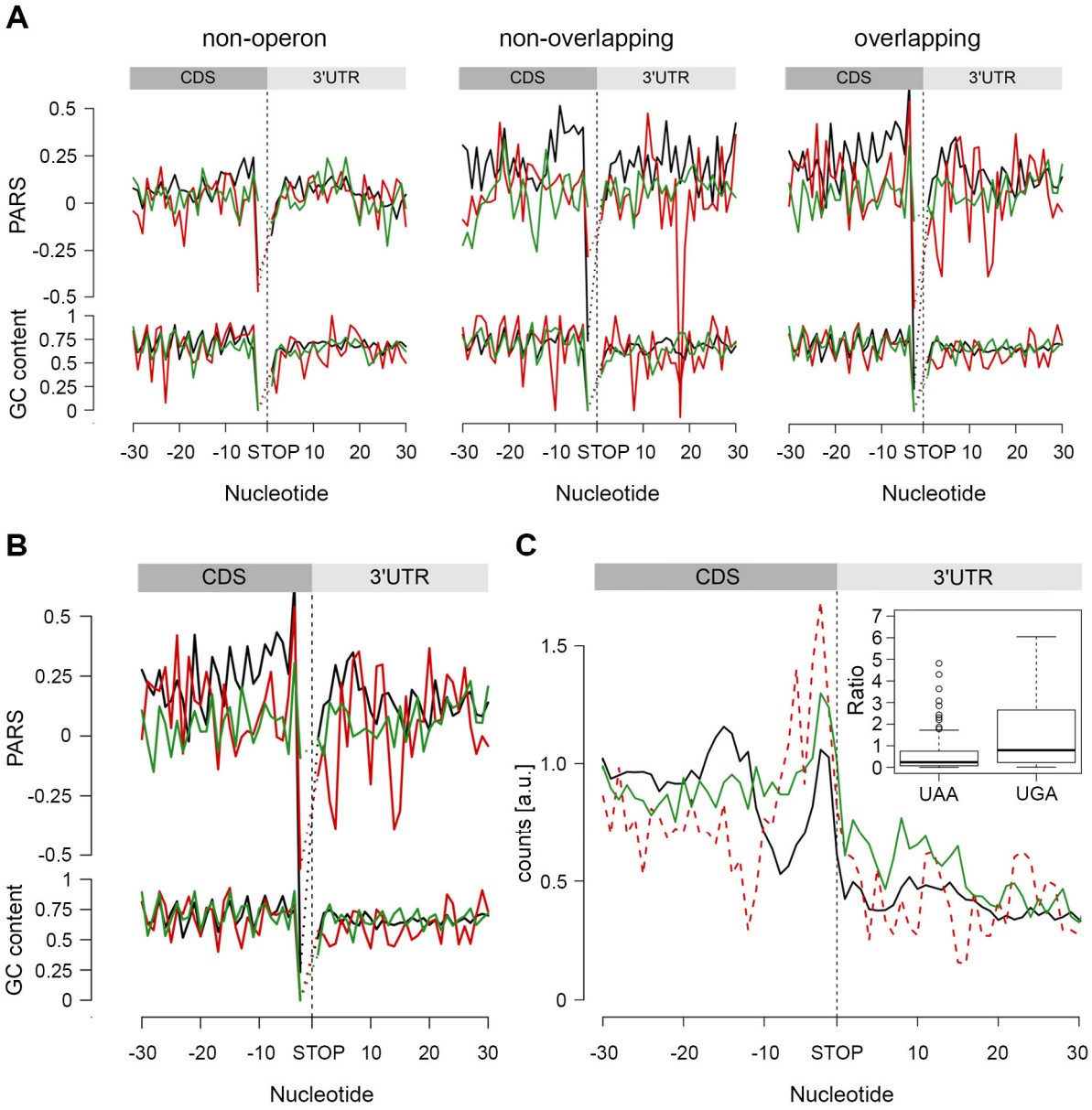


Figure 2.10 | The stop codon of operon genes is more structured than non-operon genes. (**A**) Average PARS score and GC content around the stop codon of different gene groups terminated with UAA (black), UAG (red) and UGA (green) stop codons. (**B**) Average PARS score and GC content for each position of genes terminating with UAA (black), UAG (red) and UGA (green) stop codons. (**C**) RPF coverage around the stop codon region for genes terminated by UAA (black), UAG (dashed red) and UGA (green) stop codons. Only genes with coverage over 60 reads (Fig. 7.3 D) were used; overlapping operon genes were excluded. Note, that UAG-terminated genes are included only for comparison; their low number prevents performing any statistical analysis. The inset shows, for both UAA- and UAG-terminated genes, the ratio between the RPFs downstream of the stop codon (3 to27 nt) and a mean of the CDS. The readthrough value for the majority of the genes was zero; only genes with a value higher than zero are plotted.

2.2 Probing dimensionality beyond the linear sequence of mRNA

mRNA is more than a mere information-carrying molecule between DNA and translating ribosomes. The intrinsic propensity of the linear single-stranded mRNA to form higher order structures, i.e. secondary and tertiary folded motifs, adds another layer of information to guide its cellular localization, regulate gene expression, or fine-tune the stress response. Pioneering attempts to probe local Watson-Crick geometries and more distant non-Watson-Crick base-pairings started in the fifties (Cox & Littauer, 1959). Since then, the chemical toolbox of reagents that modify unpaired nucleotides has grown and currently comprises dimethylsulfate (DMS), diethylpyrocarbonate (DEPC), 1-cyclohexyl-3-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate (CMCT), ethylnitrosourea (ENU) and the recently added N-methylisatoic anhydride (NMIA) and 2-methylnicotinic acid imidazolide (NAI). The chemical modification of a nucleotide blocks reverse-transcription and is precisely detected on a sequencing gel together with a Sanger sequencing reaction. The reactivity of these chemical substances is restricted to single-stranded nucleotides (Ziehler & Engelke, 2001). The structure-probing spectrum has been expanded by using enzymes that specifically recognize single-stranded (ribonuclease A, T, S1) or double-stranded (ribonuclease V1) regions (Ziehler & Engelke, 2001), thus yielding more complete information by detecting both paired and unpaired nucleotides. Furthermore, X-ray crystallography and nuclear magnetic resonance spectroscopy provide the most precise structural information, particularly in resolving distant tertiary contacts and hierarchical mRNA assemblies. Although laborious for a variety of reasons, such as the requirement for a large quantity of the starting material, solubility issues, and the large size of the mRNA molecule, collectively these approaches have yielded some structures benchmarking our understanding of the impact of secondary mRNA structure in regulating cellular processes. These include *Tetrahymena* ribozyme, RNase P catalytic domain, and *rpoH* RNA thermometers.

Coupling the susceptibility of specific nucleotides to chemical modification as well as enzyme cleavage with massively parallel sequencing approaches adds a new dimension towards global transcriptome-wide analysis of the mRNA structures. Importantly, this new twist overcomes the limitations of the single-molecule approaches largely restricted to probe small-size mRNA fragments. The first transcriptome-wide approach linked the enzyme probing of paired and

unpaired nucleotides of isolated yeast total mRNAs with the nucleotide resolution of deep sequencing (i.e., parallel analysis of RNA structure, PARS (Kertesz *et al.*, 2010)). Further twists of PARS, e.g. Frag-Seq (Underwood *et al.*, 2010) and ds/ssRNA-seq (Zheng *et al.*, 2010, Li *et al.*, 2012a, Li *et al.*, 2012b) also exploited the catalytic specificity of enzymes to assess the transcriptome of mouse, plants and metazoans. Although an *in vitro* methodology without the full *in vivo* interaction profile, these studies revealed new insights into RNA structure valid for a very large fraction of mRNA, such as different structure propensity of the coding regions and non-coding regions in yeast and human (Kertesz *et al.*, 2010, Wan *et al.*, 2014). This approach analyzes isolated protein-free total mRNA and, its power is in unraveling the intrinsic propensity of each nucleotide to participate in double-stranded (ds) or single-stranded (ss) structure.

Cellular components (e.g., RNA-interacting proteins, translating ribosomes) (Kwok *et al.*, 2013) or simply the crowded cellular environment (Tyrrell *et al.*, 2013) may introduce some differences to the *in vivo* mRNA structure compared to that *in vitro*. To probe mRNA structure in its native environment and gain more biologically meaningful interpretations, chemical probing of nucleotide accessibility has been applied directly to cells and coupled to the detection power of deep sequencing (Fig. 2.11). Though a step forward in determining mRNA structure directly in the cellular environment, some of the chemical reagents, e.g., DMS (Ding *et al.*, 2014, Rouskin *et al.*, 2014, Talkish *et al.*, 2014), CMCT (Incarnato *et al.*, 2014), are limited in resolution as they react with only few of the four nucleotides. Only the reagents for selective 2'-hydroxyl acylation, NMIA and 1M7, and analyzed by primer extension (SHAPE), modify all four single-stranded nucleotides (Wilkinson *et al.*, 2006, Lucks *et al.*, 2011), but these are poorly soluble in water-based solutions and highly unstable in living cells (Spitale *et al.*, 2013). Recent synthetic developments allowed extracting information on all unpaired bases using 2-methylnicotinic acid imidazolide (NAI) (Spitale *et al.*, 2015). The toxicity and side effects of the chemical probes remain to be disclosed. For example, DMS targets not only RNAs, but also DNA and proteins (Moio *et al.*, 2011), raising the question as to whether the treatment itself puts the cells under stress and some specific aspects in the stress biology might be captured and co-interpreted.

Indeed, there is no Holy-Grail approach for *in vivo* application. Combined approaches that assess the role of secondary structure from multiple angles should be preferred over a single approach (Fig. 2.11). In the first step, the intrinsic propensity of each mRNA to participate in

secondary interactions can be determined by means of PARS (Kertesz *et al.*, 2010). Combining PARS with ribosome profiling (Ribo-Seq) which reveals the position of translating ribosomes with codon resolution specifically address the impact of mRNA secondary structure for translating ribosomes (Del Campo *et al.*, 2015). In the chemical modifications, nucleotide inaccessibility might result from protein (RNA-binding proteins) or ribosome shielding (Tijerina *et al.*, 2007, Spitale *et al.*, 2014), and thus misinterpreted as non-accessible or nucleotides participating in secondary interactions. Extracting conserved secondary structure patterns in the absence of proteins and ribosomes and then exclusively looking at their translatability by ribosome profiling enabled the discovery of additional regulatory element of translation initiation in bacteria (Del Campo *et al.*, 2015), which would have remained invisible for chemical imprinting as translation *in vivo* is not synchronized. Additional complementation of the PARS approach with RNA sequencing allows deconvoluting the role of the secondary structure propensity and transcript abundance and extract cleavage signatures of different nucleases (Del Campo *et al.*, 2015).

Application of sequencing technologies for mRNA structure determination bears great potential to become a high-throughput, global and unbiased approach. However, although these techniques identify unpaired and paired nucleotides (or regions), they do not specify interaction partners. RNA duplexes can be confidently determined using hiCLiP (combining cross-linking) approach (Sugimoto *et al.*, 2015); this technique extracts only double-stranded mRNA directly involved in interactions with RNA-binding proteins. Furthermore, sophisticated computational analyses are needed to construct mRNA secondary structure; in these approaches the experimental data are used to constrain the predicted secondary structures. The application of parallel sequencing for secondary structure analysis drives forward the development of various other algorithms for analyzing and predicting RNA structures whose application is largely restricted when applied to large mRNAs (Eddy, 2014). Collectively, these computational tools reveal more local, secondary RNA structure and are limited in defining the global tertiary architecture of RNA. Conceptually, phylogenetic analysis can be used to predict new structures by extracting patterns of conserved and covariant nucleotides in comparison to known 3D structures. But for how many RNA molecules are the complete structures known? A first step towards identifying distant tertiary interactions with non-Watson-Crick geometry is the RNA proximity ligation (RPL) approach (Ramani *et al.*, 2015). Proximity ligation identifies the three-dimensional proximities of the

each nucleotide, and coupling it to deep sequencing enables high-throughput and treatment of large data sets (Ramani *et al.*, 2015). Although the data are quite noisy and mRNA is analyzed *ex vivo* upon extraction, it has a potential, combined with *in vivo* approaches, to unravel more complex mRNA architectures. This information is of particular importance to understand the conformations of translationally inactive mRNA, e.g., during transport and localization, within stress or neuronal granules and in p-bodies.

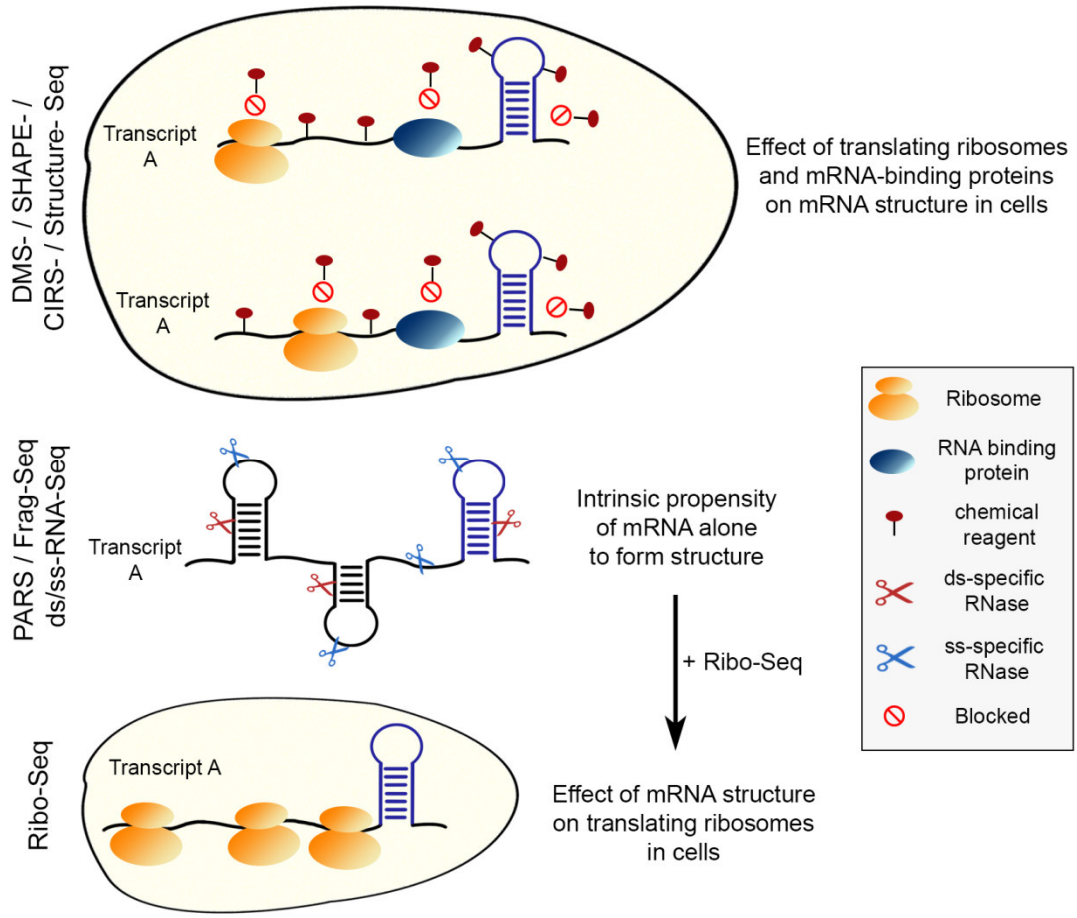


Figure 2.11 | Overview of the approaches used to probe secondary mRNA structure. The chemical probing directly in the cell reveals persistent structures *in vivo*; the binding sites of RNA-binding proteins also protect from modification with the chemical reagent and might be considered as double-stranded structure. Combined enzymatic probing *in vitro* with ribosome profiling (Ribo-Seq) allows for distinguishing of persistent mRNA structures with effect only on translation.

2.3 Mapping the non-standardized biases of ribosome profiling

Ribosome profiling is a new emerging technology that uses massively parallel amplification of ribosome-protected fragments and next-generation sequencing to monitor translation *in vivo* with codon resolution. Studies using this approach provide insightful views on the regulation of translation on a global cell-wide level. In this review, we compare different experimental set-ups and current protocols for sequencing data analysis. Specifically, we review the pitfalls at some experimental steps and highlight the importance of standardized protocol for sample preparation and data processing pipeline, at least for mapping and normalization.

2.3.1 Introduction

At any given time, the amounts and types of proteins reflect the functional status of the cell. The protein composition is a balance between protein synthesis and degradation. On the synthesis side, protein production is controlled at the level of transcription and translation and the messenger RNA (mRNA) is the connecting entity between these two processes. Moreover, emerging evidence suggests that the mRNA open reading frame bears far more information than just the amino acid sequence of the synthesized protein. Codon choice to encode one amino acid (Plotkin & Kudla, 2011), tRNA modifications (Nedialkova & Leidel, 2015, Tyagi & Pedrioli, 2015) or secondary structures (Wen *et al.*, 2008, Chen *et al.*, 2013) modulate the local speed at which mRNA is translated and link it to protein biogenesis or stress response. Recent developments in the next-generation sequencing (NGS) technologies revealed additional layers embedded in the mRNA to regulate its translatability and consequently the downstream processes in protein biogenesis including cotranslational folding, insertion into membranes and interactions with auxiliary factors (Kramer *et al.*, 2009, Zhang & Ignatova, 2011, Pechmann *et al.*, 2014). Specifically, a recent twist of the NGS technologies to capture translating ribosomes, named ribosome profiling (Ingolia *et al.*, 2009), has significantly advanced our understanding on translation regulation in various organisms (reviewed in (Ingolia, 2014)). Ribosome profiling is based on high-throughput sequencing of ribosome-protected RNA fragments, or ribosomal 'footprints', which specifically report on the position of the translating ribosomes with a nucleotide resolution (Ingolia *et al.*, 2009). A

growing body of published literature illustrates the power of this approach to unravel new aspects on translation regulation, for example identification of extensive upstream initiation at non-AUG codons in eukaryotes (Ingolia *et al.*, 2009, Ingolia *et al.*, 2011, Fritsch *et al.*, 2012, Lee *et al.*, 2012) and specific regulation of the stress response at translation level (Liu *et al.*, 2013, Shalgi *et al.*, 2013, Andreev *et al.*, 2015). Further development of the profiling technology to isolate a fraction of ribosomes that are involved in specific cellular processes revealed new insights into the localized protein synthesis in yeast (Jan *et al.*, 2014) or the interaction with a trigger factor, an auxiliary factor facilitating co-translational folding in bacteria (Oh *et al.*, 2011).

Without doubt ribosome profiling is a powerful technology to address various aspects of translation regulation on a genome-wide scale, and several excellent reviews summarize the power of this technology (Morris, 2009, Kuersten *et al.*, 2013, Michel & Baranov, 2013, Ingolia, 2014). However, this approach is relatively young with steadily evolving experimental protocol and a non-standardized platform for data analysis. The pace of exploration creates some difficulties in comparing results produced in different laboratories. In addition, different approaches to analyze the data disclose variations in their interpretation (Gerashchenko & Gladyshev, 2014). Here, we focus on the ribosome profiling procedure and data analyses and critically review the biases of the various steps in the profiling protocol as a potential source of variation. We also provide examples on how variations in the ribosome profiling procedure put restrictions on the downstream analysis and determine the information that can be extracted from the data. We suggest standardizing ribosome profiling protocol and adjusting only a step (or few steps) depending on the specific scientific question.

2.3.2 Isolation of intact translating ribosomes

At the core of ribosome profiling is a nuclease digestion of mRNA unprotected by the ribosome and recovering ribosome-protected mRNA fragments (i.e., ribosome footprints) (Steitz, 1969) and their conversion into a DNA library which is further analyzed by deep sequencing (Ingolia *et al.*, 2009) (Fig. 2.12). Thus, this approach maps the position of the translating ribosomes on each mRNA and provides a snap-shot of translation.

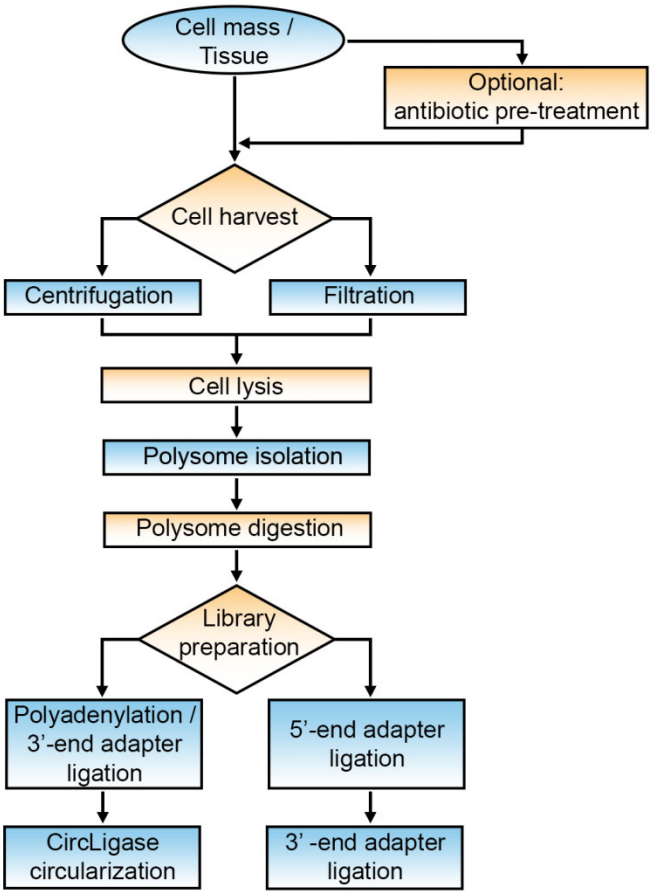


Figure 2.12 | Flow-chart of isolation of intact ribosome-mRNA complexes and library preparation for the ribosome profiling experiment. Crucial steps at which specific decisions need to be taken are color coded in orange. Detailed knowledge of the bias of each of those procedures is essential for the careful interpretation of the sequencing data.

2.3.3 Harvesting the cells and antibiotic pre-treatment

The most delicate step in the sample preparation is the isolation of intact ribosome-mRNA complexes. Ideally, the isolation procedure should faithfully freeze the translating ribosomes and avoid conditions that stimulate ribosomal drop-off and, most importantly, ribosome relocation on the mRNA during the sample processing. Early in the development of the ribosome profiling approach, cells were pre-incubated with elongation inhibitors (mainly chloramphenicol for bacteria and cychloheximide for eukaryotes) to inhibit further movement of the elongating ribosomes along the mRNA (Ingolia *et al.*, 2009). The antibiotic treatment markedly affects the coverage profiles and introduces some bias in the results; the elongation

inhibitors do not uniformly stall elongating ribosomes but rather show a codon-dependent mode of action (Orelle *et al.*, 2013). Cycloheximide also allows one complete translocation cycle before blocking the ribosome (Pestova & Hellen, 2003, Schneider-Poetsch *et al.*, 2010) and thus diffuses the read-out when determining codon-dependent stalling (Nedialkova & Leidel, 2015), while non-antibiotic treated cells deliver much sharper pause sites corresponding to rare codons (Pelechano *et al.*, 2015). In addition, a broad cumulative peak downstream of the start codon has been seen in the earlier profiling papers that use elongation inhibitors and interpreted as slow initiation (Ingolia *et al.*, 2009). The initial peak, albeit still present, significantly decreases when cell mass is flash frozen and elongation inhibiters are omitted (Guydosh & Green, 2014, Lareau *et al.*, 2014). The disproportionately high accumulation of reads at initiation is rather an artifact of the antibiotic pre-treatment (Becker *et al.*, 2013) and results from inhibition of translation elongation with ongoing initiation (Ingolia *et al.*, 2011). The antibiotic does not immediately reach the threshold of complete inhibition of elongation; instead its concentration increases gradually in the cell (Gerashchenko & Gladyshev, 2014). Hence upon treatment, some initiating ribosomes continue into the elongation cycle until they encounter the drug, which results in an excess of ribosomal footprints over the first five to ten codons from the coding sequence (Gerashchenko & Gladyshev, 2014). Additionally, an 80S ribosome stalled in the proximity of the start codon will prevent any subsequent scanning ribosome from reaching the initiation codon, which may result in an apparent stalling at an upstream open-reading frame (uORF). Thus, an initiation site with mediocre context in uORF will be occupied because of the highly efficient but blocked downstream start site (Jackson & Standart, 2015) which may lead to an erroneous interpretation of alternative uORF-induced initiation. Careful consideration of the effect of antibiotics on ribosome coverage offers little support that the large number of genes with uORFs is involved in shaping the resistance to oxidative stress (Gerashchenko & Gladyshev, 2014). Ribosome profiling without antibiotics prior to cell harvesting revealed that translation of only a small fraction of uORF-bearing mRNA was refractory to oxidative stress (Andreev *et al.*, 2015). Elongation inhibitors added prior to harvesting the ribosome-mRNA complexes alter the distribution of reads in the cumulative ribosome profiles (namely, the aligned and averaged profiles of many genes). For example emetine-stalled elongating ribosomes give slightly longer fragments than those isolated from cycloheximide-treated mammalian cells suggesting that various antibiotics stabilize different ribosome conformation (Ingolia *et al.*,

2011). Conversely, drug pre-treatment may eliminate some features of biological importance in the cumulative ribosome profiles. For example, antibiotic pretreatment in mammalian cells eliminates the ribosomal peak at the end of the open-reading frames, which is observed in untreated cells (Ingolia *et al.*, 2011).

The most widely applied cell harvesting procedure involves rapid cooling of the cell suspension and centrifugation (Becker *et al.*, 2013) (Fig. 2.12). Bacteria are cooled by pouring the cell suspension over crushed ice, while eukaryotic (mammalian) cells cultured in monolayer are re-suspended in ice-cold PBS supplemented with elongation inhibitor and immediately pelleted by centrifugation (Guo *et al.*, 2010). Tissues are usually flash-frozen and grinded in the lysis buffer supplemented with elongation inhibitor (Gonzalez *et al.*, 2014). An alternative approach for harvesting of cells growing in suspension is a rapid filtration of the cells in a pre-warmed glass nitrocellulose filtration system and flash-freezing the membrane with the cells (Fig. 2.12). So far, this filtration approach has been mainly used in unicellular organisms (yeast and *E. coli*, for example) (Ingolia *et al.*, 2009, Oh *et al.*, 2011, Li *et al.*, 2012c). Both harvesting protocols show good reproducibility between biological replicates (r = 0.99, Pearson correlation coefficient) (Becker *et al.*, 2013). Importantly, however, the RPF accumulation at native stalling sites, e.g. SecM and TnaC, is higher using the filtration harvesting (Becker *et al.*, 2013). Most likely, the filtration approach compared to the centrifugation is less susceptible to variations and faithfully halts the translating ribosomes. Still, harvesting by centrifugation might be the only option for cells that cannot be rapidly filtered. However, it is important to perform it as quickly as possible using pre-chilled devices.

In sum, the procedure for isolation of ribosome-mRNA complexes is of crucial importance. While drug pre-treatment may not influence differential expression analysis, since the expression of each gene is compared under two different conditions with an otherwise uniform protocol, the use of elongation inhibitors or the harvesting procedure may alter the interpretation of position-specific information.

2.3.4 Cell lysis

Similar to the cell harvesting procedure, the aim at this step is to recover the ribosome-mRNA complexes with minimal losses from ribosomal dissociation (or drop-off) and mRNA degradation. The composition of the lysis buffer is optimized to stabilize the ribosome-mRNA complexes with high concentration of magnesium (between 5 and 20 mM) and an additional salt, such as KCl or NaCl and NH_4Cl.

The isolation of intact polysomes is a procedure established in early ribosome research and is still applied today almost unchanged (Wettstein *et al.*, 1963, Dresden & Hoagland, 1965). The composition of the lysis buffer underwent several variations. However, some components of the lysis buffer, if overdosed, may distort the ribosome profiles. For example, high NaCl concentration decreases the monosome peak and enhances the fraction of dissociated ribosomal subunits (Becker *et al.*, 2013); high salt concentration increases the fraction of vacant ribosomes that are not engaged in translation (Blobel & Sabatini, 1971) and consequently decreases the number of RPF. Magnesium stabilizes the translating ribosomes (Ron *et al.*, 1968) and at high concentrations freezes the conformational changes in the bacterial ribosome (Blanchard *et al.*, 2004). Moreover, high magnesium concentration induces folding of the mRNA which hinders the subsequent nucleolytic digestion (Andreev *et al.*, 2015). Lowering the magnesium concentration from 15 mM to 5 mM greatly improves the codon positioning of the footprints and the resolution of the ribosome profiling (Ingolia *et al.*, 2012). Also, low magnesium conditions permit conformational flexibility of the ribosome and create heterogeneity in the length ribosomal footprints (Lareau *et al.*, 2014); the variant ribosomal footprints are informative on distinct stages of the translating ribosome during the elongation cycle.

The lysis buffer also contains an elongation inhibitor to additionally stabilize the ribosome-mRNA complexes during sample processing. The binding kinetics of the antibiotic when present in the cell lysis is rapid compared to the diffusion-driven process of antibiotic enrichment in intact cells during the pre-treatment procedure. Generation of cell extracts from *Saccharomyces* cells in the cycloheximide-containing lysis buffer faithfully halted the ribosomes along the mRNA with no distortion (Guydosh & Green, 2014). Cycloheximide should be preferred over alternative substances that stabilize eukaryotic ribosome-mRNA complexes, e.g. the non-hydrolyzable GTP analog GMP-PNP, as they slightly increase the size of the ribosome footprints (Guydosh & Green, 2014). Although such studies with

bacterial elongation inhibitors are missing, it can be expected that their mode of action will be similar to that of the cycloheximide when added to the lysis buffer.

Along with variations in the composition of the lysis buffer, the lysis procedure also varies. In general, despite the presence of components stabilizing the ribosome-mRNA complexes (e.g. elongation inhibitors, magnesium) to avoid ribosomal reallocation or dissociation, lysis is usually carried out at low temperatures by either adding frozen drops of lysis buffer to a frozen cell powder or flash-freezing with the cell mass. When this is not applicable, i.e. by ribosome profiling of tissues, the lysis buffer is generally added to the sample ice-cold (Gonzalez *et al.*, 2014).

Eukaryotic cells are lysed on ice by repeated micropipetting or homogenization (Guo *et al.*, 2010, Becker *et al.*, 2013, Chew *et al.*, 2013). Pulverized bacteria or monocellular eukaryotes are homogenized in a mill with liquid nitrogen (Oh *et al.*, 2011, Guydosh & Green, 2014, Woolstenhulme *et al.*, 2015). This method is transferrable to any cell type and frozen tissue and should be the preferred lysis approach as it allows treatment of the sample at very low temperatures. During the homogenization, local temperature fluctuations in the sample should be avoided by careful choice of the conditions, i.e. short homogenization pulses and pre-cooling the grinder jar before and after each homogenization cycle (Oh *et al.*, 2011, Guydosh & Green, 2014, Woolstenhulme *et al.*, 2015).

2.3.5 Nucleolytic generation of ribosomal footprints

The clarified lysate is then digested with a nuclease to generate monosomes (Fig. 2.12). RNase I has been exclusively used in eukaryotic ribosome profiling (Ingolia *et al.*, 2012) and micrococcal nuclease (MNase) from *Staphylococcus aureus* in bacteria; RNAse I is inactive in bacteria (Datta & Burma, 1972). MNase can also be used in eukaryotic lysates (Reid & Nicchitta, 2012, Dunn *et al.*, 2013), and, in fact, it leads to a reduced amount of ribosomal RNA (rRNA) contamination compared to RNase I treatment (Oh *et al.*, 2011, Miettinen & Bjorklund, 2015). The activity of the MNase is modulated by calcium ions. A disadvantage of MNase is its preferential cleavage at A or T nucleotides (Dingwall *et al.*, 1981) and consequently, the MNase-generated ribosome footprints might be enriched in A or T nucleotides at their 5' ends. Compared to fragments derived from yeast lysates treated with RNase I, the MNase-generated footprints are more heterogeneous in length (Becker *et al.*,

2013) due to steric effects and less precise 5' cleavage (Woolstenhulme *et al.*, 2015). In contrast, MNase cleaves precisely at the 3' end contour of the ribosome, thus the calibration of the reads in bacterial system should be preferably done using the 3' ends of the reads (Woolstenhulme *et al.*, 2015) (see the section 'Analysis of the sequencing data'). RNase I cleavages are precise at both 5' and 3' ends, enabling calibration using both termini. On the other hand, RNase I-treated samples show a slight bias towards enrichment of short genes (Miettinen & Bjorklund, 2015), though the reason for this remains unclear.

Contamination with rRNA fragments released by the nucleolytic digestion substantially decreases the amount of informative sequencing data. Importantly, the rRNA fragments generated during the nucleolysis of the polysomes are species-specific, but are limited to only few fragments and can be efficiently removed to near completeness by using few complementary oligonucleotides. Thus in setting up a protocol for ribosome profiling in a new cell line or species, it is recommendable by to perform a pioneer sequencing run to identify the contaminant rRNA species and design specific oligonuclotides for the depletion of rRNA-derived fragments.

Finally, the amount of each nuclease needs a careful determination; enhanced nuclease activity (caused either by large amounts of enzyme, pH variations or long digestion times) leads primarily to an increased contamination of the ribosome footprint libraries with rRNA fragments. By contrast, insufficient amount of MNase causes less stringent cleavage of the mRNA and results in longer fragments which migrate outside of the range selected for ribosomal fragments during the gel purification procedure. Consequently, it will yield lower depth and coverage of the mRNAs and it will decrease the accuracy in determining ribosome positions along mRNAs (Becker *et al.*, 2013).

2.3.6 Generation of the deep-sequencing library

The preparation of libraries for deep sequencing involves fusion of adapters to the generated small DNA or RNA fragments. This process also contains biases and a detailed knowledge is of crucial importance to avoid erroneous interpretation of the data. A recent review summarizes the critical caveats in each step of library preparation (van Dijk *et al.*, 2014). Here, we only compare various methods for adaptor ligations to the ribosomal footprints, which are unique to the ribosome profiling procedure. In principle, after nucleolytic digestion

the ribosome profiling follows the typical steps of library preparation in the micro RNA-Seq methodology (Guo *et al.*, 2010), including sequential adaptor ligation, reverse transcription of the RNA fragments and PCR amplification of the transcribed DNA. The earliest approach uses circularization of the fragments to fuse adaptors at both ends (Ingolia *et al.*, 2009). Prior to this, each fragment is polyadenylated at its 3' ends with poly(A)-polymerase (Ingolia *et al.*, 2009) which serves as a priming site for the reverse transcription. Polyadenylation was also introduced to produce uniform 3' ends of all fragments and to reduce the bias in the ligation (Ingolia, 2010), however the sequenced fragments are enriched in adenines at their 3' termini (Artieri & Fraser, 2014). Furthermore, in the circularization procedure an additional preference for adenine at the first 5'-position is observed (Lamm *et al.*, 2011, Artieri & Fraser, 2014): it does not depend on the polyA-tails of the fragments and the origin of this bias is unknown.

Later developments in the library preparation of ribosomal footprints use ligation approaches established in the sequencing of miRNAs, in which 3' and 5' adaptors are ligated sequentially to the fragments without circularization (Guo *et al.*, 2010). This allowed capture of low-abundance fragments and omitted the sequence bias (i.e. the preference for adenines at 5' and 3' positions). Note that direct ligation of a 3' adapter might be applied also as an alternative to polyA-tailing, preceding the circularization approach. However, some sequences in the libraries generated with sequential adaptor ligation were overrepresented compared to a sequencing in which the adaptors were ligated using the circularization protocol. The overrepresented fragments are a consequence of local secondary structure preferences (Hafner *et al.*, 2011, Zhuang *et al.*, 2012) and their propensity to co-fold with the adaptor sequences (Jackson *et al.*, 2014). Using truncated T4 RNA Ligase 2 instead of the previously used full-length, non-truncated version decreased the amount of those fragments by a half (Jackson *et al.*, 2014). Introducing short (2-4 nt) randomized sequences at the 5' and 3' ends of the adaptors also reduced the adaptor ligation bias (Jayaprakash *et al.*, 2011, Sorefan *et al.*, 2012, Zhang *et al.*, 2013).

2.3.7 Analysis of the sequencing results

The ribosomal footprints are very short (25-35 nt dependent on the organism, nucleolytic digestion protocol and manually excised region of the gel) and are usually sequenced by a single-end sequencing approach. The maximum number of total reads coming from a sequencing machine vary between sequencing samples (Mortazavi *et al.*, 2008, Garber *et al.*, 2011): for example our experience with various organisms (bacteria, mouse cell lines and tissues, plants and human samples) for which we performed ribosome profiling on Illumina HiSeq2000, have generated 40-195 million reads per sequencing lane. The final amount of reads correlates with the quality and quantity of the input material. The first step in the data processing undergoes an initial quality and adaptor trimming (Fig. 2.13). There is no uniform quality cut-off score and most ribosome profiling data are processed with a Phred score in the range ~20-30 or with 99.0-99.9% base accuracy (Ingolia *et al.*, 2012, Zhang *et al.*, 2012). In NGS data sets the quality drops towards the 3' end of the reads (Dohm *et al.*, 2008) which is also mirrored in the ribosome profiling libraries despite the short length of the fragments. Most of the tools used for this initial data processing (https://code.google.com/p/cutadapt/; http://hannonlab.cshl.edu/fastx_toolkit) (Lindgreen, 2012, Bolger *et al.*, 2014) also offer removal of reads with length shorter than expected upon adaptor cutting.

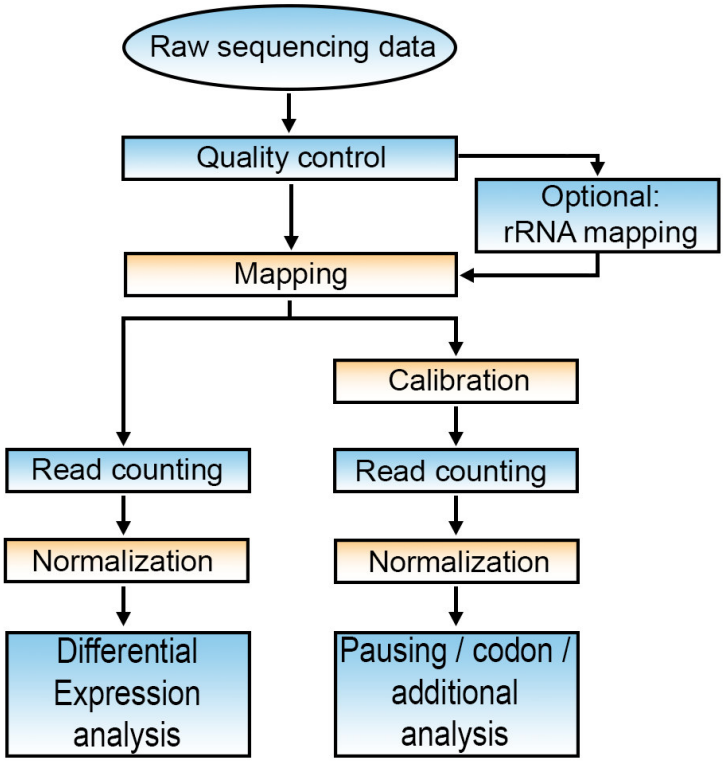


Figure 2.13 | Flow-chart of data analysis in ribosome profiling. Crucial steps are colorcoded in orange.

2.3.8 Read mapping

Read mapping is the most crucial procedure. Although principally the ribosomal footprints are in their core an RNA-Seq data set, there is no standardized pipeline with recommended mapping parameters. Mapping can be performed to genomes or transcriptomes, but the short single-end reads generated in the ribosome profiling experiment cannot be used for *de novo* assembly of genomes or transcriptomes (Simpson & Pop, 2015). Mapping to the genome should be preferred as it is unbiased towards known exon and intron annotations and allows for discovery of previously undescribed ORFs (Andreev *et al.*, 2015). Usually genome mapping gives greater coverage than mapping to transcriptomes (the loss of reads on exon junctions is minor) (Oshlack *et al.*, 2010). Furthermore, genomes are better defined than transcriptomes, which are constructed in several different ways (reviewed in (Garber *et al.*, 2011)). Also, mapping to genomes is less computationally intense and thus faster.

A prerequisite to good results is complete genome annotation, i.e. the availability of the gene coordinates. Genome annotation is a subject of intensive and constant improvement. For example the *E.coli* genome hosted on the NCBI server (Freddolino *et al.*, 2012) is updated daily and the number of genes constantly changes. Although this fast adjustment makes new findings immediately available, it creates a gap with the hand-curated databases, some of which may offer more precise annotation of additional features. For example, *RegulonDB* (Salgado *et al.*, 2013) offers more information on additional features than the NCBI annotations, including genes organized in operons, 5' and 3' UTRs. For eukaryotes the development is equally fast with frequently updated versions of genomes and their annotations. Three important webservers host various eukaryotic genomes: NCBI reference sequences, RefSeq (Pruitt *et al.*, 2007), *ensembl* (Cunningham *et al.*, 2015) and UCSC (Kent *et al.*, 2002). The genome annotation choice may significantly influence the downstream quantification of expression and differential analysis (Zhao & Zhang, 2015), although a simple advice on which database to use is not possible and should be driven by the purpose of the analysis. For research aiming at reproducible and robust gene expression estimates RefSeq might be preferred (Wu *et al.*, 2013). More exploratory questions may rely on more complex annotations, e.g. *ensembl*.

The mapping tools can be classified into two major groups: hash-table based (Li *et al.*, 2008, Homer *et al.*, 2009) or Burrows-Wheeler Transform (BWT) algorithms (Langmead *et al.*, 2009, Li & Durbin, 2009). While BWT-based approaches are faster and less computationally

demanding, the hash table-based algorithms are more flexible in aligning reads with non-perfect matches. Also the efficiency of BWT-based mapping approaches inversely correlates with the number of mismatches [reviewed in (Li & Homer, 2010, Garber *et al.*, 2011)]. Comparison of the tools is not trivial and differs depending on the data set, thus only few objective investigations have been performed so far (Giannoulatou *et al.*, 2014). In the majority of ribosome profiling experiments (Ingolia *et al.*, 2009, Guo *et al.*, 2010, Ingolia *et al.*, 2011, Gerashchenko *et al.*, 2012, Li *et al.*, 2012c, Chew *et al.*, 2013, Guttman *et al.*, 2013, Aspden *et al.*, 2014, Baudin-Baillieu *et al.*, 2014, Bazzini *et al.*, 2014, Subramaniam *et al.*, 2014), *Bowtie* (Langmead *et al.*, 2009) is used as a BWT-based mapping program. *Bowtie* offers two ways of mapping a read to a reference sequence: seed- (parameter n) and mismatch-based approach (parameter v, Table 2.1). The seed approach aligns first a seed (or core) of a read and then extends the alignment further along the read length. Thereby, the mismatches in the seed count stronger than those in the extensions. Mostly, default *Bowtie* parameters (parameter n for the seed-based approach) are used (Guo *et al.*, 2010, Li *et al.*, 2012c, Baudin-Baillieu *et al.*, 2014, Subramaniam *et al.*, 2014). Some studies apply the mismatch approach (Ingolia *et al.*, 2011, Gerashchenko *et al.*, 2012) which scores every base of each read equally. Since the default seed length of 28 nt remains unchanged when using the default parameter settings, the seed-based strategy effectively works as a mismatch approach.

A general drawback of *Bowtie* is its inability to map splice junctions. One commonly used tool to align short reads across junctions is *TopHat* (Trapnell *et al.*, 2009, Kim *et al.*, 2013) which can also find junctions de novo. First, the *TopHat* pipeline maps to all reads to a reference genome using *Bowtie* and allows reporting more than one alignment of a read (i.e. m=inf k=20 [translated to *Bowtie* parameters]). *TopHat* then assembles the mapped reads using the assembly module in Maq (Li *et al.*, 2008) in contiguous sequences inferring them to be putative exons, then uses seed and extended alignment to match reads to possible splice sites (Trapnell *et al.*, 2009). The pipeline of *TopHat* is more structured with fewer possibilities for changing the mapping parameters (Table 2.1), whereas *Bowtie* allows flexible adjustment of the mapping parameters. A new version of *Bowtie*, *Bowtie*2, has been launched (Langmead & Salzberg, 2012) which however differs conceptually from *Bowtie* and can find gapped alignments of reads resulting from insertions or deletions or sequencing errors. Note that *Bowtie*2 is suitable for reads longer than 50 nt.

Table 2.1 | *Bowtie* and *TopHat* mapping parameters and their effects.

Bowtie		*TopHat*		Description	Comments
Parameter	Default	Parameter	Default		
n	yes,2	--bowtie-n	no	Seed-based mapping approach	Mutually exclusive with v parameter
l	28	--b2-L (only *Bowtie2*)	20	Length of the seed	Only usable with n; default – too long for RPF reads
v	no	-N	yes,2	Mismatch-based mapping approach	Mutually exclusive with n parameter
m	infinite	NA, must be filtered after mapping		Maximum number of multiple positions per read	Uniquely mapping; m=1; try to avoid large numbers
best	no	always on		Reports alignment in best to worst order	Should be used always
strata	no	different scoring concept		Must be combined with best; reports best positions only	Should be used always
k	1	-g	20	Maximum number of reported alignment	Should be chosen carefully
a	no	NA		As k, but reports all alignment	Same as k=m
--	--	--bowtie1	no	Use *Bowtie* instead of *Bowtie2*	*Bowtie* does not allow changing mapping parameters

In general, mapping can be defined as a procedure to find the unique position of each read in the reference genome (Oshlack *et al.*, 2010). The logical consequence of this is to discard all reads with more than one best position. Since ribosomal footprints are very short, the proportion of reads mapping at more than one position increases with the size of the genome. In the ribosome profiling datasets a large fraction of the reads map at multiple positions to the genome (in the range of 30% and more), however there is no uniform strategy how to handle them. Strategies range from considering only reads uniquely mapping to the genome (Guo *et al.*, 2010, Baudin-Baillieu *et al.*, 2014), to allowing more than 200 or unlimited number of positions (Ingolia *et al.*, 2011). One strategy to estimate the true location of reads that map to many (multimappable) locations involves proportional assignment of reads based on the read density of the neighboring positions (Trapnell *et al.*, 2010). We varied the number of the mapping positions when aligning reads from ribosome profiling data (Fig. 2.14) and observed a clear difference between unique mapping (Fig. 2.14 B-C) and allowing multiple positions

(Fig. 2.14 D-E). The number of the mapped reads increases when multiple mapping positions are allowed (Fig. 2.14 A). Such scenarios are relevant mostly to genes with duplications or highly homologous isoforms, nevertheless the fraction of the discarded non-uniquely mappable reads can be in the range of 30%. Stringent criterion leads to sparse and incomplete coverage of each gene (Fig. 2.14 B-C), while allowing mapping to multiple positions in the genome improves significantly the coverage of a single gene (Fig. 2.14 D-E). The assignment of the reads mapping to multiple positions is of crucial importance also. Multimappable reads can be assigned randomly to one of the possible positions they map (Fig. 2.14 D) or to all possible positions (Fig. 2.14 E). However, choosing the mapping parameter in such a way that the first hit position is reported (Fig. 2.14 D) bears some caveats as the origin of the reads is unclear, i.e. whether they are from the same gene or originate from another position in the genome. Thus, it might artificially increase the total number of reads on a gene. In this context, mapping to multiple positions with equal weighting of all positions (Fig. 2.14 E) might be a better choice as it does not prefer between positions and maps uniformly to all best mappable positions. For some analysis, to avoid overinterpretation of the data (for example by differential analysis), the most conservative mapping with uniquely mappable reads (Fig. 2.14 C) might be the best choice.

The majority of the ribosome profiling datasets mapped with *Bowtie* do not set parameters to evaluate the quality of the alignments for a read (e.g., strata best) which compares for example, whether a zero-mismatch mapping is better than an alignment with two mismatches. Usually, the first encountered alignment of a read is assigned to it (Gerashchenko *et al.*, 2012, Li *et al.*, 2012c, Subramaniam *et al.*, 2014). Thus, when multimapping is allowed, a read with zero mismatches in a certain position may also be mapped to a different position with two mismatches. Consequently, it creates a bias since the best alignment would not be satisfied, but a read is randomly assigned to one of the two positions independent of the number of the mismatches. The choice of the parameters for the mapping are of crucial importance as they can result in significant variations in the mapping and gene coverage profiles (Fig. 2.14 B-E).

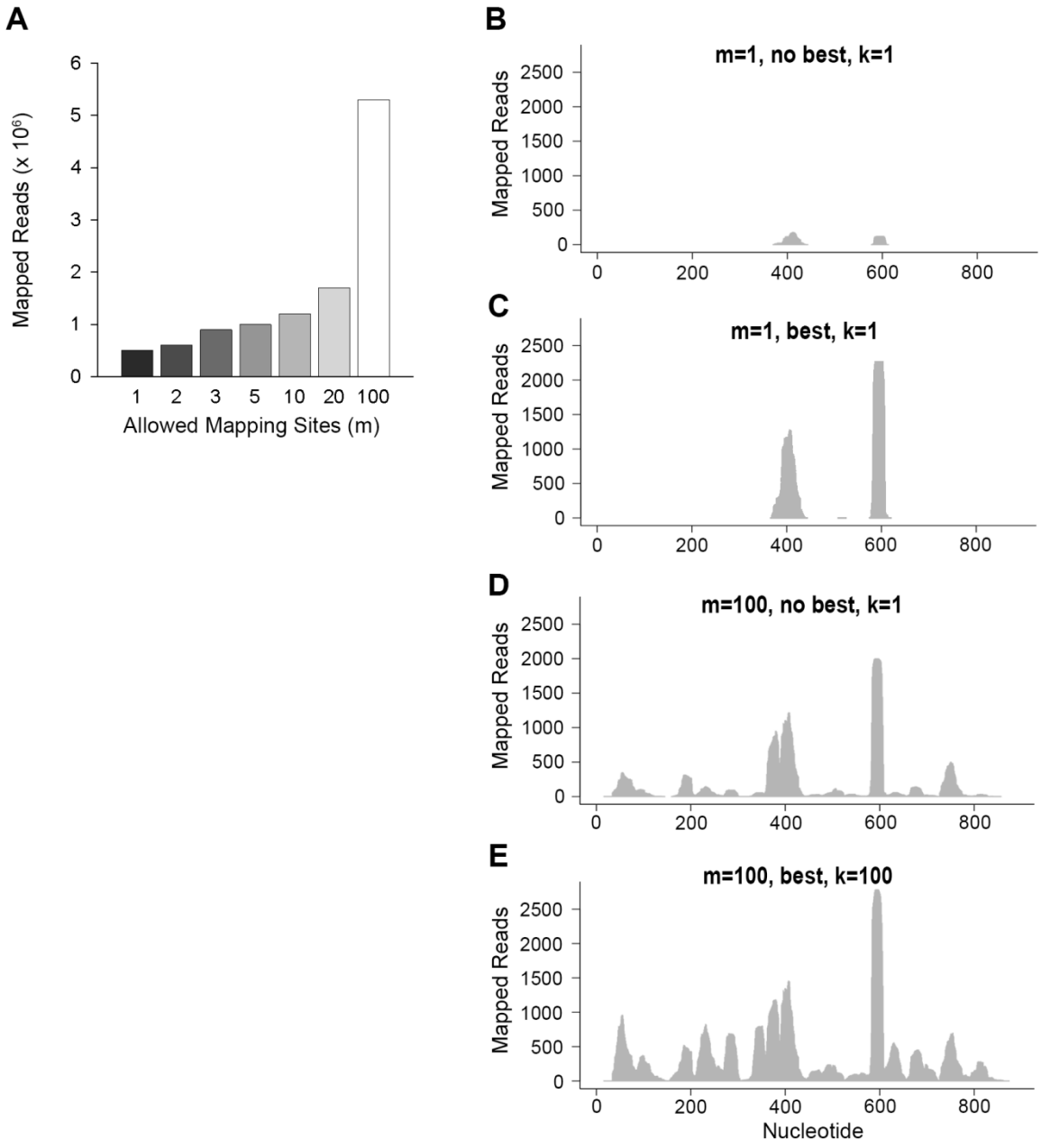


Figure 2.14 | Effects of different mapping strategies on gene coverage. (**A**) Total reads mapped to the genome allowing different number of maximal best mappable maximal best mapping positions per read (at m=1, a read with one best position (uniquely mappable) is allowed, otherwise discarded; at m=100 a read can be mapped to the 100 best positions and all positions are recorded). Mapping of ribosome profiling data of mouse brain was performed with Bowtie using the mouse genome (assembly GRCm38) allowing two mismatches per read (-v2 –strata –best). (**B-E**) Different mapping strategies result in variations in the read coverage of a human *rplA* gene mapped from ribosome profiling data with Bowtie using the human genome (assembly GRCm38). Mapping with a single hit (uniquely mappable) fulfilling the restrictions of maximum 2 mismatches (-v2 –m1) (**B**), with a single best hit (uniquely mappable) (-v2 –m1 –strata –best) (**C**), with default parameters, with restrictions to maximum 100 positions with 2 mismatches, but with only one listed in the output (-v2 –m100) (**D**), with multimapping restricted to 100 multimappable best positions (i.e. lowest number of mismatches) and best positions listed in the output (-v2 –m100 –k100) (**E**). Best, the parameters strata best are given to ensure that a multimappable position is counted as such only by the same minimum number of mismatches; no best, default mode with no strata best parameters chosen; m, maximum number of multiple positions per read; k, maximum number of reported alignment. Note settings as in B should be avoided. Parameters as in C show the most

reliable data, although the coverage is incomplete. The loss of reads in D compared to E is most likely due to the reporting of only one of the valid alignment positions.

For reproducibility of the results it is advisable to clearly state the mapping parameter in each publication.

Since both nucleases (RNase I for eukaryotic and MNase for bacterial systems) that are used to produce ribosomal footprints, cleave also rRNA, the rRNA reads comprise a large fraction of the sequencing reads, despite their removal in the experimental procedure. rRNA mapping and subtraction of those reads can be done in an extra round before or after mapping to the genome. Thereby, the mapping of the rRNA reads should be strict, i.e. allowing only a single mismatch.

In summary, mapping defines the shape of dataset to be used for further analysis and hence is a crucial step for which the parameters should be chosen carefully. In studies aiming at reproducible and robust gene expression estimates, uniquely mappable reads aligned to a reference genome (i.e. m=1 strata best) should be selected as they bear the lowest bias. However, for some genes (e.g. isoforms, duplicates), using only uniquely mapped reads may result in a partial coverage of a gene (Fig. 2.14 B-C); an incomplete coverage cannot be used to extract specific positions on which the ribosomes may pause or enrichment of reads over specific codons. For such analysis a multiple alignment of the multimappable reads (i.e. m=10 a strata best) might be chosen to ensure a maximal gene coverage. This parameter set bears drawbacks in analyzing the coverage of simultaneous expression of genes sharing large sequence identity (i.e. isoforms and duplicates). Such genes should be carefully assessed and might be separately compared only with their uniquely mappable reads or their expression should be confirmed with alternative methods (e.g. qRT-PCR).

2.3.9 Normalization of the read counts

Following mapping, read counts, also called gene counts, are collected and assigned to each gene or non-coding RNAs. Overlapping genes can be an issue here. Since the ribosome profiling protocol is strand-specific, overlapping genes on different strands are well resolved. For genes overlapping on the same strand, as commonly observed in *E.coli* in which the

coding sequence of the one gene falls into the end of the coding sequence of another, some read counting tools correct for this by randomly distributing the reads to the two overlapping genes (Anders *et al.*, 2015), while other tools do not recognize overlapping features (Quinlan, 2014).

A commonly applied approach for normalization of the read counts is reads per kilobase of exon per million mapped reads, rpkM, (Mortazavi *et al.*, 2008) which accounts for the differences in the sequencing depth (i.e. total number of the mapped reads) between sequencing libraries and for the length variation of each gene (i.e. per kilobase). Note that for short genes this normalization can give quite high rpkM values despite the presence of only few raw counts. Thus the detection limit should be set up using the raw counts (Ingolia *et al.*, 2011). Other normalization approaches frequently applied in the RNA sequencing (RNA-Seq) might be applied too (Anders & Huber, 2010, Robinson *et al.*, 2010, Dillies *et al.*, 2013) but the statistical behavior of ribosome profiling data with those normalization procedures has not yet been tested (Olshen *et al.*, 2013).

2.3.10 Further downstream analysis and post-processing

In the RNA-Seq datasets, several tools are used to identify differentially expressed (DE) genes (Guo *et al.*, 2013), some of which (e.g., *DESeq* tool) have been applied in a few ribosome profiling studies (Baudin-Baillieu *et al.*, 2014, Sidrauski *et al.*, 2015). Still they require a test that the ribosome profiling read counts follow the underlying distributions required by many tools designed for DE analysis of RNA-Seq, for example *DESeq* (Anders & Huber, 2010), *EdgeR* (Robinson *et al.*, 2010), and *baySeq* (Hardcastle & Kelly, 2010). In all cases, a careful and conservative interpretation of the data is needed since, unlike RNA-seq (Dillies *et al.*, 2013), no uniform pipeline exists for ribosome profiling data. So far, only one tool has been developed specifically for ribosome profiling data (Olshen *et al.*, 2013). Instead of performing DE analysis, a simple fold-change analysis can be carried out (Dunn *et al.*, 2013) with the assumption that most of the genes are unchanged.

Still, a fascinating issue of ribosome profiling is the ability to record the position of ribosomes with single nucleotide resolution (Ingolia *et al.*, 2009, Woolstenhulme *et al.*, 2015) which enables detecting ribosomal pausing (i.e. specific positions at which ribosomes pause) or encoding events (e.g. readthrough or frameshifting) (Li *et al.*, 2012c, Michel *et al.*, 2012,

O'Connor *et al.*, 2013). The alignment of the ribosomal reads to the open-reading frame is called calibration in which the start codon is assigned to the ribosomal P-site (Ingolia *et al.*, 2009) or the stop codon is assigned to the A-site (Woolstenhulme *et al.*, 2015). If the ribosomes are not completely halted during the isolation procedure, it will compromise the calibration and would not allow for codon resolution (Ingolia *et al.*, 2009). While ribosomal footprints of eukaryotic ribosomes can be calibrated using both stop and start codons, i.e. both 5' and 3' of the reads, reads from bacterial systems give only codon resolution when calibrated using their 3' ends, most likely because of the sharp cleavage of the MNase at the 3' of the reads but not at the 5' ends (Woolstenhulme *et al.*, 2015). Another approach to gain positional information of the translating ribosomes is center-weighted or center-assigned approach (Li *et al.*, 2012c). A defined number of nucleotides are excluded from both 5' and 3' sides of a read and the remaining centrally positioned nucleotides are weighted equally. This approach delivers less sharp resolution and defines the position of the ribosomal A- or P-sites with a subcodon resolution. Thus, it has limited applications and cannot be used for determining the reading frame (Woolstenhulme *et al.*, 2015). Both, calibrated and center-weighted ribosomal reads can be used to assess ribosomal enrichment over specific codons (Li *et al.*, 2012c, Ishimura *et al.*, 2014) or to determine sequences over which ribosomes transiently pause (Li *et al.*, 2012c, Woolstenhulme *et al.*, 2015).

In the library preparation, usually RNA fragments over a length range of 25-35 nt, tightly distributed around a peak of ~28 nt, are selected from the gel upon ribonucleolytic digestion (Ingolia *et al.*, 2009, Guydosh & Green, 2014). It should be noted that reads outside this range may also bear some biological information and, dependent on the specific question, might also be included in the library preparation. Reads shorter than the average length of ~28 nt represent different conformational states of the elongating ribosome (Lareau *et al.*, 2014) or report on ribosomes stalled over 3' truncated mRNAs (Guydosh & Green, 2014). In turn, longer reads may be informative on frameshifting events (O'Connor *et al.*, 2013). When comparing expression level on a gene basis in the DE analysis, all reads independent of their length might be considered under the assumption that each ribosome read produces one protein. For more specific analysis, including ribosomal stalling at specific positions, the reads should be separated by their length and each length group should be treated separately.

2.3.11 Computational demand and infrastructure

Raw data from one sequencing lane of Illuminas HiSeq machine can reach a size of more than 20 GB (uncompressed). Pre-processing and mapping of these raw files easily exceeds another 20 GB; discarding the intermediate preprocessing file and keeping only compressed raw files requires hard disk space for one lane of about 20 GB. The demand of RAM varies dependent on the type of analysis and programming languages. For example, using a simple Perl hash index build on each of the ~4 million nucleotides of the relatively small *E.coli* genome requires more than 4 GB of RAM. Mapping with BWT-based algorithms demands relatively low memory (Langmead *et al.*, 2009, Li & Durbin, 2009). For example, the human genome can be mapped with less than 8 GB of RAM (Langmead *et al.*, 2009, Li & Durbin, 2009). The mapping programs offer an option to use more than one CPU in parallel to increase speed (Langmead *et al.*, 2009). Many of pre- and post-processing steps are not implemented as full programs but as a collection of scripts or even in-house scripts (Anders *et al.*, 2015).

2.3.12 Conclusions

Ribosome profiling is a powerful technology to study translation *in vivo* on a cell-wide scale. While introducing this approach we are beginning to appreciate the variety of mechanisms that control translation and gene expression. However, non-standardized sample preparation and ambiguous processing of the data has produced some inconsistencies and has challenged direct comparisons between different studies. Experimentally, ribosome profiling is a multistep procedure which is in constant development and improvement of the single experimental steps. The task would be to understand the intrinsic bias of each step in order to carefully design the experimental protocol and interpret the data.

The analysis of data is complex, in part because of the short read lengths. Particularly crucial is the mapping procedure and normalization which defines the data set for further downstream analysis. The goal in the data analysis is to develop a uniform protocol, at least for mapping and normalization, as the broadness of the downstream analysis does not allow full standardization of this part of the pipeline. With the development of more standardized ribosome profiling technology and optimized sample preparation, we will move to a higher reproducibility of the data and a more accurate quantitative understanding of the mechanisms of translational control.

3. Discussion and conclusions

RNA molecules have much wider role than just being an information carrier between DNA and translating ribosomes. The possibility to carry out different regulatory function in the cell stems from its ability to fold in secondary or tertiary structures, which are essential for control of RNA transcription, splicing, translation, localization and turnover (Kozak, 2005, Garneau *et al.*, 2007, Cruz & Westhof, 2009, Martin & Ephrussi, 2009, Warf & Berglund, 2010, McManus & Graveley, 2011, Mauger *et al.*, 2013).

Although many recent studies evidenced the wide range of impact of mRNA folding on cellular processes, a precious knowledge on the transcriptome structure and its role in prokaryotes was missing. With this work, we aimed to fill this gap, providing a comprehensive analysis of the intrinsic structure propensity of the *E. coli* transcriptome, which brought us to identify structural features implicated in regulating translation initiation, termination and mRNA degradation.

The process of ribosome assembling on the mRNA to initiate translation has been well studied (Marintchev & Wagner, 2004, Simonetti *et al.*, 2009) and most of the regulatory factors have been already identified, however current models that predict translational rate and protein yields are able to predict only up to 70% of the translation efficiency (Salis *et al.*, 2009), suggesting that additional player(s), or interaction(s) are still escaping the identification.

For long time, the Shine-Dalgarno sequence was thought to be the main determinant of protein synthesis in bacteria (Ringquist *et al.*, 1992, Chen *et al.*, 1994, Kozak, 1999, Ma *et al.*, 2002, Osterman *et al.*, 2013); SD modulates the expression through the variation of two parameters: the distance of the SD from the start codon and the length of nucleotide stretch interacting with the aSD on the 16S rRNA, proportional to the hybridization energy. An optimal spacing was identified in a range of 5 to 13 nt (Chen *et al.*, 1994), with an optimal distance of 8–10 nt, if an adenine in SD sequence core is used as a reference point (Ringquist *et al.*, 1992, Chen *et al.*, 1994). Additionally, *in vitro* experiments mutating the SD and quantifying the expression of a reporter gene showed that the length of the SD sequence is proportional to the protein yield (Ringquist *et al.*, 1992, Osterman *et al.*, 2013), especially when the interaction with the 30S subunit involves the core the of anti-SD sequence than the off-center region (Ma *et al.*, 2002). Longer SD sequences are indeed more effective (Ringquist *et al.*, 1992), however too long SD have an inhibitory effect (Komarova *et al.*,

2002), which in turn augments translation only if the mRNA contains a secondary structure that limits ribosomal access to the AUG codon (de Smit & van Duin, 1994b) or if an alternative codon substitutes for AUG (Weyens *et al.*, 1988). Computing the minimum hybridization free energy (MHE) to the anti-SD sequence for all *E. coli* transcripts, we identified all SD sequences and noticed that they are clustered in three distinct classes of strengths (strong, medium and weak). In addition, there is a fourth class which lacks a SD. To our surprise, we did not observe any correlation between the SD-strength and translation efficiency, which we determined by the density of ribosomes (RPF) per mRNA. Highly translated genes did not preferably cluster in any of the SD groups resembling the distribution of all genes. Notably, some genes lacking an SD sequence also exhibit highly efficient translation, suggesting that additional factors are involved in translational initiation control.

Analysis of the secondary structure of the four SD groups showed how folding of this region was proportional to the SD strength, an observation easily explainable through the highest G content of longest (i.e. strongest) SD. Thus, our analysis reveals that SD sequences are generally occluded in secondary structures, a somewhat counterintuitive finding. Indeed, stable base-pairing at a translational initiation site in *Escherichia coli* can inhibit translation by competing with the binding of ribosomes. Van Duin and co-workers showed how SD folding is kinetically highly flexible and can be outcompeted by the ribosomes, though only if the 30S subunit is already in contact with the mRNA, to shift into place as soon as the structure opens (de Smit & van Duin, 2003). Although this model was also tested in single-molecule experiments (Studer & Joseph, 2006), whether this is a cell-wide regulatory strategy or a mechanism selected only for few genes it was still unclear. By analyzing the secondary structure of the four SD classes, we noticed that all the SD classes shared one striking feature: independent of the SD-strength and its secondary structure, the region upstream of the SD sequence, i.e. ~20 nt upstream of the start codon, is the most unstructured region of any gene. By swapping this region between different mRNAs, with more or less structured sequences in this region, we showed that the single-stranded level of this site positively regulates protein synthesis (Del Campo *et al.*, 2015). Although in this work neither ribosome profiling nor PARS analysis bear kinetic information or can reveal a sequence of binding events, we envision that the unfolded site upstream of the SD sequence may act as a primary unspecific docking site of the 30S subunit to enable interactions with the SD sequence within its unfolding window. Thus, we suggest that the level of "structureness" of the docking site is

indeed one of the factors influencing protein synthesis and that predicted "stand-by" model predicted by Van Duin is a general mechanism shared by all mRNAs in the transcriptome. Additionally, the contacts with the mRNA might be established by the essential S1 protein, which is the only ribosomal protein with an mRNA-binding affinity, explaining how the small-sized region can interact with the 30S subunit. S1 protein, which is essential for unfolding of stably structured SD (Duval *et al.*, 2013), attaches to the mRNA 11 nt upstream of the SD (Sengupta *et al.*, 2001) which is approximately the position of the unpaired region.

We also observed an enrichment of ribosomes upstream of a persistent secondary structure, which is found ~4-8 nt 5'-adjacent of the UAA stop codon. Previous research on termination regulation provides appropriate context for the interpretation of these results. The efficiency of translation termination (or conversely, the rate of termination suppression) is sensitive to the 5' and 3' sequence in immediate proximity of the stop codon (Bonetti *et al.*, 1995). An evident codon bias is present in many organisms and differs drastically between them (Cridge *et al.*, 2006), however the relationship between the nucleotide sequence and the effect on the ribosome was still unclear. The observation that the highest termination efficiency was achieved when codons encoding bulky amino acids were inserted upstream of the stop codon, despite the absence of a conserved sequence features, raised the hypothesis that interactions of these amino acids of the nascent peptide with the ribosomal tunnel would slow down terminating ribosome prior to termination, thus enhancing the termination fidelity (Bjornsson *et al.*, 1996). Further findings revealed that the accurate positioning of the A-site over the stop codon determines the accuracy in termination and suppresses the read through: A-rich sequences preceding the stop codons distort the ribosomes in the P-site, which alters the stop-codon decoding in the A-site (Tork *et al.*, 2004). In comparison, our analysis features a persistent mRNA secondary structure upstream of the UAA stop codon, which is not encoded by a universal sequence motif but is similarly responsible for a ribosomal slowdown. By drawing an analogy to these studies, we suggest that the secondary structure upstream of the UAA stop codon slows down the elongating ribosome, which may assist the accurate positioning of the ribosomal A-site for accurate decoding of the UAA stop codon. Nevertheless, additional experiments are needed to determine the interaction of these different elements and to clarify the mechanism through which the ribosome can "feel" the presence of a secondary structure close to the stop codon.

Differently from the region of the start and the stop codon, secondary structures present within the coding sequence do not affect the elongating ribosome *in vivo*, a feature observed in human and mouse cells (Rouskin *et al.*, 2014). This phenomenon was previously explained as the combined action of active, ATP-consuming mechanisms, like RNA helicases, and passive mechanisms, like translating ribosomes or ss-RNA binding proteins (Rouskin *et al.*, 2014). An alternative factor that could mask the impeding effect of RNA folding on translational speed is the selection for codon that pair to high-abundance tRNAs within highly structure regions, generating a trade-off between the two elements, which smooth the overall translation speed (Gorochowski *et al.*, 2015). Although generally not influencing the CDS, we identified a small set of persisting structured regions that transiently stall the ribosomes and may regulate protein integration into the membrane. These structures were probably selected for a functional reason and deserve further investigation to clarify their role in bacteria.

Another striking aspect of our analysis is the identification of a global signature of RNase E cleavage site. Earlier single-gene studies proposed the importance of secondary structures 5' upstream of the single-stranded cleavage site (Ehretsmann *et al.*, 1992, McDowall *et al.*, 1994, Moll *et al.*, 2003, Callaghan *et al.*, 2005). Our analysis corroborates those observations and features a structured region upstream of the A/U rich unpaired site as common signature of RNase E cleavage sites on a transcriptome-wide scale. This signature can be reconciled with the RNase E crystal structure: while a single-stranded segment only fits in the shallow channel leading to the RNase E active site (Callaghan *et al.*, 2005), the internal flexibility of the quaternary structure (Koslover *et al.*, 2008) can clearly accommodate secondary mRNA structures. The latter significantly shortens the distance between the cleavage site and 5' terminus and may explain how distant 5' termini of the mRNA facilitate catalysis (Callaghan *et al.*, 2005).

In summary, our approach of structural probing of bacterial mRNA *in vitro* with PARS, complemented with RNA-Seq and ribosome profiling, reveals structural features of importance for a variety of cellular processes. As also discussed in chapter 2.3, combined approaches that assess the role of secondary structure from multiple angles should be preferred over a single approach (Del Campo & Ignatova, 2015). Indeed, *in vivo* techniques, even though aiming at a picture directly in the living cell, could be biased by factors also acting *in vivo*: for example, nucleotide inaccessibility might result from protein or ribosome shielding and thus lead to misinterpretation of the results. Combining PARS with ribosome

profiling, which reveals the position of translating ribosomes with codon resolution, specifically addresses the impact of mRNA secondary structure for translating ribosomes and enabled us to discover additional regulatory elements of translation initiation in bacteria, which would have remained invisible for chemical imprinting as translation *in vivo* is not synchronized (Del Campo *et al.*, 2015). Additional complementation of the PARS approach with RNA sequencing allows deconvoluting the role of the secondary structure propensity and transcript abundance and extract cleavage signatures of different nucleases.

4. Methods

4.1 Materials

4.1.1 Chemicals and Reagents

All chemicals used in this study were purchased in the highest quality available. Chemicals used for RNA work were purchased with RNase-free quality and handled under RNase-free conditions. All the chemicals were purchased from Roth, except for the one listed below.

2x RNA Loading Dye	Thermo Scientific
Acid phenol-chloroform (5:1, pH 4.5)	Ambion
Chloroform:Isoamylalcohol (24:1)	Sigma-Aldrich
Glycogen (20 mg/ml) T	Thermo Scientific
Isopropanol	Merck
pCp-Cy3	Jena Bioscience
ddNTPs	Affymetrix
SYBR® Gold Nucleic Acid Gel Stain	Invitrogen
TRIzol reagent	Invitrogen

4.1.2 Enzymes

Enzymes used for RNA work were purchased RNase-free quality and handled under RNase-free conditions. Unless differently specified in the methods, all the enzymatic reactions were performed according to manufacturer instructions.

RNase V1 (0.1 U/µl)	Life Technologies
RNase A\T1	Thermo Scientific
T4 PNK	NEB
T4 RNA Ligase 1	NEB
T4 RNA Ligase 2, truncated	NEB
RevertAid™ H Minus Reverse Transcriptase	Thermo Scientific
Pfu DNA Polymerase	Thermo Scientific
Lysozyme	Roth
DNase I (RNase-free)	Thermo Scientific
Micrococcal nuclease	Thermo Scientific
T7 RNA Polymerase	Thermo Scientific
SUPERase•In™ RNase Inhibitor (20 U/µl)	Life Technologies

4.1.3 Oligonucleotides

All oligos were purchased from Metabion in desalted quality. Lyophilized oligos were reconstituted in molecular grade water (nuclease-free) and stored at -20 °C (DNA) or -80 °C (RNA). Sequences of oligos used in this study are shown in Table 4.1. The adenylated 3' adapter was purchased from Trilink Biotechnologies Inc., resuspended in DEPC treated H2O and stored at -80 °C.

Table 4.1 | List of primers used in this study.

Name	Sequence (5' → 3')	Application	Comments
CD_FW_ppiC_rt	TAATACGACTCACTATAGGGatgGCAAAAACAGCAGCAG	Assessing in vitro structure of labeled RNA	
CD_RV_ppiC	TTAGTTGCGGTACAGCACCTT		
CD_Fw_pET11b_WT	AGGGGAATTGTGAGCGGATAA	Sequencing primers for pET-Duet-1	Primers anneal at the ends of the multicloning site.
CD_Rv_pET11b_WT	ATGCTAGTTATTGCTCAGCGGT		
CD_FW_adhE	AAAAA**TCTAGA**ATACTCTCGTATTCGAGCAGATGATTTACTAAAAAAGTTTAACATTATCAGGAGAG CATTATGGCTGTTACTAATG	Cloning differently structured docking sites into pET-Duet-1-YFP	These oligos are designed to be be annealed, extended and then cloned between XbaI (orange) and NcoI (green). All the Fw oligos anneal to the Rv_adhE_all oligo, on the underlined regions. The FW oligos contain the 5'UTR (before SD) of the indicated genes, plus the SD (blue), ATG (red) and +42 nt of *adhE* gene.
CD_Rv_adhE_all	TTTTT**CCATGG**GCTCTACGAGTGCGTTAAGTTCAGCGACATTAGTAACAGCCATAATGCTCTCCT		
CD_FW_ppiD	AAAAA**TCTAGA**TGCGCGCATCGATACGTTGCGTGAGGTACACAGTCATCTACAGCAGGAGAGCATT ATGGCTGTTACTAATG		
CD_FW_cspE	AAAAA**TCTAGA**GACACAGCATTTGTGTCTATTTTTCATGTAAAGGAGAGCATTATGGCTGTTACTAA TG		
CD_FW_accD	AAAAATCTAGACATTATGCGTCCCCAAAGATAAAACTGGCATCGAACCAGGTTCAGACAGAAAGGA GAGCATTATGGCTGTTACTAATG		
CD_Rv_ppiD_all	TTTTTCCATGGTGAGCACGAGACTGTTTGCAGCCGTGCGTAAGCTGTCCATCATGGTGTAACAACAC TCC	Cloning differently structured docking sites into pET-Duet-1-YFP	These oligos will be annealed, extended and then cloned between XbaI (orange) and NcoI (green). All the Fw oligos anneal to the Rv_ppiD_all oligo, on the underlined regions. The FW oligos contain the 5'UTR (before SD) of the indicated genes, plus the SD (blue), ATG (red) and some CDS of ***ppiD*** gene.
CD_Fw_pp_ppiD	AAAAATCTAGATGCGCGCATCGATACGTTGCGTGAGGTACACAGTCATCTACAGCGGAGTGTTGTT ACACCATGATGGACAG		
CD_Fw_pp_adhE	AAAAATCTAGAATACTCTCGTATTCGAGCAGATGATTTACTAAAAAAGTTTAACATTATCGGAGTGT TGTTACACCATGATGGACAG		
CD_Fw_pp_cspE	AAAAATCTAGAGACACAGCATTTGTGTCTATTTTTCATGTAAGGAGTGTTGTTACACCATGATGGAC AG		
CD_Fw_pp_accD	AAAAATCTAGACATTATGCGTCCCCAAAGATAAAACTGGCATCGAACCAGGTTCAGACAGAAGGAG TGTTGTTACACCATGATGGACAG		

Table 4.1 (continuation)

Name	Sequence (5' → 3')	Application	Comments
CD_FW_ac_accD	AAAAATCTAGACATTATGCGTCCCCAAAGATAAAACTGGCATCGAACCAGGTTCAGACAGAAAGGT CCCTAatgAGCTGGATTGAACG	Cloning differently structured docking sites into pET-Duet-1-YFP	These oligos will be annealed, extended and then cloned between XbaI and NcoI. All the Fw oligos anneal to the Rv_accD_all oligo, on the underlined regions. The FW oligos contain the 5'UTR (before SD) of the indicated genes, plus the SD (blue), ATG (red) and some CDS of **accD** gene.
CD_FW_ac_adhE	AAAAATCTAGAATACTCTCGTATTCGAGCAGATGATTTACTAAAAAAGTTTAACATTATCAGGTCCCT AatgAGCTGGATTGAACG		
CD_Rv_accD_all	TTTTTCCATGGGGGTGGGAGTAATGTTGCTTTTAATTCGTTCAATCCAGCTcatTAGGGACCT		
SubHyb RPF	Biotin- GCCTCGTCATCACGCCTCAGCC	Subractive hybridization	
3' adapter	rApp/TGGAATTCTCGGGTGCCAAGG/3ddC/	TruSeq library preparation	rApp indicates 5' adenylation /3ddC/ indicates 3' dideoxycytosine (Guo *et al.,* 2010)
5' adapter	GUUCAGAGUUCUACAGUCCGACGAUC		
RT primer	CCTTGGCACCCGAGAATTCCA		
PCR primer 1	AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA		
PCR primer 2	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA		

4.1.3 Buffers

Buffers used for RNA work were prepared under RNase-free conditions from stock solutions using DEPC-treated H2O, filter sterilized and stored at 4 °C (unless stated otherwise).

10x RNA-structure buffer
100 mM Tris pH 7.0
1 M KCl
100 mM $MgCl_2$

Cold sucrose buffer
0.5 M RNase-free sucrose
50 mM KCl
16 mM Tris-HCl pH 8.1

Polysome lysis buffer
10 mM Tris pH 7.8
50 mM NH_4Cl
10 mM $MgCl_2$
0.2% triton X-100
100 µg/ml chloramphenicol
20 mM $CaCl_2$
10 U/ml DNase I (RNase-free)

Polysome lysis buffer with pH 9.2
10 mM Tris pH 11
50 mM NH_4Cl
10 mM $MgCl_2$
0.2% triton X-100
100 µg/ml chloramphenicol
20 mM $CaCl_2$

70% (w/v) Sucrose
35 g sucrose were dissolved step by step in 20-30 ml DEPC-treated H2O in a water bath (~70°C). DEPC-treated H2O was added to a final volume of 50 ml.

Sucrose gradients
Gradients for ultracentrifugation contain sucrose to a final concentration of 15%, 23%, 31%, 40%, 50%. Each concentration was supplemented with 1X Polysome lysis buffer and 0.35 mg/ml chloramphenicol.

2× Alkaline fragmentation solution
0.5 Vol 0.5 M EDTA
15 Vol 100 mM Na2CO3
110 Vol 100 mM NaHCO3
Solution was prepare fresh before using from stock solutions combined in the indicated ratios, resulting in an unadjusted pH of ~9.2.

Stop/Precipitation solution
60 µl 3 M NaACo (pH 5.5)
1.5 µl Glycogen
500 µl DEPC-H2O

20× SSC buffer
3 M NaCl
0.3 M Na-citrate, pH 7.0

10× T4 RNA ligase 2 buffer
500 mM Tris/HCl (pH 7.5)
20 mM MgCl2
10 mM DTT
The buffer was stored at -20 °C. SUPERase•InTM was added freshly prior to use.

10× T4 RNA ligase 1 buffer
500 mM Tris/HCl (pH 7.8)
100 mM MgCl2
100 mM DTT
Buffer was stored at -20 °C. SUPERase•InTM was added freshly prior to use.

DNA elution buffer
10 mM Tris/HCl (pH 8.0)
300 mM NaCl
1 mM EDTA

4.1.4 Kits

RevertAid H Minus First Strand cDNA Synthesis kit	Thermo
ERCC RNA Spike-In Control Mix	Ambion
TruSeq SBS Kit v3 – HS	Illumina
TruSeq SR Cluster Kit v3 cBot – HS	Illumina
Qubit dsDNA HS Assay	Life Technologies
RNA 6000 Nano kit	Agilent
DNA 1000 kit	Agilent
GeneJET™ RNA Purification Kit	Thermo Scientific
MICROBExpress™ Bacterial mRNA Enrichment Kit	Ambion
RNA Clean & Concentrator™ kit	Zymo Research
DNA1000 Chips	Agilent
µMACS Streptavidin Kit	Myltenyi Biotec
GeneJET Plasmid Miniprep Kit	Thermo Scientific

4.2 Methods

4.2.1 Enzymatic reaction and molecular biology techniques

All the enzymatic reactions were conducted according to the enzyme manufacturer instructions.

All the standard molecular biology methods (e.g., bacterial transformation, cloning) were performed according to "Molecular Cloning", by Green and Sambrook (Green & Sambrook, 2012).

4.2.2 RNA structural probing by deep sequencing

The *E. coli* MC4100 strain was cultured at 37 °C to mid-log phase (OD_{600} ~ 0.4) in LB media. Total RNA was extracted using TRIzol reagent and the sample was enriched in mRNA by depleting small RNAs with GeneJET™ RNA Purification Kit and ribosomal RNA with two cycle of MICROBExpress™ Bacterial mRNA Enrichment Kit, which reduces the amount of rRNA to appr. 25% of the total sequencing reads. To probe the RNA structure, two µg of enriched mRNA were resuspended in 45 µl of DEPC water and denatured for 2 min at 95°C, cooled on ice and slowly refolded increasing the temperature from 4° to 23°C in 20 min, after addition of 10x RNA-structure buffer with pH 7.0. The samples were digested for 1 min at 23°C with either 0.05 U RNase V1 or a combination of 2 µg RNase A and 5 U RNase T1 (1:5000 dilution of RNase A/T1 mix from Thermo scientific). The reaction was stopped by extracting the RNA with phenol-chloroform, followed by ethanol precipitation. During the optimization phase, the same steps were performed as above described, using different concentration of enzymes, as indicated in the paragraph 2.1.

The RNase A/T1-digested sample was phosphorylated with T4 PNK and purified with RNA Clean & Concentrator™ kit. Both the V1 and A/T1 digested samples were randomly fragmented in alkaline fragmentation solution for 12 min at 95°C. The reaction was stopped by adding 560 µl of stop/precipitation solution, followed by isopropanol precipitation. RNA fragments in the range of 50 to 200 nt were gel extracted from a 8M UREA, 6% PAGE and the cDNA libraries were prepared in a similar way to the one for RNA-Seq libraries (see paragraph 4.2.4), with small modification. First, the 5'-adapter is ligated with T4 RNA Ligase 1 overnight at 22 °C, followed by gel selection of ligated product in the range of 75-225 nt. Then, the 3'-end is dephosphorylated with T4 polynucleotide kinase (NEB) in the

corresponding buffer without ATP and the 3'-adapter is ligated, using T4 RNA Ligase 2, truncated. The fragments with adaptors at both termini were reverse transcribed with RevertAid™ H Minus Reverse Transcriptase using 5'-CAAGCAGAAGACGGCATACGA-3' primer and PCR-amplified with *Pfu* DNA Polymerase for 10 to 20 cycles (see table 4.1 a list of for primers). The DNA library size was determined with the Bioanalyzer using DNA1000 Chips and the concentration was measured using the Qubit dsDNA HS Assay. Sequencing was done with the TruSeq SBS Kit on a HiSeq2000 sequencing machine provided at the Sequencing Core Unit of the Max Delbrück Center for Molecular Medicine (Berlin, Germany).

4.2.3 Ribosome profiling

To isolate mRNA-bound ribosome complexes and extract the RPFs we used a previously described approach (Cozzone & Stent, 1973) with some modifications. MC4100 *E. coli* cells were cultivated to OD ~ 0.5, chloramphenicol added to 100 ug/ml and the culture was immediately rapidly cooled down by pouring through the crushed ice. Cells were harvested by centrifugation at 5000g for 5 min at 4°C. The pellet was resuspended in 12 ml of sucrose-buffer solution. To produce protoplasts, 0.3 ml of 10% EDTA (pH 8.0) and 0.3 ml of freshly dissolved lysozyme (50mg/ml) were added. The suspension was gently stirred for 5 min. 0.3 ml of 1M MgCl2 were added to stop lysozyme action and the cells were collected at 6000g for 10 min at 4 °C. The protoplasts from 100 ml culture were resuspended in 0.7ml freshly prepared lysis buffer. The mixture was clarified by centrifugation at 10000g for 10min at 4 °C, isolating the polysomes. For the isolation of RPFs, an aliquot of 100 A_{260} units of ribosome-bound mRNA fraction was subjected to nucleolytic digestion with 10 units/µl micrococcal nuclease for 10 min at room temperature in buffer with pH 9.2. The monosomal fraction was separated by sucrose density gradient (15-50% w/v) and collected. The RNA protected fragments were isolated from monosomes using the hot SDS/phenol method. Since micrococcal nuclease also cleaved rRNA into fragments with a size similar to the RPFs, an additional depletion step was introduced. The sample was enriched predominantly in one rRNA fragment which was removed by subtractive hybridization at 70 °C using a 5'-biotin-5'-GCCTCGTCATCACGCCTCAGCC-3' DNA oligonucleotide along with µMACS Streptavidin Kit to remove the biotin-labeled DNA/rRNA hybrids. Both randomly fragmented mRNAs and RPFs extracted from monosomes were denatured for 2 min at 80°C, and 3'-

dephosphorylated with T4 PNK for 90 min at 37°C in the corresponding buffer without ATP. RNA was precipitated by standard methods. Subsequently, 20-35-nt RNA fragments were size selected on a denaturing 15% polyacrylamide gel. The gel was extracted, precipitated and resuspended in DEPC water.

4.2.4 Random mRNA fragmentation and cDNA libraries

To generate the RNA-Seq sample to which the ribosome profiling data are compared, 20 µl of the enriched mRNA (as described above) was mixed with equal volume of 2x alkaline fragmentation solution and incubated for 40 min at 95°C. The reaction was stopped by adding 560 µl of stop/precipitation solution, followed by isopropanol precipitation.

Gel-purified RNA fragments from RPFs and fragmented mRNAs were dissolved in RNAse-free water and used for the preparation of the cDNA library via direct adapter ligation (Guo *et al.*, 2010), including some additional steps. A first adapter was ligated to the 3' end of the fragments, using T4 RNA Ligase 2, truncated, with a different formulation of reaction buffer. As both mRNA fragments and RPFs were hydroxylated at their 5'- and 3'-termini, they were initially 5'-phosphorylated with T4 polynucleotide kinase in ATP-containing buffer and successively the 5'-adapter was ligated at the 5'-termini by the T4 RNA Ligase 1, with a different formulation of reaction buffer. The fragments with adaptors at both termini were size selected on a denaturing 15% polyacrylamide gel, extracted and reverse transcribed with RevertAid™ H Minus Reverse Transcriptase and PCR-amplified with *Pfu* DNA Polymerase for 10 to 20 cycles. The DNA library size was determined with the Bioanalyzer using DNA1000 Chips (Agilent) and the concentration was measured using the Qubit dsDNA HS Assay (Life Technology). Sequencing was done with the TruSeq SBS Kit on a HiSeq2000 (Illumina) sequencing machine provided at the Sequencing Core Unit of the Max Delbrück Center for Molecular Medicine (Berlin, Germany).

4.2.5 Mapping of the sequencing reads

Sequenced reads were quality trimmed using *fastx-toolkit* (0.0.13.2; quality threshold: 20) and sequencing adapters were cut using *cutadapt* (1.2.1; minimal overlap: 1 nt) discarding reads shorter than 12 nucleotides. Processed reads were mapped to the *E. coli* genome (strain MG1655, version U00096.2, downloaded from NCBI) using Bowtie (0.12.9) allowing a maximum of two mismatches for the RNA-Seq and ribosome profiling data and a maximum

of three mismatches for the PARS data. Strain MC4100 is a derivative of MG1655 with four major deletions (Peters *et al.*, 2003)

The number of raw reads unambiguously aligned to ORFs in both RNA-Seq and ribosome profiling data sets, from two biological and one technical replicates were used to generate gene read counts, by counting the number of reads whose middle nucleotide (for even read length the nucleotide 5' of the mid-position) fell in the CDS. Gene read counts were normalized by the length of the unique CDS per kilobase (rpkM) and the total mapped reads per million (rpM) (Mortazavi *et al.*, 2008). In this mapping round, reads aligning to rRNA and tRNA genes were excluded since a large fraction of them map non-uniquely due to the multiple copies of those genes. Mapping of 5S and 16S RNA was done separately allowing no mismatches to only one copy of the rRNA reference sequence.

4.2.6 Computing the PARS score

The first nucleotide of the mapped reads from V1 or A/T1 digested samples, each derived from two biological replicates, was assigned to a nucleotide position in the genome and the counts were normalized to the sequencing depth. For each position, we computed the PARS score which is defined as the $\log_2$ of the ratio between the number of reads per million (rpM) from the V1-treated and the A/T1-treated samples (to each we added a small number 1, to avoid division by zero and to reduce the potential overestimating of low-coverage bases (Kertesz *et al.*, 2010)). RNase A hydrolyzes at single-stranded C and U nucleotides and RNase T1 at single-stranded G nucleotides, thus we excluded all adenines from the analyses. In addition, zero PARS score may result at positions with the same count values for A/T1 and V1 digestion, which are usually located in regions with highly flexible structure. As a minimum PARS coverage per transcript we used a threshold of 1.0 per transcript length (Fig. 2.4 A) termed transcript load (Kertesz *et al.*, 2010) which is defined as the sum of combined PARS readouts of the biological replicates per transcript divided by the effective transcript length (that is the annotated transcript length minus the number of unmappable nucleotides); the same threshold was used in yeast PARS analysis named as load of a transcript (Kertesz *et al.*, 2010). For the cumulative plots, all genes were aligned either to the start or the stop codon and for each position the mean of the PARS score of the two biological replicates was calculated. The GC content was calculated considering only the non-zero PARS score entries.

Periodicity of average PARS score in the CDSs and 5'UTR and 3'UTR was analyzed by Discrete Fourier transform (Fig. 7.2 A). The following regions were analyzed: over 10 to 99 nt downstream of the start codon, 99 to10 nt upstream of the stop codon (i.e. excluding possible influences of the initiation and termination codons but keeping the translation reading frame) for the CDSs, and 50 to 11 nt upstream of the start codon or downstream of the stop codon for the 5'UTR and 3'UTR, respectively. The periodicity for each of the three nucleotides in a codon was calculated also over the same region of the CDSs (Fig. 7.2 B).

4.2.7 Modeling the sampling error between biological replicates

To select a reliable minimum of read counts per gene and to assess the influence of counting noise, we computed the binomial partitioning of total counts between two independent biological replicates (Ingolia *et al.*, 2009) of the RNA-Seq and ribosome profiling from bacteria grown in LB. Genes were binned logarithmically based on the total number of their reads. The standard deviation of the ratio (repl#1/(repl#1 + repl#2)) across each bin was computed as a function of the mean sum of reads in each bin. In addition, a constant variance was added to the theoretical predictions accounting for other sources of error, yielding:

$$\sqrt{\frac{p(1-p)}{n} + s^2} \qquad (1)$$

where p represents the probability to assign a read to replicate #1, n is the total number of sequencing reads from replicate #1 and replicate #2 and s was obtained by fitting Eq. 1 to the data (Fig. 7.3 C,D).

4.2.8 Detection of RPF enrichment upstream of secondary structures

To determine positions whose secondary structure may influence elongation we used two approaches: CDS were systematically screened for double-stranded stretches (1) with a window of 10 nt containing 4 to 8 structured nt (i.e. with positive PARS score), or (2) using the mean PARS score within a window with different size (10 or 20 nt) (Fig. 7.4). A 10-nt-window with 6 structured nt delivered the best result considering the number of the selected positions (908 positions, Fig. 7.4) and was chosen in the analysis.

To define RPF enrichment upstream of a selected secondary structure (L_1), the RPF counts over 29 nt upstream of the double-stranded stretch (RPF1) were compared to the RPF counts

over 29 nt (1st-30th nt) downstream (L_2) of the detected stretch (RPF2). Read counts were normalized by the total number of reads for the whole region (Zhang *et al.*, 2014):

$$L_1 = \frac{RPF_1}{RPF_1 + RPF_2} \quad (2)$$

$$L_2 = \frac{RPF_2}{RPF_1 + RPF_2} \quad (3)$$

4.2.9 Determination of codon periodicity in the RPF and RNA-Seq data sets

Reads with length of 23-25 nt which were unambiguously mapped to the 1000 most expressed genes were combined for the RNA-Seq or ribosome profiling and binned by their length. To compute the codon periodicity in the RNA-Seq and ribosome profiling data sets, we used the reads mapped to the 3'-ends of the corresponding ORFs which were positioned at one of the three stop codons (UAG, UAA and UGA).

4.2.10 Detection of SD sequences

For all annotated genes, the MHE was calculated between sequences 1-25 nt upstream of the start codon and anti-SD sequence (3'-UCCUCCAC-5') using *RNAsubopt* (2.1.5; default parameters) from the *Vienna RNA Package* (Lorenz *et al.*, 2011). For each 8mer, the calculated MHE was assigned to the 8th base as described (Li *et al.*, 2012c) and the minimum of the calculated MHE of all 8mers was taken as an identifier for the SD sequence and used to determine the corresponding spacing. To designate different SD groups based on their MHE we used a randomization control. The random sample was created in two different ways: (1) by generating all possible random 8-mer sequences (65,536 sequences) or (2) by choosing each nucleotide randomly within the 8-mer (444,000 sequences). For both randomized groups we received similar results. For comparison to the natural SD, 4,400 random sequences were selected which resemble the *E. coli* gene number in Fig. 7.5 A.

4.2.11 Footprint analysis with fluorescently-labeled mRNA

In vitro transcribed RNA of *ppiC* was 3' end-labeled with 10 µM pCp-Cy3 using 15 units of T4 Ligase 1 (NEB). 2 µg of fluorescently-labeled RNA was structure probed with 0.05 U of RNase V1, in conditions identical to the PARS experiment. The digestion was stopped with

phenol/chlorophorm extraction, precipitated overnight at 4°C and resuspended in 10 µl of 2x RNA Loading Dye. In parallel, a ddNTP-Sanger sequencing PCR reaction was performed using 20 pmol of a 3'-fluorescently(Cy3)-labeled primer, in the presence of 400 ng of DNA template, 10 µM dNTPs, 1.25 U Pfu DNA Polymerase, Pfu Polymerase Buffer and 1 mM of each ddNTP. PCR was performed according to the manufacturer instructions in a volume of 15 µl. After addition of 2x RNA Loading Dye, all samples were boiled for 3 min at 95°C and loaded on a 6% PAGE, 7M UREA gel (50x40 cm), already pre-run for 30 min at 50W. The gel was then run for 3 h at 50W and the fluorescence was detected using a fluorescent gel imager.

4.2.12 Cloning and expression analysis

To create the construct bearing differently structured docking sites, genes were selected according to the 5 highest or lowest mean PARS score of the region comprised between -12 and -30 nt before the start codon, of genes having at least 5 bases of these region covered by information. Next, the genes were ordered according to the GC content and the one showing the highest and lowest GC percentage were chosen. Oligonucleotides were designed in order to clone the sequence in a pET-Duet-1 (Novagen), in which a YFP reporter gene was previously cloned (Bentele *et al.*, 2013). The oligonucleotides comprise the region from -60 to the Shine-Dalgarno of each selected gene, plus the region from the Shine-Dalgarno to +42 downstream of the start codon for the constant gene (*adhe, ppiD* or *accD*). Oligos were designed in order to be annealed and directly cloned, using a unique reverse primer and different forward primers. Cloning was performed according to standard molecular biology techniques (Green & Sambrook, 2012) and the obtained plasmids were transformed into competent *E. coli* DH5α cells. Plasmids were purified with GeneJET Plasmid Miniprep Kit, verified by Sanger sequencing and the one containing the wanted clone were transformed into *E. coli* BL21 cells.

To assess the differences in gene reporter expression, *E. coli* Bl21 cells were grown at 37 °C in LB medium till OD_{600}=0.5 and induced with 1 mM IPTG for 90 minutes. The median expression of the YFP-fused constructs was quantified in a population of approximately 10^5 cells by flow cytometry on a FACSCalibur (BD Bioscience), with the following settings: FSC=E01, log; SSC=381, log; FL1=600, log; threshold: SSC=90, primary, FL1=61. The fluorescence of the different constructs was recorded on a total number of 100000 events and

the data were processed by Flowing software 2. The values were normalized to the autofluorescence background of untransformed cells transformed. The mean of the median of 3 biological replicates was plotted, together with the standard error of the mean.

4.2.13 Statistical analysis

All data analyses were performed with in-house algorithms in Pearl and R. Differences between the distributions were assessed for significance by a nonparametric Mann-Whitney test, and enrichment of RPF was assessed by a Kolmogorov-Smirnov test. Note that we used Mann-Whitney U test, also called Wilcoxon rank-sum test, which is suitable for unpaired data for which no normal distribution can be assumed. To determine codon periodicity, Kullback-Leibler divergence was used to measure the deviation of the observed distribution of the 3'-end of the mapped read from a uniform distribution. Differences in the expression (FACS experiments) were evaluated using two-tailed Student's *t*-test. Differences were considered statistically significant when $P < 0.05$.

4.2.14 Data access

The sequencing data have been submitted to Gene Express Omnibus database currently as a read-only access for the peer-review:
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=ohmfcsowvvaxrwj&acc=GSE63817

5. References

Akashi, H., (1994) Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**: 927-935.

Al-Hashimi, H.M. & N.G. Walter, (2008) RNA dynamics: it is about time. *Curr Opin Struct Biol* **18**: 321-329.

Allen, G.S., A. Zavialov, R. Gursky, M. Ehrenberg & J. Frank, (2005) The cryo-EM structure of a translation initiation complex from Escherichia coli. *Cell* **121**: 703-712.

Anders, S. & W. Huber, (2010) Differential expression analysis for sequence count data. *Genome biology* **11**: R106.

Anders, S., P.T. Pyl & W. Huber, (2015) HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166-169.

Andreev, D.E., P.B. O'Connor, C. Fahey, E.M. Kenny, I.M. Terenin, S.E. Dmitriev, P. Cormican, D.W. Morris, I.N. Shatsky & P.V. Baranov, (2015) Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. *eLife* **4**: e03971.

Artieri, C.G. & H.B. Fraser, (2014) Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res* **24**: 2011-2021.

Aspden, J.L., Y.C. Eyre-Walker, R.J. Phillips, U. Amin, M.A. Mumtaz, M. Brocard & J.P. Couso, (2014) Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife* **3**: e03528.

Bakshi, S., A. Siryaporn, M. Goulian & J.C. Weisshaar, (2012) Superresolution imaging of ribosomes and RNA polymerase in live Escherichia coli cells. *Mol Microbiol* **85**: 21-38.

Baudin-Baillieu, A., R. Legendre, C. Kuchly, I. Hatin, S. Demais, C. Mestdagh, D. Gautheret & O. Namy, (2014) Genome-wide translational changes induced by the prion [PSI+]. *Cell reports* **8**: 439-448.

Bazzini, A.A., T.G. Johnstone, R. Christiano, S.D. Mackowiak, B. Obermayer, E.S. Fleming, C.E. Vejnar, M.T. Lee, N. Rajewsky, T.C. Walther & A.J. Giraldez, (2014) Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal* **33**: 981-993.

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

Bentele, K., P. Saffert, R. Rauscher, Z. Ignatova & N. Bluthgen, (2013) Efficient translation initiation dictates codon usage at gene start. *Mol Syst Biol* **9**: 675.

Berg, O.G. & C.G. Kurland, (1997) Growth rate-optimised tRNA abundance and codon usage. *J Mol Biol* **270**: 544-550.

Bjornsson, A., S. Mottagui-Tabar & L.A. Isaksson, (1996) Structure of the C-terminal end of the nascent peptide influences translation termination. *EMBO J* **15**: 1696-1704.

Blanchard, S.C., H.D. Kim, R.L. Gonzalez, Jr., J.D. Puglisi & S. Chu, (2004) tRNA dynamics on the ribosome during translation. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 12893-12898.

Blobel, G. & D. Sabatini, (1971) Dissociation of mammalian polyribosomes into subunits by puromycin. *Proceedings of the National Academy of Sciences of the United States of America* **68**: 390-394.

Bolger, A.M., M. Lohse & B. Usadel, (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Bonetti, B., L. Fu, J. Moon & D.M. Bedwell, (1995) The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in Saccharomyces cerevisiae. *J Mol Biol* **251**: 334-345.

Boni, I.V., V.S. Artamonova, N.V. Tzareva & M. Dreyfus, (2001) Non-canonical mechanism for translational control in bacteria: synthesis of ribosomal protein S1. *EMBO J* **20**: 4222-4232.

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

Brehm, S.L. & T.R. Cech, (1983) Fate of an intervening sequence ribonucleic acid: excision and cyclization of the Tetrahymena ribosomal ribonucleic acid intervening sequence in vivo. *Biochemistry* **22**: 2390-2397.

Brion, P. & E. Westhof, (1997) Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct* **26**: 113-137.

Bulmer, M., (1987) Coevolution of codon usage and transfer RNA abundance. *Nature* **325**: 728-730.

Burkhardt, N., R. Junemann, C.M. Spahn & K.H. Nierhaus, (1998) Ribosomal tRNA binding sites: three-site models of translation. *Crit Rev Biochem Mol Biol* **33**: 95-149.

Butler, J.S., M. Springer & M. Grunberg-Manago, (1987) AUU-to-AUG mutation in the initiator codon of the translation initiation factor IF3 abolishes translational autocontrol of its own gene (infC) in vivo. *Proc Natl Acad Sci U S A* **84**: 4022-4025.

Buxbaum, A.R., G. Haimovich & R.H. Singer, (2015) In the right place at the right time: visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol* **16**: 95-109.

Callaghan, A.J., M.J. Marcaida, J.A. Stead, K.J. McDowall, W.G. Scott & B.F. Luisi, (2005) Structure of Escherichia coli RNase E catalytic domain and implications for RNA turnover. *Nature* **437**: 1187-1191.

Cannarozzi, G., N.N. Schraudolph, M. Faty, P. von Rohr, M.T. Friberg, A.C. Roth, P. Gonnet, G. Gonnet & Y. Barral, (2010) A role for codon order in translation dynamics. *Cell* **141**: 355-367.

Carrier, T.A. & J.D. Keasling, (1997) Controlling messenger RNA stability in bacteria: strategies for engineering gene expression. *Biotechnol Prog* **13**: 699-708.

Cerretti, D.P., L.C. Mattheakis, K.R. Kearney, L. Vu & M. Nomura, (1988) Translational regulation of the spc operon in Escherichia coli. Identification and structural analysis of the target site for S8 repressor protein. *J Mol Biol* **204**: 309-329.

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

Charneski, C.A. & L.D. Hurst, (2013) Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol* **11**: e1001508.

Chen, C., H. Zhang, S.L. Broitman, M. Reiche, I. Farrell, B.S. Cooperman & Y.E. Goldman, (2013) Dynamics of translation by single ribosomes through mRNA secondary structures. *Nat Struct Mol Biol* **20**: 582-588.

Chen, H., M. Bjerknes, R. Kumar & E. Jay, (1994) Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. *Nucleic Acids Res* **22**: 4953-4957.

Chen, J., A. Petrov, M. Johansson, A. Tsai, S.E. O'Leary & J.D. Puglisi, (2014) Dynamic pathways of -1 translational frameshifting. *Nature* **512**: 328-332.

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

Chiba, S., A. Lamsa & K. Pogliano, (2009) A ribosome-nascent chain sensor of membrane protein biogenesis in Bacillus subtilis. *EMBO J* **28**: 3461-3475.

Chursov, A., D. Frishman & A. Shneider, (2013) Conservation of mRNA secondary structures may filter out mutations in Escherichia coli evolution. *Nucleic Acids Res* **41**: 7854-7860.

Clarke, J.E., L. Kime, A.D. Romero & K.J. McDowall, (2015) Direct entry by RNase E is a major pathway for the degradation and processing of RNA in Escherichia coli. *Nucleic Acids Res* **42**: 11733-11751.

Coleman, J.R., D. Papamichail, S. Skiena, B. Futcher, E. Wimmer & S. Mueller, (2008) Virus attenuation by genome-scale changes in codon pair bias. *Science* **320**: 1784-1787.

Corbin, R.W., O. Paliy, F. Yang, J. Shabanowitz, M. Platt, C.E. Lyons, Jr., K. Root, J. McAuliffe, M.I. Jordan, S. Kustu, E. Soupene & D.F. Hunt, (2003) Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc Natl Acad Sci U S A* **100**: 9232-9237.

Cox, R.A. & U.Z. Littauer, (1959) Secondary structure of ribonucleic acid in solution. *Nature* **184(Suppl 11)**: 818-819.

Cozzone, A.J. & G.S. Stent, (1973) Movement of ribosomes over messenger RNA in polysomes of rel + and rel - Escherichia coli strains. *J Mol Biol* **76**: 163-179.

Cridge, A.G., L.L. Major, A.A. Mahagaonkar, E.S. Poole, L.A. Isaksson & W.P. Tate, (2006) Comparison of characteristics and function of translation termination signals between and within prokaryotic and eukaryotic organisms. *Nucleic Acids Res* **34**: 1959-1973.

Cruz, J.A. & E. Westhof, (2009) The dynamic landscapes of RNA architecture. *Cell* **136**: 604-609.

Cunningham, F., M.R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C.G. Giron, L. Gordon, T. Hourlier, S.E. Hunt, S.H. Janacek, N. Johnson, T. Juettemann, A.K. Kahari, S. Keenan, F.J. Martin, T. Maurel, W. McLaren, D.N. Murphy, R. Nag, B. Overduin, A. Parker, M. Patricio, E. Perry, M. Pignatelli, H.S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S.P. Wilder, A. Zadissa, B.L. Aken, E. Birney, J. Harrow, R. Kinsella, M. Muffato, M. Ruffier, S.M. Searle, G. Spudich, S.J. Trevanion, A. Yates, D.R. Zerbino & P. Flicek, (2015) Ensembl 2015. *Nucleic acids research* **43**: D662-669.

Dahan, O., H. Gingold & Y. Pilpel, (2011) Regulatory mechanisms and networks couple the different phases of gene expression. *Trends Genet* **27**: 316-322.

Dana, A. & T. Tuller, (2014) The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res* **42**: 9171-9181.

Datta, A.K. & D.P. Burma, (1972) Association of ribonuclease I with ribosomes and their subunits. *The Journal of biological chemistry* **247**: 6795-6801.

de Smit, M.H. & J. van Duin, (1994a) Control of translation by mRNA secondary structure in Escherichia coli. A quantitative analysis of literature data. *J Mol Biol* **244**: 144-150.

de Smit, M.H. & J. van Duin, (1994b) Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol* **235**: 173-184.

de Smit, M.H. & J. van Duin, (2003) Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J Mol Biol* **331**: 737-743.

Del Campo, C., A. Bartholomaus, I. Fedyunin & Z. Ignatova, (2015) Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS Genet* **11**: e1005613.

Del Campo, C. & Z. Ignatova, (2015) Probing dimensionality beyond the linear sequence of mRNA. *Curr Genet*.

Dillies, M.A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, F. Jaffrezic & C. French StatOmique, (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics* **14**: 671-683.

Ding, Y., Y. Tang, C.K. Kwok, Y. Zhang, P.C. Bevilacqua & S.M. Assmann, (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**: 696-700.

Dingwall, C., G.P. Lomonossoff & R.A. Laskey, (1981) High sequence specificity of micrococcal nuclease. *Nucleic acids research* **9**: 2659-2673.

Dohm, J.C., C. Lottaz, T. Borodina & H. Himmelbauer, (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research* **36**: e105.

Dong, H., L. Nilsson & C.G. Kurland, (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* **260**: 649-663.

Draper, D.E. & P.H. von Hippel, (1978) Nucleic acid binding properties of Escherichia coli ribosomal protein S1. II. Co-operativity and specificity of binding site II. *J Mol Biol* **122**: 339-359.

Dresden, M. & M.B. Hoagland, (1965) Polyribosomes from Escherichia Coli: Enzymatic Method for Isolation. *Science* **149**: 647-649.

Dressaire, C., B. Laurent, P. Loubiere, P. Besse & M. Cocaign-Bousquet, (2010) Linear covariance models to examine the determinants of protein levels in Lactococcus lactis. *Mol Biosyst* **6**: 1255-1264.

Dunn, J.G., C.K. Foo, N.G. Belletier, E.R. Gavis & J.S. Weissman, (2013) Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife* **2**: e01179.

Duret, L., (2000) tRNA gene number and codon usage in the C. elegans genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet* **16**: 287-289.

Duval, M., A. Korepanov, O. Fuchsbauer, P. Fechter, A. Haller, A. Fabbretti, L. Choulier, R. Micura, B.P. Klaholz, P. Romby, M. Springer & S. Marzi, (2013) Escherichia coli ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol* **11**: e1001731.

Duval, M., A. Simonetti, I. Caldelari & S. Marzi, (2015) Multiple ways to regulate translation initiation in bacteria: Mechanisms, regulatory circuits, dynamics. *Biochimie* **114**: 18-29.

Eddy, S.R., (2014) Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu Rev Biophys* **43**: 433-456.

Ehresmann, C., F. Baudin, M. Mougel, P. Romby, J.P. Ebel & B. Ehresmann, (1987) Probing the structure of RNAs in solution. *Nucleic Acids Res* **15**: 9109-9128.

Ehretsmann, C.P., A.J. Carpousis & H.M. Krisch, (1992) Specificity of Escherichia coli endoribonuclease RNase E: in vivo and in vitro analysis of mutants in a bacteriophage T4 mRNA processing site. *Genes Dev* **6**: 149-159.

Espah Borujeni, A., A.S. Channarasappa & H.M. Salis, (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res* **42**: 2646-2659.

Fitch, W.M., (1974) The large extent of putative secondary nucleic acid structure in random nucleotide sequences or amino acid derived messenger-RNA. *J Mol Evol* **3**: 279-291.

Fluitt, A., E. Pienaar & H. Viljoen, (2007) Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput Biol Chem* **31**: 335-346.

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

Freddolino, P.L., S. Amini & S. Tavazoie, (2012) Newly identified genetic variations in common Escherichia coli MG1655 stock cultures. *Journal of bacteriology* **194**: 303-306.

Freistroffer, D.V., M.Y. Pavlov, J. MacDougall, R.H. Buckingham & M. Ehrenberg, (1997) Release factor RF3 in E.coli accelerates the dissociation of release factors RF1 and RF2 from the ribosome in a GTP-dependent manner. *EMBO J* **16**: 4126-4133.

Fritsch, C., A. Herrmann, M. Nothnagel, K. Szafranski, K. Huse, F. Schumann, S. Schreiber, M. Platzer, M. Krawczak, J. Hampe & M. Brosch, (2012) Genome-wide search for novel human uORFs and N-terminal protein extensions using ribosomal footprinting. *Genome research* **22**: 2208-2218.

Garber, M., M.G. Grabherr, M. Guttman & C. Trapnell, (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* **8**: 469-477.
Gardin, J., R. Yeasmin, A. Yurovsky, Y. Cai, S. Skiena & B. Futcher, (2014) Measurement of average decoding rates of the 61 sense codons in vivo. *Elife* **3**.
Garneau, N.L., J. Wilusz & C.J. Wilusz, (2007) The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol* **8**: 113-126.
Geissmann, T.A. & D. Touati, (2004) Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J* **23**: 396-405.
Gerashchenko, M.V. & V.N. Gladyshev, (2014) Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic acids research* **42**: e134.
Gerashchenko, M.V., A.V. Lobanov & V.N. Gladyshev, (2012) Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 17394-17399.
Giannoulatou, E., S.H. Park, D.T. Humphreys & J.W. Ho, (2014) Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie. *BMC bioinformatics* **15 Suppl 16**: S15.
Gingold, H. & Y. Pilpel, (2011) Determinants of translation efficiency and accuracy. *Mol Syst Biol* **7**: 481.
Godinic-Mikulcic, V., J. Jaric, B.J. Greber, V. Franke, V. Hodnik, G. Anderluh, N. Ban & I. Weygand-Durasevic, (2014) Archaeal aminoacyl-tRNA synthetases interact with the ribosome to recycle tRNAs. *Nucleic Acids Res* **42**: 5191-5201.
Gold, L., (1988) Posttranscriptional regulatory mechanisms in Escherichia coli. *Annu Rev Biochem* **57**: 199-233.
Gonzalez, C., J.S. Sims, N. Hornstein, A. Mela, F. Garcia, L. Lei, D.A. Gass, B. Amendolara, J.N. Bruce, P. Canoll & P.A. Sims, (2014) Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **34**: 10924-10936.
Goodman, D.B., G.M. Church & S. Kosuri, (2013) Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**: 475-479.
Gorochowski, T.E., Z. Ignatova, R.A. Bovenberg & J.A. Roubos, (2015) Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic Acids Res* **43**: 3022-3032.
Green, M.R. & J. Sambrook, (2012) *Molecular Cloning: A Laboratory Manual*. New York: Cold Spring Harbor Laboratory Press.
Groeneveld, H., K. Thimon & J. van Duin, (1995) Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? *RNA* **1**: 79-88.
Gualerzi, C.O. & C.L. Pon, (2015) Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cell Mol Life Sci* **72**: 4341-4367.
Guo, H., N.T. Ingolia, J.S. Weissman & D.P. Bartel, (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835-840.
Guo, Y., C.I. Li, F. Ye & Y. Shyr, (2013) Evaluation of read count based RNAseq analysis methods. *BMC genomics* **14 Suppl 8**: S2.
Gustafsson, C., S. Govindarajan & J. Minshull, (2004) Codon bias and heterologous protein expression. *Trends Biotechnol* **22**: 346-353.
Gutell, R.R., J.C. Lee & J.J. Cannone, (2002) The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol* **12**: 301-310.
Gutman, G.A. & G.W. Hatfield, (1989) Nonrandom utilization of codon pairs in Escherichia coli. *Proc Natl Acad Sci U S A* **86**: 3699-3703.
Guttman, M., P. Russell, N.T. Ingolia, J.S. Weissman & E.S. Lander, (2013) Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154**: 240-251.

Guydosh, N.R. & R. Green, (2014) Dom34 rescues ribosomes in 3' untranslated regions. *Cell* **156**: 950-962.
Hafner, M., N. Renwick, M. Brown, A. Mihailovic, D. Holoch, C. Lin, J.T. Pena, J.D. Nusbaum, P. Morozov, J. Ludwig, T. Ojo, S. Luo, G. Schroth & T. Tuschl, (2011) RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *Rna* **17**: 1697-1712.
Hajnsdorf, E. & I.V. Boni, (2012) Multiple activities of RNA-binding proteins S1 and Hfq. *Biochimie* **94**: 1544-1553.
Hardcastle, T.J. & K.A. Kelly, (2010) baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics* **11**: 422.
Hess, A.K., P. Saffert, K. Liebeton & Z. Ignatova, (2015) Optimization of translation profiles enhances protein expression and solubility. *PLoS One* **10**: e0127039.
Homer, N., B. Merriman & S.F. Nelson, (2009) BFAST: an alignment tool for large scale genome resequencing. *PloS one* **4**: e7767.
Hussmann, J.A., S. Patchett, A. Johnson, S. Sawyer & W.H. Press, (2015) Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet* **11**: e1005732.
Ikemura, T., (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol* **2**: 13-34.
Incarnato, D., F. Neri, F. Anselmi & S. Oliviero, (2014) Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biol* **15**: 491.
Ingolia, N.T., (2010) Genome-wide translational profiling by ribosome footprinting. *Methods in enzymology* **470**: 119-142.
Ingolia, N.T., (2014) Ribosome profiling: new views of translation, from single codons to genome scale. *Nature reviews. Genetics* **15**: 205-213.
Ingolia, N.T., G.A. Brar, S. Rouskin, A.M. McGeachy & J.S. Weissman, (2012) The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature protocols* **7**: 1534-1550.
Ingolia, N.T., G.A. Brar, N. Stern-Ginossar, M.S. Harris, G.J. Talhouarne, S.E. Jackson, M.R. Wills & J.S. Weissman, (2014) Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep* **8**: 1365-1379.
Ingolia, N.T., S. Ghaemmaghami, J.R. Newman & J.S. Weissman, (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**: 218-223.
Ingolia, N.T., L.F. Lareau & J.S. Weissman, (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**: 789-802.
Iost, I. & M. Dreyfus, (2006) DEAD-box RNA helicases in Escherichia coli. *Nucleic Acids Res* **34**: 4189-4197.
Ishimura, R., G. Nagy, I. Dotu, H. Zhou, X.L. Yang, P. Schimmel, S. Senju, Y. Nishimura, J.H. Chuang & S.L. Ackerman, (2014) RNA function. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science* **345**: 455-459.
Jackson, R. & N. Standart, (2015) The awesome power of ribosome profiling. *Rna* **21**: 652-654.
Jackson, T.J., R.V. Spriggs, N.J. Burgoyne, C. Jones & A.E. Willis, (2014) Evaluating bias-reducing protocols for RNA sequencing library preparation. *BMC genomics* **15**: 569.
Jacob, F. & J. Monod, (1961) Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**: 318-356.
Jan, C.H., C.C. Williams & J.S. Weissman, (2014) Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* **346**: 1257521.

Jayaprakash, A.D., O. Jabado, B.D. Brown & R. Sachidanandam, (2011) Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing. *Nucleic acids research* **39**: e141.

Julian, P., P. Milon, X. Agirrezabala, G. Lasso, D. Gil, M.V. Rodnina & M. Valle, (2011) The Cryo-EM structure of a complete 30S translation initiation complex from Escherichia coli. *PLoS Biol* **9**: e1001095.

Kaminishi, T., D.N. Wilson, C. Takemoto, J.M. Harms, M. Kawazoe, F. Schluenzen, K. Hanawa-Suetsugu, M. Shirouzu, P. Fucini & S. Yokoyama, (2007) A snapshot of the 30S ribosomal subunit capturing mRNA via the Shine-Dalgarno interaction. *Structure* **15**: 289-297.

Kanaya, S., Y. Yamada, Y. Kudo & T. Ikemura, (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**: 143-155.

Katz, L. & C.B. Burge, (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res* **13**: 2042-2051.

Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler & D. Haussler, (2002) The human genome browser at UCSC. *Genome research* **12**: 996-1006.

Kertesz, M., Y. Wan, E. Mazor, J.L. Rinn, R.C. Nutter, H.Y. Chang & E. Segal, (2010) Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**: 103-107.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley & S.L. Salzberg, (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**: R36.

Kim, H.K., F. Liu, J. Fei, C. Bustamante, R.L. Gonzalez, Jr. & I. Tinoco, Jr., (2014) A frameshifting stimulatory stem loop destabilizes the hybrid state and impedes ribosomal translocation. *Proc Natl Acad Sci U S A* **111**: 5538-5543.

Komar, A.A., (2009) A pause for thought along the co-translational folding pathway. *Trends Biochem Sci* **34**: 16-24.

Komarova, A.V., L.S. Tchufistova, E.V. Supina & I.V. Boni, (2002) Protein S1 counteracts the inhibitory effect of the extended Shine-Dalgarno sequence on translation. *RNA* **8**: 1137-1147.

Korkmaz, G., M. Holm, T. Wiens & S. Sanyal, (2014) Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem* **289**: 30334-30342.

Kortmann, J. & F. Narberhaus, (2012) Bacterial RNA thermometers: molecular zippers and switches. *Nat Rev Microbiol* **10**: 255-265.

Koslover, D.J., A.J. Callaghan, M.J. Marcaida, E.F. Garman, M. Martick, W.G. Scott & B.F. Luisi, (2008) The crystal structure of the Escherichia coli RNase E apoprotein and a mechanism for RNA degradation. *Structure* **16**: 1238-1244.

Kozak, M., (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene* **234**: 187-208.

Kozak, M., (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* **361**: 13-37.

Kramer, F.R. & D.R. Mills, (1981) Secondary structure formation during RNA synthesis. *Nucleic Acids Res* **9**: 5109-5124.

Kramer, G., D. Boehringer, N. Ban & B. Bukau, (2009) The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nature structural & molecular biology* **16**: 589-597.

Kudla, G., A.W. Murray, D. Tollervey & J.B. Plotkin, (2009) Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**: 255-258.

Kuersten, S., A. Radek, C. Vogel & L.O. Penalva, (2013) Translation regulation gets its 'omics' moment. *Wiley interdisciplinary reviews. RNA* **4**: 617-630.

Kunec, D. & N. Osterrieder, (2016) Codon Pair Bias Is a Direct Consequence of Dinucleotide Bias. *Cell Rep* **14**: 55-67.
Kwok, C.K., Y. Ding, Y. Tang, S.M. Assmann & P.C. Bevilacqua, (2013) Determination of in vivo RNA structure in low-abundance transcripts. *Nature communications* **4**: 2971.
Kwok, C.K., Y. Tang, S.M. Assmann & P.C. Bevilacqua, (2015) The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem Sci* **40**: 221-232.
Labuda, D., G. Striker & D. Porschke, (1984) Mechanism of codon recognition by transfer RNA and codon-induced tRNA association. *J Mol Biol* **174**: 587-604.
Lai, D., J.R. Proctor & I.M. Meyer, (2013) On the importance of cotranscriptional RNA structure formation. *RNA* **19**: 1461-1473.
Lamm, A.T., M.R. Stadler, H. Zhang, J.I. Gent & A.Z. Fire, (2011) Multimodal RNA-seq using single-strand, double-strand, and CircLigase-based capture yields a refined and extended description of the C. elegans transcriptome. *Genome research* **21**: 265-275.
Langmead, B. & S.L. Salzberg, (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357-359.
Langmead, B., C. Trapnell, M. Pop & S.L. Salzberg, (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**: R25.
Lareau, L.F., D.H. Hite, G.J. Hogan & P.O. Brown, (2014) Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife* **3**: e01257.
Laursen, B.S., H.P. Sorensen, K.K. Mortensen & H.U. Sperling-Petersen, (2005) Initiation of Protein Synthesis in Bacteria. *Microbiology and Molecular Biology Reviews* **69**: 101-123.
LeCuyer, K.A. & D.M. Crothers, (1994) Kinetics of an RNA conformational switch. *Proc Natl Acad Sci U S A* **91**: 3373-3377.
Lee, S., B. Liu, S. Lee, S.X. Huang, B. Shen & S.B. Qian, (2012) Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America* **109**: E2424-2432.
Li, F., Q. Zheng, P. Ryvkin, I. Dragomir, Y. Desai, S. Aiyer, O. Valladares, J. Yang, S. Bambina, L.R. Sabin, J.I. Murray, T. Lamitina, A. Raj, S. Cherry, L.S. Wang & B.D. Gregory, (2012a) Global analysis of RNA secondary structure in two metazoans. *Cell Rep* **1**: 69-82.
Li, F., Q. Zheng, L.E. Vandivier, M.R. Willmann, Y. Chen & B.D. Gregory, (2012b) Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *Plant Cell* **24**: 4346-4359.
Li, G.W., E. Oh & J.S. Weissman, (2012c) The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**: 538-541.
Li, H. & R. Durbin, (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.
Li, H. & N. Homer, (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics* **11**: 473-483.
Li, H., J. Ruan & R. Durbin, (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**: 1851-1858.
Lindgreen, S., (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. *BMC research notes* **5**: 337.
Liu, B., Y. Han & S.B. Qian, (2013) Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. *Molecular cell* **49**: 453-463.
Lorenz, R., S.H. Bernhart, C. Honer Zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler & I.L. Hofacker, (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.

Lu, P., C. Vogel, R. Wang, X. Yao & E.M. Marcotte, (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* **25**: 117-124.

Lucks, J.B., S.A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G.P. Schroth, L. Pachter, J.A. Doudna & A.P. Arkin, (2011) Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci U S A* **108**: 11063-11068.

Ma, J., A. Campbell & S. Karlin, (2002) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J Bacteriol* **184**: 5733-5745.

Mackie, G.A., (2013) RNase E: at the interface of bacterial RNA processing and decay. *Nat Rev Microbiol* **11**: 45-57.

Mahen, E.M., J.W. Harger, E.M. Calderon & M.J. Fedor, (2005) Kinetics and thermodynamics make different contributions to RNA folding in vitro and in yeast. *Mol Cell* **19**: 27-37.

Mahen, E.M., P.Y. Watson, J.W. Cottrell & M.J. Fedor, (2010) mRNA secondary structures fold sequentially but exchange rapidly in vivo. *PLoS Biol* **8**: e1000307.

Mao, Y., Q. Li, Y. Zhang, J. Zhang, G. Wei & S. Tao, (2013) Genome-wide analysis of selective constraints on high stability regions of mRNA reveals multiple compensatory mutations in Escherichia coli. *PLoS One* **8**: e73299.

Marintchev, A. & G. Wagner, (2004) Translation initiation: structures, mechanisms and evolution. *Q Rev Biophys* **37**: 197-284.

Martin, K.C. & A. Ephrussi, (2009) mRNA localization: gene expression in the spatial dimension. *Cell* **136**: 719-730.

Marzi, S., A.G. Myasnikov, A. Serganov, C. Ehresmann, P. Romby, M. Yusupov & B.P. Klaholz, (2007) Structured mRNAs regulate translation initiation by binding to the platform of the ribosome. *Cell* **130**: 1019-1031.

Masse, E., C.K. Vanderpool & S. Gottesman, (2005) Effect of RyhB small RNA on global iron use in Escherichia coli. *J Bacteriol* **187**: 6962-6971.

Mauger, D.M., N.A. Siegfried & K.M. Weeks, (2013) The genetic code as expressed through relationships between mRNA structure and protein function. *FEBS Lett* **587**: 1180-1188.

McDowall, K.J., S. Lin-Chao & S.N. Cohen, (1994) A+U content rather than a particular nucleotide order determines the specificity of RNase E cleavage. *J Biol Chem* **269**: 10790-10796.

McManus, C.J. & B.R. Graveley, (2011) RNA structure and the mechanisms of alternative splicing. *Curr Opin Genet Dev* **21**: 373-379.

Melnikov, S., A. Ben-Shem, N. Garreau de Loubresse, L. Jenner, G. Yusupova & M. Yusupov, (2012) One core, two shells: bacterial and eukaryotic ribosomes. *Nat Struct Mol Biol* **19**: 560-567.

Michel, A.M. & P.V. Baranov, (2013) Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley interdisciplinary reviews. RNA* **4**: 473-490.

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

Miettinen, T.P. & M. Bjorklund, (2015) Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions. *Nucleic acids research* **43**: 1019-1034.

Miller, O.L., Jr., B.A. Hamkalo & C.A. Thomas, Jr., (1970) Visualization of bacterial genes in action. *Science* **169**: 392-395.

Milon, P., C. Maracci, L. Filonava, C.O. Gualerzi & M.V. Rodnina, (2012) Real-time assembly landscape of bacterial 30S translation initiation complex. *Nat Struct Mol Biol* **19**: 609-615.

Mohammad, F., C.J. Woolstenhulme, R. Green & A.R. Buskirk, (2016) Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling. *Cell Rep* **14**: 686-694.
Moio, P., A. Kulyyassov, D. Vertut, L. Camoin, E. Ramankulov, M. Lipinski & V. Ogryzko, (2011) Exploring the use of dimethylsulfate for in vivo proteome footprinting. *Proteomics* **11**: 249-260.
Moll, I., T. Afonyushkin, O. Vytvytska, V.R. Kaberdin & U. Blasi, (2003) Coincident Hfq binding and RNase E cleavage sites on mRNA and small regulatory RNAs. *RNA* **9**: 1308-1314.
Morris, D.R., (2009) Ribosomal footprints on a transcriptome landscape. *Genome biology* **10**: 215.
Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer & B. Wold, (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.
Mortimer, S.A., M.A. Kidwell & J.A. Doudna, (2014) Insights into RNA structure and function from genome-wide studies. *Nat Rev Genet* **15**: 469-479.
Nakahigashi, K., Y. Takai, Y. Shiwa, M. Wada, M. Honma, H. Yoshikawa, M. Tomita, A. Kanai & H. Mori, (2014) Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. *BMC Genomics* **15**: 1115.
Nakatogawa, H., A. Murakami & K. Ito, (2004) Control of SecA and SecM translation by protein secretion. *Curr Opin Microbiol* **7**: 145-150.
Narberhaus, F., T. Waldminghaus & S. Chowdhury, (2006) RNA thermometers. *FEMS Microbiol Rev* **30**: 3-16.
Nedialkova, D.D. & S.A. Leidel, (2015) Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell* **161**: 1606-1618.
Nichols, N.M. & D. Yue, (2008) Ribonucleases. *Curr Protoc Mol Biol* **Chapter 3**: Unit3 13.
Noeske, J. & J.H. Cate, (2012) Structural basis for protein synthesis: snapshots of the ribosome in motion. *Curr Opin Struct Biol* **22**: 743-749.
O'Connor, P.B., G.W. Li, J.S. Weissman, J.F. Atkins & P.V. Baranov, (2013) rRNA:mRNA pairing alters the length and the symmetry of mRNA-protected fragments in ribosome profiling experiments. *Bioinformatics* **29**: 1488-1491.
Oh, E., A.H. Becker, A. Sandikci, D. Huber, R. Chaba, F. Gloge, R.J. Nichols, A. Typas, C.A. Gross, G. Kramer, J.S. Weissman & B. Bukau, (2011) Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* **147**: 1295-1308.
Olshen, A.B., A.C. Hsieh, C.R. Stumpf, R.A. Olshen, D. Ruggero & B.S. Taylor, (2013) Assessing gene-level translational control from ribosome profiling. *Bioinformatics* **29**: 2995-3002.
Orelle, C., S. Carlson, B. Kaushal, M.M. Almutairi, H. Liu, A. Ochabowicz, S. Quan, V.C. Pham, C.L. Squires, B.T. Murphy & A.S. Mankin, (2013) Tools for characterizing bacterial protein synthesis inhibitors. *Antimicrobial agents and chemotherapy* **57**: 5994-6004.
Oshlack, A., M.D. Robinson & M.D. Young, (2010) From RNA-seq reads to differential expression results. *Genome biology* **11**: 220.
Osterman, I.A., S.A. Evfratov, P.V. Sergiev & O.A. Dontsova, (2013) Comparison of mRNA features affecting translation initiation and reinitiation. *Nucleic Acids Res* **41**: 474-486.
Pan, T. & T. Sosnick, (2006) RNA folding during transcription. *Annu Rev Biophys Biomol Struct* **35**: 161-175.
Pechmann, S., J.W. Chartron & J. Frydman, (2014) Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nat Struct Mol Biol* **21**: 1100-1105.
Pelechano, V., W. Wei & L.M. Steinmetz, (2015) Widespread Co-translational RNA Decay Reveals Ribosome Dynamics. *Cell* **161**: 1400-1412.

Pestova, T.V. & C.U. Hellen, (2003) Translation elongation after assembly of ribosomes on the Cricket paralysis virus internal ribosomal entry site without initiation factors or initiator tRNA. *Genes & development* **17**: 181-186.

Peters, J.E., T.E. Thate & N.L. Craig, (2003) Definition of the Escherichia coli MC4100 genome by use of a DNA array. *J Bacteriol* **185**: 2017-2021.

Petrelli, D., C. Garofalo, M. Lammi, R. Spurio, C.L. Pon, C.O. Gualerzi & A. La Teana, (2003) Mapping the active sites of bacterial translation initiation factor IF3. *J Mol Biol* **331**: 541-556.

Picard, F., H. Milhem, P. Loubiere, B. Laurent, M. Cocaign-Bousquet & L. Girbal, (2012) Bacterial translational regulations: high diversity between all mRNAs and major role in gene expression. *BMC Genomics* **13**: 528.

Plotkin, J.B. & G. Kudla, (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* **12**: 32-42.

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

Proshkin, S., A.R. Rahmouni, A. Mironov & E. Nudler, (2010) Cooperation between translating ribosomes and RNA polymerase in transcription elongation. *Science* **328**: 504-508.

Pruitt, K.D., T. Tatusova & D.R. Maglott, (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **35**: D61-65.

Qian, W., J.R. Yang, N.M. Pearson, C. Maclean & J. Zhang, (2012) Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet* **8**: e1002603.

Qu, X., J.D. Wen, L. Lancaster, H.F. Noller, C. Bustamante & I. Tinoco, Jr., (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* **475**: 118-121.

Quax, T.E., N.J. Claassens, D. Soll & J. van der Oost, (2015) Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell* **59**: 149-161.

Quinlan, A.R., (2014) BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **47**: 11 12 11-11 12 34.

Ramani, V., R. Qiu & J. Shendure, (2015) High-throughput determination of RNA structure by proximity ligation. *Nat Biotechnol* **33**: 980-984.

Reichmann, M.E., M.W. Rees, R.H. Symons & R. Markham, (1962) Experimental evidence for the degeneracy of the nucleotide triplet code. *Nature* **195**: 999-1000.

Reid, D.W. & C.V. Nicchitta, (2012) Primary role for endoplasmic reticulum-bound ribosomes in cellular translation identified by ribosome profiling. *The Journal of biological chemistry* **287**: 5518-5527.

Repsilber, D., S. Wiese, M. Rachen, A.W. Schroder, D. Riesner & G. Steger, (1999) Formation of metastable RNA structures by sequential folding during transcription: time-resolved structural analysis of potato spindle tuber viroid (-)-stranded RNA by temperature-gradient gel electrophoresis. *RNA* **5**: 574-584.

Resch, A., B. Vecerek, K. Palavra & U. Blasi, (2010) Requirement of the CsdA DEAD-box helicase for low temperature riboregulation of rpoS mRNA. *RNA Biol* **7**: 796-802.

Reuter, J.A., D.V. Spacek & M.P. Snyder, (2015) High-throughput sequencing technologies. *Mol Cell* **58**: 586-597.

Ringquist, S., S. Shinedling, D. Barrick, L. Green, J. Binkley, G.D. Stormo & L. Gold, (1992) Translation initiation in Escherichia coli: sequences within the ribosome-binding site. *Mol Microbiol* **6**: 1219-1229.

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

Rodnina, M.V., A. Savelsbergh & W. Wintermeyer, (1999) Dynamics of translation on the ribosome: molecular mechanics of translocation. *FEMS Microbiol Rev* **23**: 317-333.

Rodnina, M.V. & W. Wintermeyer, (2011) The ribosome as a molecular machine: the mechanism of tRNA-mRNA movement in translocation. *Biochem Soc Trans* **39**: 658-662.

Romier, C., R. Dominguez, A. Lahm, O. Dahl & D. Suck, (1998) Recognition of single-stranded DNA by nuclease P1: high resolution crystal structures of complexes with substrate analogs. *Proteins* **32**: 414-424.

Ron, E.Z., R.E. Kohler & B.D. Davis, (1968) Magnesium ion dependence of free and polysomal ribosomes from Escherichia coli. *Journal of molecular biology* **36**: 83-89.

Rouskin, S., M. Zubradt, S. Washietl, M. Kellis & J.S. Weissman, (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**: 701-705.

Salgado, H., M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muniz-Rascado, J.S. Garcia-Sotelo, V. Weiss, H. Solano-Lira, I. Martinez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernandez, K. Alquicira-Hernandez, A. Lopez-Fuentes, L. Porron-Sotelo, A.M. Huerta, C. Bonavides-Martinez, Y.I. Balderas-Martinez, L. Pannier, M. Olvera, A. Labastida, V. Jimenez-Jacinto, L. Vega-Alvarado, V. Del Moral-Chavez, A. Hernandez-Alvarez, E. Morett & J. Collado-Vides, (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic acids research* **41**: D203-213.

Salis, H.M., E.A. Mirsky & C.A. Voigt, (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* **27**: 946-950.

Sato, K. & F. Egami, (1957) [The specificity of T1 ribonuclease]. *C R Seances Soc Biol Fil* **151**: 1792-1796.

Scharff, L.B., L. Childs, D. Walther & R. Bock, (2011) Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites. *PLoS Genet* **7**: e1002155.

Schmeing, T.M. & V. Ramakrishnan, (2009) What recent ribosome structures have revealed about the mechanism of translation. *Nature* **461**: 1234-1242.

Schmidt, M., P. Zheng & N. Delihas, (1995) Secondary structures of Escherichia coli antisense micF RNA, the 5'-end of the target ompF mRNA, and the RNA/RNA duplex. *Biochemistry* **34**: 3621-3631.

Schneider-Poetsch, T., J. Ju, D.E. Eyler, Y. Dang, S. Bhat, W.C. Merrick, R. Green, B. Shen & J.O. Liu, (2010) Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nature chemical biology* **6**: 209-217.

Schuwirth, B.S., M.A. Borovinskaya, C.W. Hau, W. Zhang, A. Vila-Sanjurjo, J.M. Holton & J.H. Cate, (2005) Structures of the bacterial ribosome at 3.5 A resolution. *Science* **310**: 827-834.

Scolnick, E., R. Tompkins, T. Caskey & M. Nirenberg, (1968) Release factors differing in specificity for terminator codons. *Proc Natl Acad Sci U S A* **61**: 768-774.

Seffens, W. & D. Digby, (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res* **27**: 1578-1584.

Sengupta, J., R.K. Agrawal & J. Frank, (2001) Visualization of protein S1 within the 30S ribosomal subunit and its interaction with messenger RNA. *Proc Natl Acad Sci U S A* **98**: 11991-11996.

Shabalina, S.A., A.Y. Ogurtsov & N.A. Spiridonov, (2006) A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Res* **34**: 2428-2437.

Shabalina, S.A., N.A. Spiridonov & A. Kashina, (2013) Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res* **41**: 2073-2094.

Shalgi, R., J.A. Hurt, I. Krykbaeva, M. Taipale, S. Lindquist & C.B. Burge, (2013) Widespread regulation of translation by elongation pausing in heat shock. *Molecular cell* **49**: 439-452.

Shine, J. & L. Dalgarno, (1974) The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A* **71**: 1342-1346.

Shine, J. & L. Dalgarno, (1975) Determinant of cistron specificity in bacterial ribosomes. *Nature* **254**: 34-38.

Shultzaberger, R.K., R.E. Bucheimer, K.E. Rudd & T.D. Schneider, (2001) Anatomy of Escherichia coli ribosome binding sites. *J Mol Biol* **313**: 215-228.

Sidrauski, C., A.M. McGeachy, N.T. Ingolia & P. Walter, (2015) The small molecule ISRIB reverses the effects of eIF2alpha phosphorylation on translation and stress granule assembly. *eLife* **4**.

Siegfried, N.A., S. Busan, G.M. Rice, J.A. Nelson & K.M. Weeks, (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat Methods* **11**: 959-965.

Simonetti, A., S. Marzi, I.M. Billas, A. Tsai, A. Fabbretti, A.G. Myasnikov, P. Roblin, A.C. Vaiana, I. Hazemann, D. Eiler, T.A. Steitz, J.D. Puglisi, C.O. Gualerzi & B.P. Klaholz, (2013) Involvement of protein IF2 N domain in ribosomal subunit joining revealed from architecture and function of the full-length initiation factor. *Proc Natl Acad Sci U S A* **110**: 15656-15661.

Simonetti, A., S. Marzi, L. Jenner, A. Myasnikov, P. Romby, G. Yusupova, B.P. Klaholz & M. Yusupov, (2009) A structural view of translation initiation in bacteria. *Cell Mol Life Sci* **66**: 423-436.

Simonetti, A., S. Marzi, A.G. Myasnikov, A. Fabbretti, M. Yusupov, C.O. Gualerzi & B.P. Klaholz, (2008) Structure of the 30S translation initiation complex. *Nature* **455**: 416-420.

Simpson, J.T. & M. Pop, (2015) The Theory and Practice of Genome Sequence Assembly. *Annual review of genomics and human genetics*.

Soper, T.J. & S.A. Woodson, (2008) The rpoS mRNA leader recruits Hfq to facilitate annealing with DsrA sRNA. *RNA* **14**: 1907-1917.

Sorefan, K., H. Pais, A.E. Hall, A. Kozomara, S. Griffiths-Jones, V. Moulton & T. Dalmay, (2012) Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence* **3**: 4.

Sosnick, T.R. & T. Pan, (2003) RNA folding: models and perspectives. *Curr Opin Struct Biol* **13**: 309-316.

Spirin, A.S., (2002) Translational Control in Prokaryotes. 309-338.

Spitale, R.C., P. Crisalli, R.A. Flynn, E.A. Torre, E.T. Kool & H.Y. Chang, (2013) RNA SHAPE analysis in living cells. *Nat Chem Biol* **9**: 18-20.

Spitale, R.C., R.A. Flynn, E.A. Torre, E.T. Kool & H.Y. Chang, (2014) RNA structural analysis by evolving SHAPE chemistry. *Wiley Interdiscip Rev RNA* **5**: 867-881.

Spitale, R.C., R.A. Flynn, Q.C. Zhang, P. Crisalli, B. Lee, J.W. Jung, H.Y. Kuchelmeister, P.J. Batista, E.A. Torre, E.T. Kool & H.Y. Chang, (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**: 486-490.

Stadler, M. & A. Fire, (2011) Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA* **17**: 2063-2073.

Starosta, A.L., J. Lassak, K. Jung & D.N. Wilson, (2014) The bacterial translation stress response. *FEMS Microbiol Rev* **38**: 1172-1201.

Steitz, J.A., (1969) Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature* **224**: 957-964.

Stormo, G.D., T.D. Schneider & L.M. Gold, (1982) Characterization of translational initiation sites in E. coli. *Nucleic Acids Res* **10**: 2971-2996.

Studer, S.M. & S. Joseph, (2006) Unfolding of mRNA secondary structure by the bacterial translation initiation complex. *Mol Cell* **22**: 105-115.

Subramaniam, A.R., B.M. Zid & E.K. O'Shea, (2014) An integrated approach reveals regulatory controls on bacterial translation elongation. *Cell* **159**: 1200-1211.
Sugimoto, Y., A. Vigilante, E. Darbo, A. Zirra, C. Militti, A. D'Ambrogio, N.M. Luscombe & J. Ule, (2015) hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature* **519**: 491-494.
Sussman, J.K., E.L. Simons & R.W. Simons, (1996) Escherichia coli translation initiation factor 3 discriminates the initiation codon in vivo. *Mol Microbiol* **21**: 347-360.
Talkish, J., G. May, Y. Lin, J.L. Woolford, Jr. & C.J. McManus, (2014) Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20**: 713-720.
Taniguchi, Y., P.J. Choi, G.W. Li, H. Chen, M. Babu, J. Hearn, A. Emili & X.S. Xie, (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**: 533-538.
Thirumalai, D. & C. Hyeon, (2005) RNA and protein folding: common themes and variations. *Biochemistry* **44**: 4957-4970.
Tijerina, P., S. Mohr & R. Russell, (2007) DMS footprinting of structured RNAs and RNA-protein complexes. *Nat Protoc* **2**: 2608-2623.
Tork, S., I. Hatin, J.P. Rousset & C. Fabret, (2004) The major 5' determinant in stop codon read-through involves two adjacent adenines. *Nucleic Acids Res* **32**: 415-421.
Toulme, F., C. Mosrin-Huaman, I. Artsimovitch & A.R. Rahmouni, (2005) Transcriptional pausing in vivo: a nascent RNA hairpin restricts lateral movements of RNA polymerase in both forward and reverse directions. *J Mol Biol* **351**: 39-51.
Trapnell, C., L. Pachter & S.L. Salzberg, (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105-1111.
Trapnell, C., B.A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M.J. van Baren, S.L. Salzberg, B.J. Wold & L. Pachter, (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28**: 511-515.
Tripathy, D.R., A.K. Dinda & S. Dasgupta, (2013) A simple assay for the ribonuclease activity of ribonucleases in the presence of ethidium bromide. *Anal Biochem* **437**: 126-129.
Tuller, T., A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman & Y. Pilpel, (2010a) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**: 344-354.
Tuller, T., Y.Y. Waldman, M. Kupiec & E. Ruppin, (2010b) Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci U S A* **107**: 3645-3650.
Tuller, T. & H. Zur, (2015) Multiple roles of the coding sequence 5' end in gene expression regulation. *Nucleic Acids Res* **43**: 13-28.
Turner, D.H., N. Sugimoto & S.M. Freier, (1988) RNA structure prediction. *Annu Rev Biophys Biophys Chem* **17**: 167-192.
Tyagi, K. & P.G. Pedrioli, (2015) Protein degradation and dynamic tRNA thiolation fine-tune translation at elevated temperatures. *Nucleic acids research* **43**: 4701-4712.
Tyrrell, J., J.L. McGinnis, K.M. Weeks & G.J. Pielak, (2013) The cellular environment stabilizes adenine riboswitch RNA structure. *Biochemistry* **52**: 8777-8785.
Underwood, J.G., A.V. Uzilov, S. Katzman, C.S. Onodera, J.E. Mainzer, D.H. Mathews, T.M. Lowe, S.R. Salama & D. Haussler, (2010) FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat Methods* **7**: 995-1001.
Vakulskas, C.A., A. Pannuri, D. Cortes-Selva, T.R. Zere, B.M. Ahmer, P. Babitzke & T. Romeo, (2014) Global effects of the DEAD-box RNA helicase DeaD (CsdA) on gene expression over a broad range of temperatures. *Mol Microbiol* **92**: 945-958.
van Dijk, E.L., Y. Jaszczyszyn & C. Thermes, (2014) Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research* **322**: 12-20.

Varenne, S., J. Buc, R. Lloubes & C. Lazdunski, (1984) Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *J Mol Biol* **180**: 549-576.

Vimberg, V., A. Tats, M. Remm & T. Tenson, (2007) Translation initiation region sequence preferences in Escherichia coli. *BMC Mol Biol* **8**: 100.

Vogt, V.M., (1973) Purification and further properties of single-strand-specific nuclease from Aspergillus oryzae. *Eur J Biochem* **33**: 192-200.

Wan, Y., M. Kertesz, R.C. Spitale, E. Segal & H.Y. Chang, (2011) Understanding the transcriptome through RNA structure. *Nat Rev Genet* **12**: 641-655.

Wan, Y., K. Qu, Z. Ouyang & H.Y. Chang, (2013) Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nat Protoc* **8**: 849-869.

Wan, Y., K. Qu, Z. Ouyang, M. Kertesz, J. Li, R. Tibshirani, D.L. Makino, R.C. Nutter, E. Segal & H.Y. Chang, (2012) Genome-wide measurement of RNA folding energies. *Mol Cell* **48**: 169-181.

Wan, Y., K. Qu, Q.C. Zhang, R.A. Flynn, O. Manor, Z. Ouyang, J. Zhang, R.C. Spitale, M.P. Snyder, E. Segal & H.Y. Chang, (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**: 706-709.

Warf, M.B. & J.A. Berglund, (2010) Role of RNA structure in regulating pre-mRNA splicing. *Trends Biochem Sci* **35**: 169-178.

Wen, J.D., L. Lancaster, C. Hodges, A.C. Zeri, S.H. Yoshimura, H.F. Noller, C. Bustamante & I. Tinoco, (2008) Following translation by single ribosomes one codon at a time. *Nature* **452**: 598-603.

Wettstein, F.O., T. Staehelin & H. Noll, (1963) Ribosomal aggregate engaged in protein synthesis: characterization of the ergosome. *Nature* **197**: 430-435.

Weyens, G., D. Charlier, M. Roovers, A. Pierard & N. Glansdorff, (1988) On the role of the Shine-Dalgarno sequence in determining the efficiency of translation initiation at a weak start codon in the car operon of Escherichia coli K12. *J Mol Biol* **204**: 1045-1048.

White, H.B., 3rd, B.E. Laux & D. Dennis, (1972) Messenger RNA structure: compatibility of hairpin loops with protein sequence. *Science* **175**: 1264-1266.

Wilkinson, K.A., E.J. Merino & K.M. Weeks, (2006) Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nat Protoc* **1**: 1610-1616.

Wilson, K.S. & P.H. von Hippel, (1995) Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc Natl Acad Sci U S A* **92**: 8793-8797.

Winkler, W., A. Nahvi & R.R. Breaker, (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**: 952-956.

Winkler, W.C. & R.R. Breaker, (2005) Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol* **59**: 487-517.

Woese, C.R., L.J. Magrum, R. Gupta, R.B. Siegel, D.A. Stahl, J. Kop, N. Crawford, J. Brosius, R. Gutell, J.J. Hogan & H.F. Noller, (1980) Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence. *Nucleic Acids Res* **8**: 2275-2293.

Wolin, S.L. & P. Walter, (1988) Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J* **7**: 3559-3569.

Wong, T.N., T.R. Sosnick & T. Pan, (2007) Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proc Natl Acad Sci U S A* **104**: 17995-18000.

Wood, C.R., M.A. Boss, T.P. Patel & J.S. Emtage, (1984) The influence of messenger RNA secondary structure on expression of an immunoglobulin heavy chain in Escherichia coli. *Nucleic Acids Res* **12**: 3937-3950.

Woolhead, C.A., P.J. McCormick & A.E. Johnson, (2004) Nascent membrane and secretory proteins differ in FRET-detected folding far inside the ribosome and in their exposure to ribosomal proteins. *Cell* **116**: 725-736.

Woolstenhulme, C.J., N.R. Guydosh, R. Green & A.R. Buskirk, (2015) High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell reports* **11**: 13-21.

Woolstenhulme, C.J., S. Parajuli, D.W. Healey, D.P. Valverde, E.N. Petersen, A.L. Starosta, N.R. Guydosh, W.E. Johnson, D.N. Wilson & A.R. Buskirk, (2013) Nascent peptides that block protein synthesis in bacteria. *Proc Natl Acad Sci U S A* **110**: E878-887.

Wu, P.Y., J.H. Phan & M.D. Wang, (2013) Assessing the impact of human genome annotation choice on RNA-seq expression estimates. *BMC bioinformatics* **14 Suppl 11**: S8.

Yakhnin, A.V., H. Yakhnin & P. Babitzke, (2006) RNA polymerase pausing regulates translation initiation by providing additional time for TRAP-RNA interaction. *Mol Cell* **24**: 547-557.

Yang, J.R., X. Chen & J. Zhang, (2014) Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol* **12**: e1001910.

Young, R. & H. Bremer, (1976) Polypeptide-chain-elongation rate in Escherichia coli B/r as a function of growth rate. *Biochem J* **160**: 185-194.

Youngman, E.M., S.L. He, L.J. Nikstad & R. Green, (2007) Stop codon recognition by release factors induces structural rearrangement of the ribosomal decoding center that is productive for peptide release. *Mol Cell* **28**: 533-543.

Yu, C.H., Y. Dang, Z. Zhou, C. Wu, F. Zhao, M.S. Sachs & Y. Liu, (2015) Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol Cell* **59**: 744-754.

Yusupov, M.M., G.Z. Yusupova, A. Baucom, K. Lieberman, T.N. Earnest, J.H. Cate & H.F. Noller, (2001) Crystal structure of the ribosome at 5.5 A resolution. *Science* **292**: 883-896.

Zamecnik, P., (2005) From protein synthesis to genetic insertion. *Annu Rev Biochem* **74**: 1-28.

Zhang, G., I. Fedyunin, S. Kirchner, C. Xiao, A. Valleriani & Z. Ignatova, (2012) FANSe: an accurate algorithm for quantitative mapping of large scale sequencing reads. *Nucleic acids research* **40**: e83.

Zhang, G., M. Hubalewska & Z. Ignatova, (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat Struct Mol Biol* **16**: 274-280.

Zhang, G. & Z. Ignatova, (2009) Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PLoS One* **4**: e5036.

Zhang, G. & Z. Ignatova, (2011) Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *Current opinion in structural biology* **21**: 25-31.

Zhang, Y., R.A. Mooney, J.A. Grass, P. Sivaramakrishnan, C. Herman, R. Landick & J.D. Wang, (2014) DksA guards elongating RNA polymerase against ribosome-stalling-induced arrest. *Mol Cell* **53**: 766-778.

Zhang, Z., J.E. Lee, K. Riemondy, E.M. Anderson & R. Yi, (2013) High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome biology* **14**: R109.

Zhao, S. & B. Zhang, (2015) A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. *BMC genomics* **16**: 97.

Zheng, Q., P. Ryvkin, F. Li, I. Dragomir, O. Valladares, J. Yang, K. Cao, L.S. Wang & B.D. Gregory, (2010) Genome-wide double-stranded RNA sequencing reveals the

functional significance of base-paired RNAs in Arabidopsis. *PLoS Genet* **6**: e1001141.
Zhuang, F., R.T. Fuchs & G.B. Robb, (2012) Small RNA expression profiling by high-throughput sequencing: implications of enzymatic manipulation. *Journal of nucleic acids* **2012**: 360358.
Ziehler, W.A. & D.R. Engelke, (2001) Probing RNA structure with chemical reagents and enzymes. *Curr Protoc Nucleic Acid Chem* **Chapter 6**: Unit 6 1.
Zur, H. & T. Tuller, (2012) Strong association between mRNA folding strength and protein abundance in S. cerevisiae. *EMBO Rep* **13**: 272-277.

6. Appendix

6.1 Hazard statements (H statements)

H225 – Highly flammable liquid and vapour.

H226 – Flammable liquid and vapour.

H227 – Combustible liquid

H228 – Flammable solid.

H290 – May be corrosive to metals.

H301 – Toxic if swallowed.

H302 – Harmful if swallowed.

H311 – Toxic in contact with skin.

H312 – Harmful in contact with skin.

H314 – Causes severe skin burns and eye damage.

H315 – Causes skin irritation.

H317 – May cause an allergic skin reaction.

H318 – Causes serious eye damage.

H319 – Causes serious eye irritation.

H331 – Toxic if inhaled.

H332 – Harmful if inhaled.

H335 – May cause respiratory irritation.

H336 – May cause drowsiness or dizziness.

H340 – May cause genetic defects.

H341 – Suspected of causing genetic defects.

H350 – May cause cancer.

H351 – Suspected of causing cancer.

H361 – Suspected of damaging fertility or the unborn child.

H370 – Causes damage to organs.

H372 – Causes damage to organs through prolonged or repeated exposure.

H373 – May cause damage to organs through prolonged or repeated exposure.

H412 – Harmful to aquatic life with long lasting effects.

6.2 Precautionary statements (P statements)

P210 – Keep away from heat/sparks/open flames/hot surfaces. — No smoking.

P233 – Keep container tightly closed.

P261 – Avoid breathing dust/fume/gas/mist/vapours/spray.

P280 – Wear protective gloves/protective clothing/eye protection/face protection.

P281 – Use personal protective equipment as required.

P310 – Immediately call a POISON CENTER or doctor/physician.

P314 – Get medical advice/attention if you feel unwell.

P301 + P310 – IF SWALLOWED: Immediately call a POISON CENTER or doctor/physician.

P301 + P312 – IF SWALLOWED: Call a POISON CENTER or doctor/physician if you feel unwell.

P301 +P330 + P331 – IF SWALLOWED: rinse mouth. Do NOT induce vomiting.

P302 + P352 – IF ON SKIN: Wash with plenty of soap and water.

P303 + P361 + P353 – IF ON SKIN (or hair): Remove/Take off immediately all contaminated clothing. Rinse skin with water/shower.

P304 + P340 – IF INHALED: Remove victim to fresh air and keep at rest in a position comfortable for breathing.

P304 + P341 – IF INHALED: If breathing is difficult, remove victim to fresh air and keep at rest in a position comfortable for breathing.

P305 + P351 + P338 – IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing.

P308 + P310 – IF exposed or concerned: Immediately call a POISON CENTER or doctor/physician.

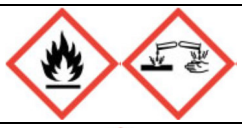
P308 + P313 – IF exposed or concerned: Get medical advice/attention.

P309 + P310 – IF exposed or if you feel unwell: Immediately call a POISON CENTER or doctor/physician.

P370 + P378 – In case of fire: Use … for extinction.

P406 – Store in corrosive resistant/… container with a resistant inner liner.

6.2 List of hazardous substances used in the study

Chemical	Pictogram	H statement	P statement
Acrylamide		H301-H312-H315-H317-H319-H332-H340-H350-H361f-H372	P280-P302+P352-P305+P351+P338
Acetic Acid		H226-H314	P280-P301+P330+P331-P305+P351+P338
Bisacrylamide		H302	
Chloroform		H302, H315, H319, H332,H336, H351, H361, H373	P261, P281, P305+351+338
Dithiothreitol		H302-H315-H319-H335	P261-P280-P301+P312-P304+P340
Ethanol		H225	P210-P233
Ethidium bromide		H332-H341	P281-P308+P313
Ethylenediaminetetraacetic acid		H319	P305+351+338
Formaldehyde		H301-H311-H314-H317-H331-H351-H370-H335	P281--P308+ P310-P303+P361+P353-P304+P340-P305+P351+P338
Hydrochloric acid		H290-H314-H335	P280-P301+P330+P331-P305+P351+P338
Isopropanol		H225-H319-H336	P210-P233-P305+P351+P338
Phenol		H301, H311, H314, H331,H341, H373	P261, P280, P301+310,P305+351+338, P310

Chemical	Pictogram	H statement	P statement
Sodium carbonate		H319	P305+351+338
Sodium dodecyl sulfate		H228-H302+H332-H315-H318-H335-H412	P210-P280-P302+P352-P304+P340-P305+P351+P338-P314
Sodium hydroxide		H290-H314	P280-P303+P361+P353-P301+P330+P331-P305+P351+P338-P309+P310-P406
Syber gold		H227	P210, P280, P370 + P378
Tetramethyl-ethylendiamin			P210-P233-P280-P301+P330+P331-P305+P351+P338
Tris		H225-H302-H314-H332	P261-P280-P302+P352-P305+P351+P338-P304+P340
Xylen cyanol		H350	

7. Supplementary materials

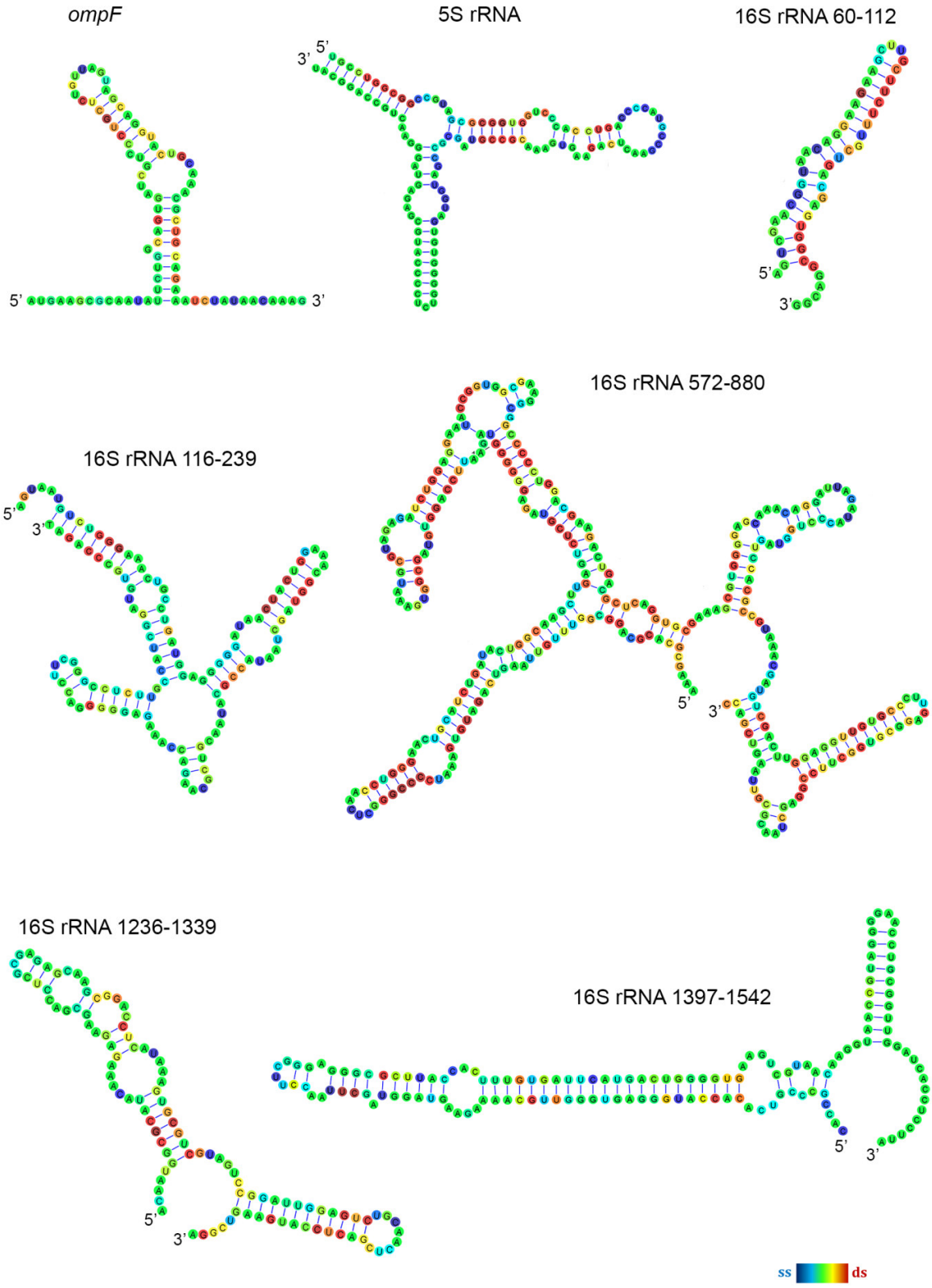


Figure 7.1 | Correlation between the PARS score and transcripts with known secondary structure. The PARS score was overlaid with the determined with OmpF (Schmidt *et al.*, 1995), 5S rRNA (Woese *et al.*, 1980) and 16S rRNA structure. The color intensity of the nucleotides reflects the PARS scores (rainbow legend). For more details on the PARS-based colorcoding see the legend to Fig 2.4 B. For 16S rRNA, PARS score was overlaid with the determined structure. Solvent exposed helices were selected from the crystal structure (Yusupov *et al.*, 2001, Gutell *et al.*, 2002) and overlaid with the experimentally determined PARS values. The solvent-exposed regions are cleaved first and this first phase of nucleolysis reports on the native structure allowing for more conservative PARS analysis. Nucleotides 60–107 –helix 6; nt 116–239 –helix 7 to 10; nt 572–880 –helix 20 to 26; nt 1236–133 –helices 41 and 42; nt 1397–1542–44 and 45. The color intensity of the 16S rRNA nucleotides reflects the magnitude of the PARS scores.

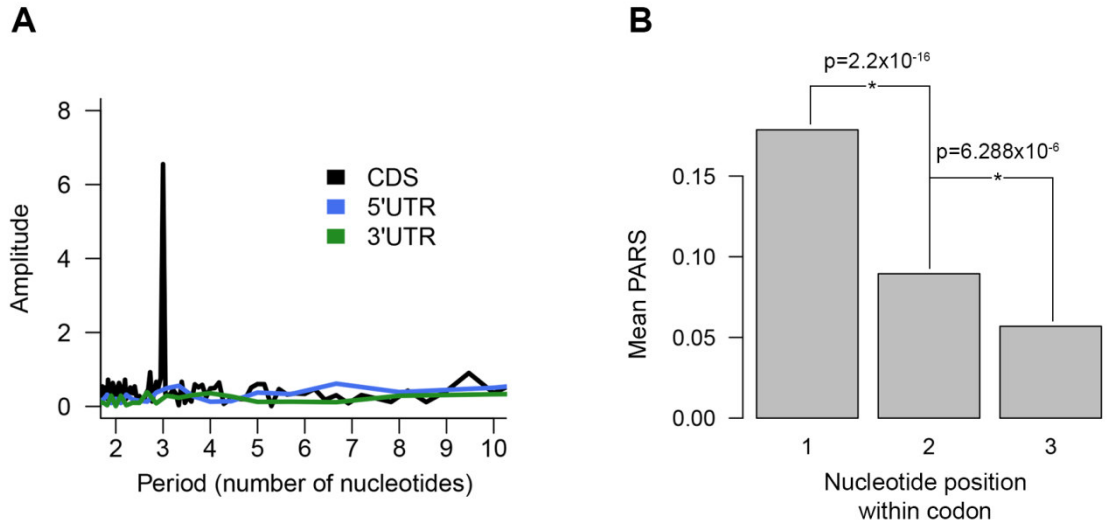


Figure 7.2 | Periodicity in the structure of the *E. coli* CDSs. (**A**) Discrete Fourier transform analysis. Analyses were performed with the average PARS score over 10 to 99 nt downstream of the start codon, 99 to10 nt upstream of the stop codon for the CDSs, and 50 to 11 nt upstream of the start codon or downstream of the stop codon for the 5'UTR and 3'UTR, respectively. (**B**) Average PARS score for each of the three nucleotides of a codon, averaged across all codons.

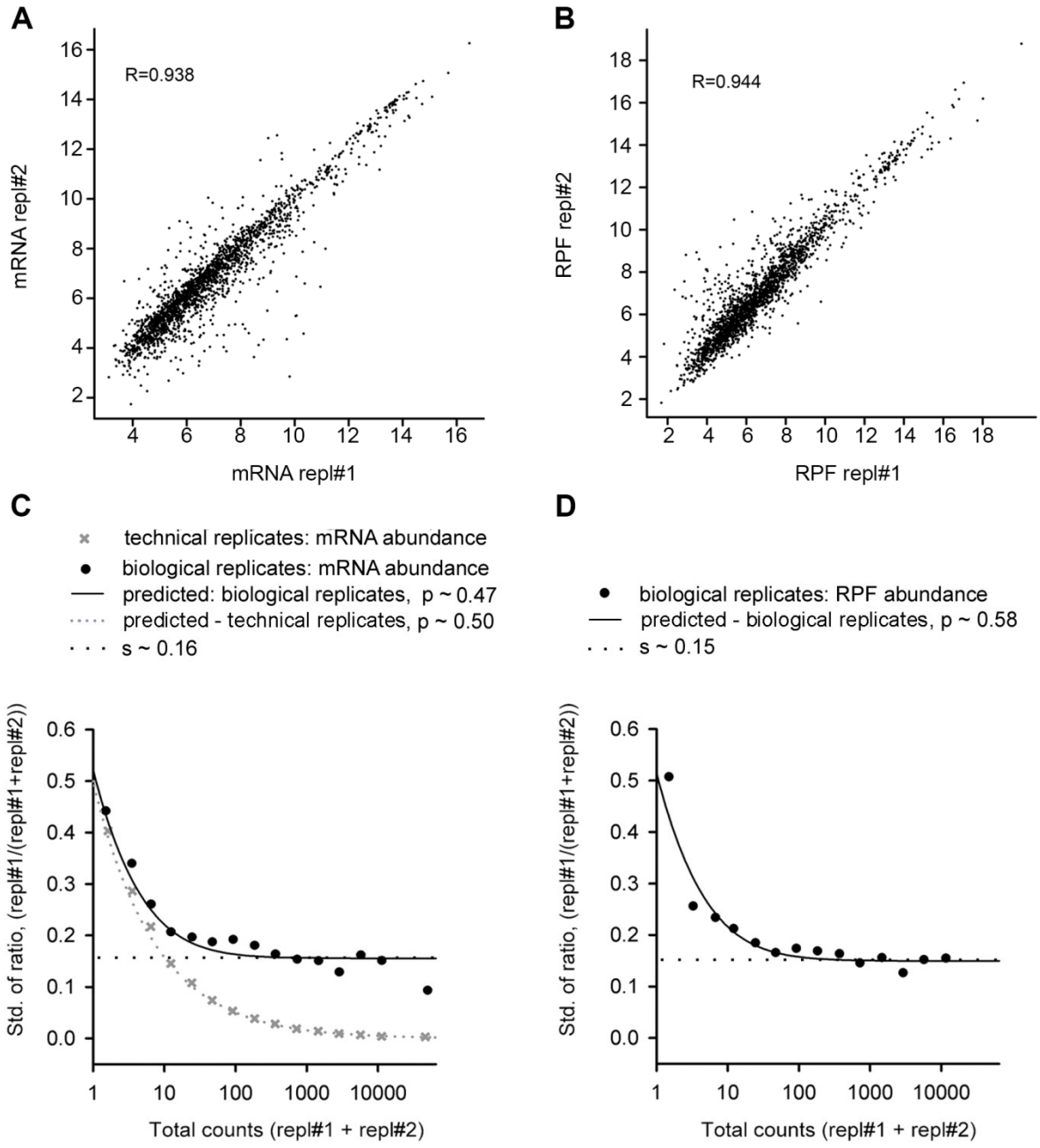


Figure 7.3 | Reproducibility and variability analysis of the RNA-Seq and Ribo-Seq. (**A, B**) Reproducibility of randomly fragmented mRNAs (**A**) and RPFs (**B**) of two biological replicates. The Pearson correlation coefficients, calculated between the $\log_2$ of the read coverage for each transcript with counts > 60 (see panel C, D) indicate that the RiboSeq and RNA-Seq analyses are highly reproducible. (**C, D**) Variability analysis of counting statistics on the error in quantification of RNA-Seq (**C**) and ribosome profiling (**D**). The two independent biological and technical mRNA (A) and RPF (B) replicates were used to estimate the biological variation compared to the technical one. The technical replicates are dominated by counting noise, thus s = 0 (Eq. 1). A threshold of 120 total counts (i.e., 60 counts for each replicate) was chosen as for total reads >120 the variability approached the infinite-counts asymptote and the contribution of the counting statistics was little. For the RNA-Seq data set the fitting parameters are $p = 0.47$ and s = 0.16, and for the RPF data set are $p = 0.58$ and s = 0.15. By setting a threshold to 60 reads both in mRNA-Seq and RPF-analysis, the technical error is smaller than 5% of the biological variation. In total, 1.955 genes have >60 mRNA and RPF reads and have PARS over the selected threshold of 1 (Fig 2.4 A).

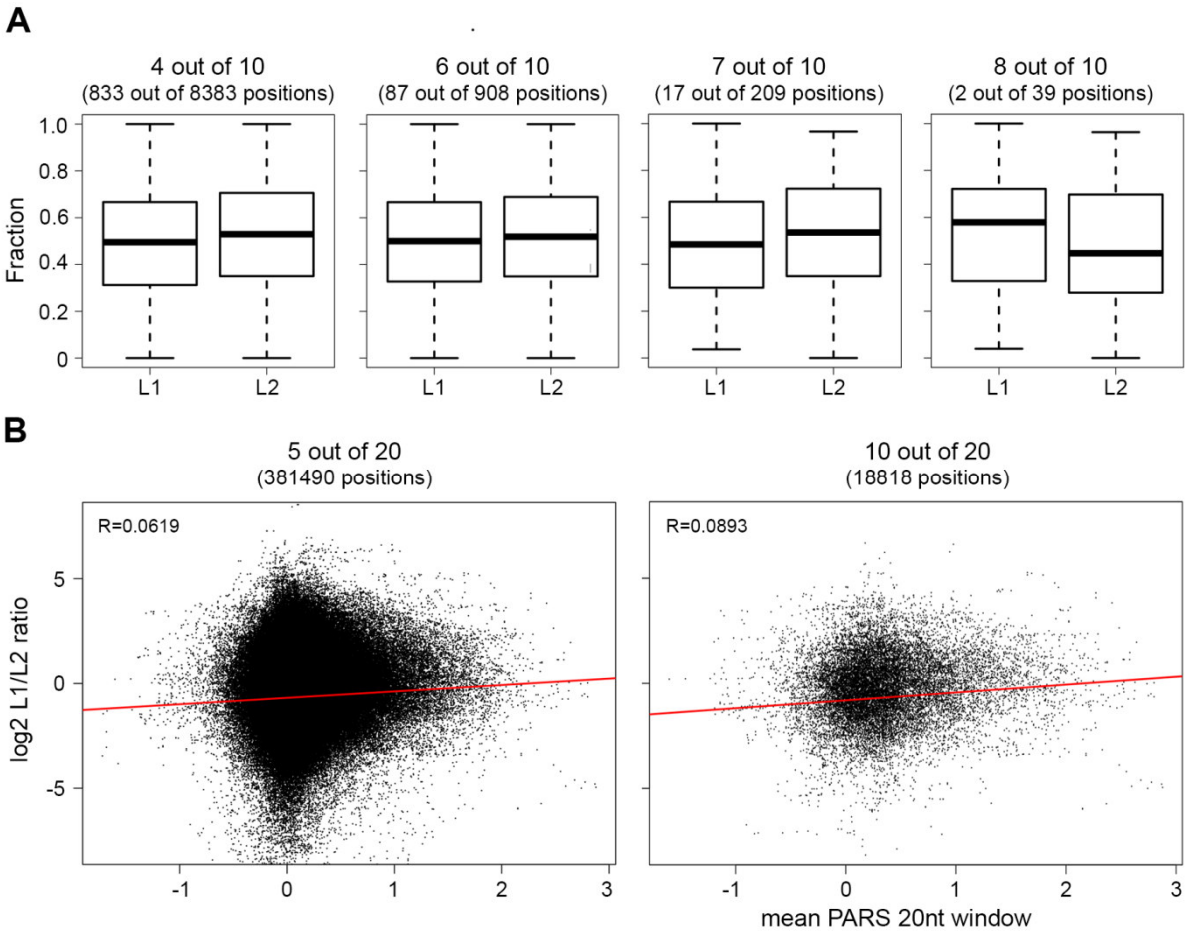


Figure 7.4 | Defining structured mRNA regions within the CDSs. (**A**) The search is performed by varying the number of structured nucleotides (i.e. with positive PARS score) within a window of 10 nt. The numbers in brackets denote the number of 80thpercentile positions within the whole set of detected structured positions. (**B**) The search is performed using the mean PARS score within a variable window (10 or 20 nt) under the restriction that within a window at least 5 nt (5 out of 10 nt or 5 out 20 nt) or 10 nt (10 out of 20 nt) have a PARS score different than zero. Note that this approach also cannot select for a minimal threshold PARS score over which the L_1/L_2 ratio becomes significant. PARS score gives the propensity of each nucleotide to partition between single or double stranded structure, therefore this propensity differs from the gain of energy which is determined by the type of nucleotide, the context and other factors.

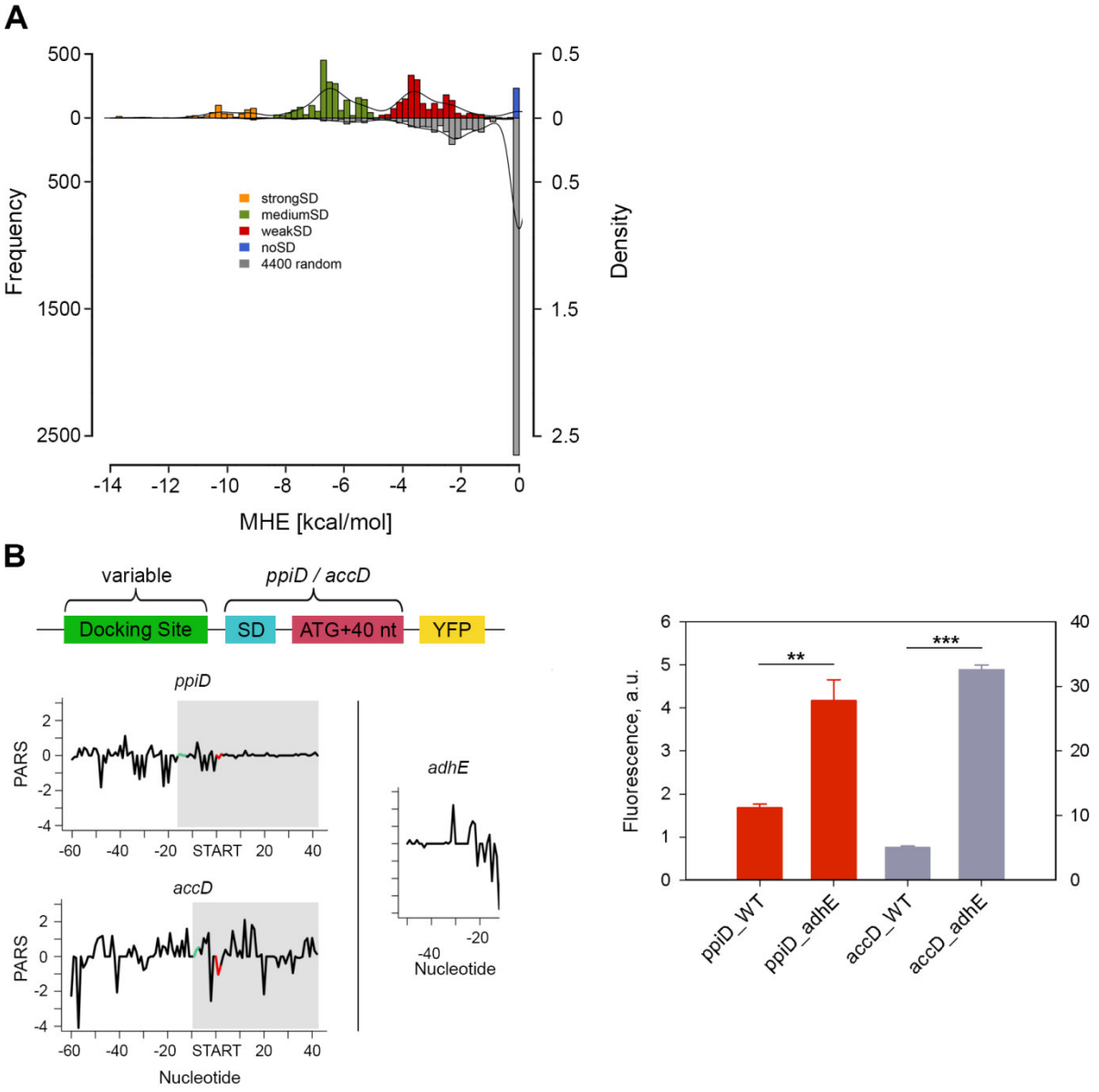


Figure 7.5 | The structure propensity of the sequence upstream of SD correlates with expression. (**A**) Randomization of SD sequences. The MHE of paring randomized sequences (gray) with the anti-SD of the 16s rRNA is compared to the MHE distributions of naturally occurring SDs (Fig. 2.9 A). The fully randomized sample of all possible variations of randomized sequences of 8-nt length was ~65,000, however only 4,400 randomly chosen sequences (gray) are plotted to match the number of *E. coli* ORFs. The smoothed lines represent kernel density estimation (right y-axis). Color coding of the naturally occurring *E. coli* SD sequences is in Fig. 2.9 A. (**B**) FACS expression analysis of *ppiD* and *accD* whose original sequence upstream of the SD (schematic) was replaced by that of *adhE* which has clearly different PARS score (*adhE*= -0.564, *ppiD*= -0.495, accD= 0.5284043). Data are means (n = 3) ± standard error of the mean (s.e.m.).**, P <0.01; ***, P <0.001.

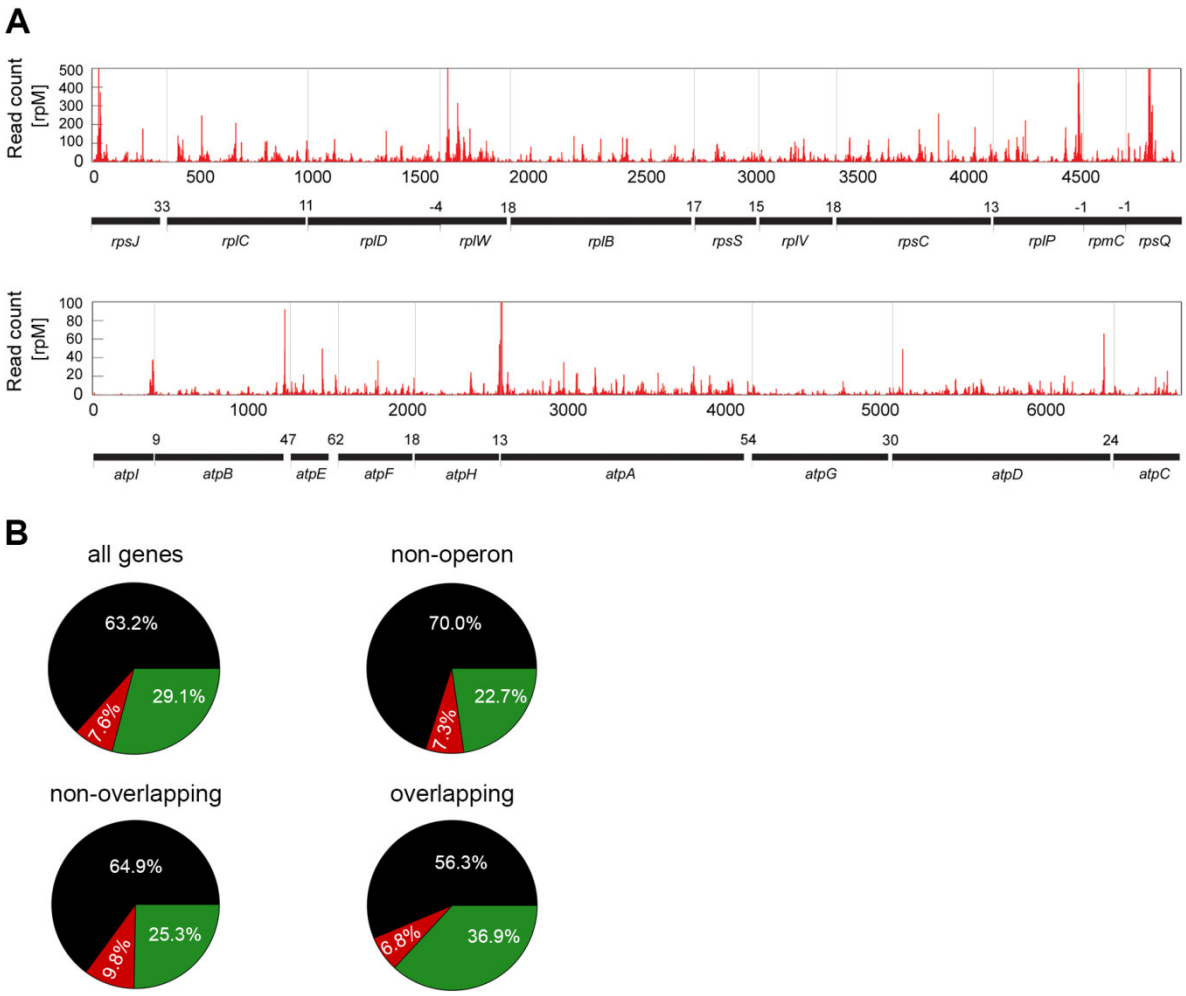


Figure 7.6 | Stop codon distributions and secondary structure of different gene groups. (**A**) Examples of genes organized in operons containing overlapping genes (upper panel) or non-overlapping (bottom panel) genes. RPF counts are plotted against the nucleotide position of operons. The gray vertical lines denote the boundaries of each ORF; the distance between the ORFs is given in nt in the schematic below the RPF-coverage profile. Negative numbers denote overlapping ORFs. (**B**) Frequency of the three stop codons in different gene groups. UAA (black), UGA (green) and UAG (red).

Table 7.1 | List of genes with identified ribosomal stalling induced by mRNA secondary structure. Green, membrane proteins; blue, ribosomal proteins; orange, cytosolic proteins.

Gene Name	Function	Detected pausing site
ECK120000097.atpD	ATP synthase	981
ECK120000340.ftsY	SRP receptor	1159
ECK120000662.ompA	outer membrane protein A	817
ECK120000663.ompC	outer membrane protein C	67
ECK120000664.ompF	outer membrane protein C	70
ECK120000763.proW	proline ABC transporter - membrane subunit	56
ECK120001178.yidC	inner-membrane protein insertion factor	341
ECK120001455.yhbE	conserved inner membrane protein	325
ECK120001816.mdoH	membrane glycosyltransferase	995
ECK120001991.nuoE	NADH:ubiquinone oxidoreductase	319
ECK120001992.nuoG	NADH:ubiquinone oxidoreductase	1514
ECK120002029.treB	trehalose PTS permease	1245
ECK120002695.nanT	NanT sialic acid MFS transporter	460, 1234
ECK120002864.mscK	potassium dependent mechanosensitive channel MscK	2293
ECK120003892.bamB	Outer Membrane Protein Assembly Complex - BamB subunit	231
ECK120000394.glpT	glycerol-3-phosphate:phosphate antiporter	1048
ECK120000772.pssA	Phosphatidylserine synthase (component)	233
ECK120001020.trxA	thioredoxin 1	79
ECK120004324.dcuS	DcuS sensory histidine kinase	1244
ECK120000353.fusA	elongation factor G	1298, 1469
ECK120000855.rplA	50S ribosomal subunit protein L1	370
ECK120000856.rplB	50S ribosomal subunit protein L2	683
ECK120000857.rplC	50S ribosomal subunit protein L3	228
ECK120000859.rplE	50S ribosomal subunit protein L5	443
ECK120000863.rplK	50S ribosomal subunit protein L11	48
ECK120000877.rpmB	50S ribosomal subunit protein L28	38
ECK120000892.rpsB	30S ribosomal subunit protein S2	253
ECK120000893.rpsC	30S ribosomal subunit protein S3	432
ECK120000898.rpsH	30S ribosomal subunit protein S8	55
ECK120000901.rpsK	30S ribosomal subunit protein S11	365
ECK120000902.rpsL	30S ribosomal subunit protein S12	236
ECK120001161.rimP	ribosome maturation protein	306
ECK120001768.typA	GTPase ribosome-associated	524
ECK120000131.carB	carbamoyl phosphate synthetase	125, 258
ECK120000209.deaD	DEAD-box RNA helicase	1148
ECK120000216.deoD	purine nucleoside phosphorylase	38
ECK120000351.fumC	fumarase C monomer	197
ECK120000387.glpD	glycerol-3-phosphate dehydrogenase subunit	449
ECK120000551.malK	maltose ABC transporter - ATP binding subunit	1108
ECK120000553.malP	maltodextrin phosphorylase monomer	304
ECK120000555.malT	MalT transcriptional activator	644
ECK120000630.nanA	N-acetylneuraminate lyase component	587
ECK120000695.pgk	phosphoglycerate kinase	986
ECK120000781.purA	Adenylosuccinate synthase (component)	220, 1056

ECK120000808.rbsD	ribose pyranase	50
ECK120000885.rpoB	RNA polymerase	671, 1085, 1596, 2916
ECK120000886.rpoC	RNA polymerase	723
ECK120000887.rpoD	RNA polymerase	1565
ECK120000970.sucC	succinyl-CoA synthetase	359
ECK120000994.tnaA	Tryptophanase	908
ECK120001032.tyrS	tyrosyl-tRNA synthetase	224
ECK120001056.valS	valyl-tRNA synthetase	1832
ECK120001401.plsX	fatty acid/phospholipid synthesis protein	797
ECK120001567.mfd	transcription-repair coupling factor	2915
ECK120001623.lptB	Lipopolysaccharide export ABC transporter ATP-binding protein	49
ECK120002177.gpmM	Phosphoglycerate mutase	356
ECK120002193.acnB	Aconitase B	35
ECK120002257.msrB	methionine sulfoxide reductase B	304
ECK120002445.pta	phosphate acetyltransferase	1300
ECK120002944.rnk	regulator of nucleoside diphosphate kinase	312
ECK120003088.rlmL	23S rRNA m^2G2445 methyltransferase	31
ECK120003163.nagZ	beta-N-Acetylglucosaminidase	954
ECK120000249.eco	ecotin monomer	444
ECK120000948.speA	arginine decarboxylase	1166

Table 7.2 | List of the 64 RNase E cleavage positions. Positions denotes the gene coordinates in the *E. coli* chromosome on either the forward (frw) or reverse (rvs) strand.

Gene	Position	Strand	Gene	Position	Strand
ftsI	93136	fwd	*mglB*	2237371	rvs
ftsI	93144	fwd	*mglB*	2238477	rvs
coaE	113122	rvs	*nuoN*	2388534	rvs
tsf	190802	fwd	*nuoL*	2392843	rvs
cyoC	447668	rvs	*nuoL*	2393046	rvs
cyoB	449135	rvs	*nuoG*	2397080	rvs
cyoB	449152	rvs	*nuoG*	2397802	rvs
yajG	452839	rvs	*nuoG*	2397961	rvs
uspG	640655	rvs	*nuoC*	2401345	rvs
fur	709635	rvs	*nuoB*	2402226	rvs
sdhC	754611	fwd	*nuoB*	2402633	rvs
sdhD	754977	fwd	*nuoA*	2403043	rvs
sdhD	755103	fwd	*maeB*	2576334	rvs
ybgF	778815	fwd	*rnc*	2701527	rvs
glnH	846518	rvs	*pssA*	2721316	fwd
ihfB	963323	fwd	*rimM*	2743400	rvs
ompF	985413	rvs	*nlpD*	2865691	rvs
ompF	985851	rvs	*nlpD*	2865699	rvs
ompF	985860	rvs	*hybO*	3144295	rvs
ompA	1018886	rvs	*ispB*	3331735	fwd
rne	1143045	rvs	*smg*	3430051	rvs
minD	1224121	rvs	*rpsM*	3440310	rvs
minD	1224577	rvs	*envZ*	3533575	rvs
oppB	1300973	fwd	*malT*	3551863	fwd
uspF	1433290	rvs	*tnaA*	3887308	fwd
aldA	1487678	fwd	*tnaA*	3887483	fwd
aldA	1487690	fwd	*tnaA*	3888162	fwd
manX	1900641	fwd	*rbsC*	3934182	fwd
cspC	1905439	rvs	*fdhE*	4079159	rvs
cspC	1905470	rvs	*metL*	4130357	fwd
rfbD	2109053	rvs	*frdA*	4379660	rvs
gatZ	2173243	rvs	*ytfK*	4437515	fwd

8. Acknowledgments

The present thesis is the result of three years of exciting work. In this time, I had the opportunity to learn a lot about science and doing science, to discuss my ideas with professors and scientists from all over the world, to present my research in conferences and publish my work in scientific journals. But I also had the opportunity to grow up as a man, to experience how to “restart” a life in a new country, to learn about many cultures of the world and appreciate the beauty and the wealth of cultural differences.
For all of this, and much more, I have to thank Zoya Ignatova. For choosing me for this exciting project and giving me the opportunity to experience the joy and frustration of a PhD project. For the supervision, the emotional discussions and, in conclusion, for passing down the love for science to me.

I am thankful to the European Initial Training Network NICHE for all the great meeting we had, spanning from scientific discussions to group activities. But particularly for sharing, since the beginning, the experience of a PhD: it was of great help to share the difficulties and the successes of doing science with you guys and I feel lucky to see that science, through collaborations, can give birth to friendship. I am confident that this will last long and will generate new collaborations in turn.
I am also grateful to NICHE for the founding and the European feeling that we as a young generation could experience, particularly in these hard days for Europe.

Thank to Justin Clarke and Kenneth McDowall (Univ. Leeds, UK) for providing the full data set of the RNase E and non-RNase E cleavage sites and Kajetan Bentele for the help with modeling the sampling error between biological replicates. Furthermore, I feel grateful to Prof. Ben Luisi (Univ. of Cambridge, UK) for having carefully listened to my poster presentation and for suggesting a McDowall publication, which gave new hints for the developing of this work.
I am obliged to Claudia Langnick and Mirjam Feldkamp (Max Delbrück Center, Berlin) for assistance with deep-sequencing.

I owe my deepest gratitude to my family for supporting me in this choice, with your endless love. Despite of the distance, I always felt you with me.

A huge “thank you” goes to my Italian friends, who always managed to be close and who often came to visit and encourage me, particularly in the hardest of times. So, thank you Sara & Michele, Ilaria e Ingrid, Luca, Simone, Giulio, Lucia and all the others. You are all important.

This thesis and this PhD would have been literally impossible without the infinite help and care of the whole Biochemistry group. You guys became my family in these years and I feel deeply lucky to have shared this adventure with you. Your encouragement and comfort, advices and critics, laughs and tears were always precious, inside and outside the lab. I will feel always bound to all of you.
I would have a special thank for each of you but the list would take a whole new thesis. However, I am deeply grateful to Alex Bartholomäus, for the close and fruitful collaboration and for sharing the bad and good time of this project. Looking forward to new adventures! Also, I want to sincerely thank Iole Ferro, for experiencing together the life of an immigrant and for the long scientific discussions, during dinner time.

This work is dedicated to Theresa, who brought the sun in the often-gray sky of Germany and who was patient enough to deal with my longer days in the lab and with my working in Hamburg. Indeed, we make the best of our lives in a perfect chemistry set.

9. Declaration on oath

Hiermit erkläre ich, Cristian Del Campo, an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keinen anderen als die angegeben Quellen und Hilfsmittel benutzt habe.

Hamburg, den 22 Februar 2016

I, Cristian Del Campo, hereby declare on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids. I hereby declare that I have not previously applied or pursued for a doctorate (Ph.D. studies).

Hamburg, 22 February 2016

Cristian Del Campo