Analysing deep-seq data to reveal the hidden translational features in normal and pathological conditions

Irina Chelysheva

Dissertation
with the aim of achieving a doctoral degree
at the Faculty of Mathematics, Informatics and
Natural Sciences
Department of Chemistry of Universität Hamburg

April 2020

Gutachter:

Prof. Dr. Andrew Torda

Prof. Dr. Hans-Jürgen Kreienkamp

Tag der Disputation: 12 Juni 2020

Die vorgelegte Arbeit wurde von Juni 2016 bis October 2019 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 (faktisch: bis Mai 2019) und Herr Prof. Dr. Andrew Torda angefertigt.

Eidesstattliche Versicherung

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

Irina Chelysheva

25 April 2020

Declaration on oath

I, Irina Chelysheva, 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).

Irina Chelysheva

25 April 2020

Dedication

I dedicate this thesis to my grandfather, who passed away when I was too young, but who believed in me more than anyone else.

Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less.

(Marie Curie)

Oral and poster presentations, covering parts of this work:

Role of stress granules in translation: the bioinformatics view and analysis. Poster presentation. Dynamics of translation; Erice, Italy (2017)

Stress granules in translation: the bioinformatics view and analysis. Oral presentation. RNAtion. Computational and Experimental RNA Biology Conference for Young Scientists; Poznan, Poland (2017)

RNA chaperone Hfq is crucial for ribosome biogenesis. Oral presentation. Final progress meeting, FOR 1805, DFG; Berlin, Germany (2018)

RNA-chaperone Hfq is crucial for ribosome biogenesis – the bioinformatic view. Poster presentation. Translating Translation; Blankenese, Germany (2018)

Bioinformatic view on deep-sequencing data analysis. Oral presentation. Young Scientist Symposium; Bordeaux, France (2018)

Publications prepared during graduation and partly covered in this work:

Andrade, J. M., dos Santos, R. F., Chelysheva, I., Ignatova, Z., and Arraiano, C. M. (2018). The RNA-binding protein Hfq is important for ribosome biogenesis and affects translation fidelity. *The EMBO Journal*, e97631.

Anders, M.*, Chelysheva, I.*, Goebel, I., Trenkner, T., Zhou, J., Mao, Y., Verzini, S., Qian, S-B., Ignatova, Z. (2018). Dynamic m^6A methylation facilitates mRNA triaging to stress granules. *Life Science Alliance*, 1(4), e201800113.

*These authors contributed equally to this work.

Gorochowski, T. E., Chelysheva, I., Eriksen, M., Nair, P., Pedersen, S., and Ignatova, Z. (2019). Absolute quantification of translational regulation and burden using combined sequencing approaches. *Molecular Systems Biology*, 15(5), e8719.

List of the hazardous substances

No H&P-substances were used in this dissertation.

Table of Contents

Zusammenfassung

In dieser Arbeit zeige ich mithilfe dreier erfolgreicher deep-sequencing Studien die Wichtigkeit von gründlichen bioinformatischen Analysen als letzten und damit auch entscheidenden Schritt in die Forschungskette auf, was uns erlaubt, verschiedene experimentelle Setups in einem Schritt durchzuführen. Um die wertvollsten Informationen zu extrahieren, welche oftmals versteckt sind und dadurch von üblichen Analysepipelines nicht ermittelt werden können, muss man spezifische Algorithmen entwickeln, die exakt auf eine wissenschaftliche Frage oder Datentyp zugeschnitten sind. Rasch entwickelte "Deep sequencing" Technologien erhöhen sowohl die Präzision als auch die Tiefe der Datenbestände, die weltweit produziert werden. Die enorme Quantität an biologischen Daten in der Forschung resultiert in der Notwendigkeit für kontinuierlich neue bioinformatische Werkzeuge und Algorithmen, um diese Daten zu prozessieren.

Abstract

Rapidly developing deep sequencing technologies constantly increase both, the precision and the depth, of the datasets produced worldwide. The enormous amount of big biological data in research causes the need for continuous integration of novel bioinformatics tools and algorithms to process it. In order to extract the most valuable information, which is frequently hidden, and therefore escapes from the common pipelines of analysis, the specific algorithms fitting a particular scientific question and data-type, has to be integrated. In this thesis, using three successful deep sequencing-based studies, I am showing the importance of in-depth bioinformatics analysis as the last and frequently the crucial step in the research pipeline, which allows to combine various experimental setups into one flow.

List of Figures

List of abbreviations

ORF - Open Reading Frame

sORF - small Open Reading Frame

UTR - Untranslated region

CDS - Coding DNA Sequence

RBS - Ribosome Binding Site

r-proteins - ribosomal proteins

miRNA - microRNA

mRNA - messenger RNA

rRNA - ribosomal RNA

RPF - Ribosome Protected Fragment

nt - nucleotide

RPKM - Reads Per Kilobase of transcript, per Million mapped reads

RPM - Reads Per Million mapped reads

RD - Ribosomal Density (also known as TE - Translation Efficiency)

NGS - Next Generation Sequencing

Ribo-Seq - Ribosome Profiling

RNA-Seq - RNA sequencing

PAR-CLIP (PAR-CLIP-seq) - photo-activatable ribonucleoside cross-linking and immunoprecipitation (sequencing)

PK - Pseudoknot

SG(s) - Stress Granule(s)

m^6A-seq - m^6A Sequencing (also known as MeRIP-Seq - Methylated RNA Immunoprecipitation Sequencing)

AS - arsenite

QC - Quality Control

Chapter 1

Introduction

1.1 Background and aim of the thesis

The central dogma of molecular biology describes the flow of genetic information from DNA into protein as a two-step process of transcription and translation (Crick, 1958) having RNA as an intermediate between protein coding gene and its protein product.

Nowadays, emergent technologies allow to access a regulation of the protein biosynthesis at each of its steps. The wast majority of the technics, which are widely used for this purposes, based on the sequencing technologies. Though the variability of available deep sequencing technologies increased dramatically over the last decades (Koboldt et al, 2013), each of them is applicable to the particular scientific problem.

As DNA itself is a first molecule involved in the protein biosynthesis, the earliest sequencing attempts and approaches corresponded to DNA sequencing, which has been defined as a process of determining the nucleic acid sequence, e.g. the order of nucleotides in DNA. The technology started from Sanger in 1955, who completed a sequence of one protein - insulin (Ryle et al, 1955), and developed throughout the years into the powerful whole genome sequencing technologies, which are currently used worldwide.

Serving as a messenger between DNA and the ribosomes where the protein synthesis occurs, this type of RNA is called a messenger RNA (mRNA). The number of

copies of mRNA corresponding to each particular gene (also called "transcripts" as a product of transcription) plays a key role in the protein biosynthesis. The total pull of mRNA transcripts expressed in the cell compile a transcriptome, which can be explored with through RNA sequencing (RNA-Seq).

Besides mRNA the other types of RNA are widely present in the cell. These RNA transcripts do not encode the proteins and therefore called non-coding RNAs (ncRNAs). The most abundant and functionally important types include ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), long non-coding RNAs and short RNAs, such as microRNAs, siRNAs, piRNAs, snRNAs, snoRNAs. First two types are crucial for the cell, as they are directly involved in the protein synthesis, while some of the others are functionally important as well, since they are involved in the regulation of RNA stability, protein translation and other essential functions. Though, many of the ncRNA species and their functions remain unexplored, variety of the deep sequencing approaches has been established to study the particular type of ncRNAs (Motameny et al, 2010).

Ribosome profiling (Ribo-Seq) is another sequencing technology widely used nowadays. It is accessing the translation of RNA into proteins, as second step of the protein biosynthesis. This method has been developed by Nicholas Ingolia and Jonathan Weissman (Ingolia et al, 2009) to allow detection of the actively translated mRNAs by reporting on the positions of all the active ribosomes in the cell at any given particular moment.

Besides the primary purpose of giving a "global snapshot" on translating ribosomes, the technology has enormous potential to reveal the hidden translational features, such as precise localization of the Translation Start and Stop Sites (Lee et al, 2012), discovering of the novel Open Reading Frames (ORFs) (Mackowiak et al, 2015), measuring Translation efficiency (McGlincy and Ingolia, 2017), Translation initiation and termination rates (Baggett, 2017), the speed of translating ribosomes (Del Campo et al, 2015), evaluating specific responses to the changing growth conditions, comparing the expression levels, revealing the Gene Onthology terms (GO-terms) or pass ways

involved and affected by the treatment and much more. All of these potential applications enable researchers to deeply understand the mechanisms of translation itself and the changes occurring in response to different factors, such as various types of stress, antibiotics treatment or presence of mutations in particular genes. However, it requires constant developing of the novel pipelines and algorithms for the processing and analysis of the data.

Many other deep sequencing technologies exist and continuously develop, such as CLIP-Seq (High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) that identifies protein–RNA binding sites or RNA modification sites, or MeRIP-seq - method for detection of post-transcriptional RNA modifications, or Oxford Nanopore Sequencing Technology, which allows to sequence DNA or RNA directly without additional equipment.

The advantages of each given sequencing technic can be multiplied by using a combination of different sequencing approaches in order to provide a generic view on the processes occurring in the cell. However, this is not a straightforward task, as it requires a deep understanding of both fields - molecular biology and statistics interconnected through the programming environment. This led to the development of the bioinformatics as a field, where researchers analyze the big data, which is coming from the deep sequencing. The need of the specialists in bioinformatics is increasing every year along with an amount of data produced.

One of the major goals for the bioinformatician as the last researcher in the deep sequencing pipeline is to interconnect to variety of outputs from the different sequencing technics in order to extract the most valuable and accurate information that meets the objectives of the study. Therefore, the bioinformatics pipeline cannot be standardized even on the level of preprocessing the raw sequencing data and moreover not in the downstream analysis.

This issue leads to the primary aim of this dissertation - to develop the specific bioinformatics algorithms and pipelines for processing and analysis allowing to interconnect various deep sequencing datasets and to reveal the hidden features for

each of the studies as well as to show the importance of this step for the research outcome.

1.2 Structure of the thesis

The current dissertation is organized in six chapters and is based on the three different studies, all of which include NGS datasets. While the research questions differ, the studies are interconnected by the data analysis part, which aims to address the potentially hidden applications of various NGS technologies, particularly, RNA-seq and Ribo-seq.

Chapter 2, directly following the Introduction Chapter, describes the common pipeline of the NGS data analysis, provides an overview of the available bioinformatics tools and methods, which can be used at any of the steps of the data processing, as well as discusses the selection of each particular program for the current studies included in this thesis.

Chapter 3 corresponds to the study of post-transcriptional RNA modification — m^6A. The work has been performed in the collaboration with Maximilian Anders who led the experimental part of the study and successfully defended his Doctoral thesis on this topic while I was responsible for the analysis of the multiple sequencing datasets and their combination. The results have been published with a shared first authorship.

> Anders, M.*, Chelysheva, I.*, Goebel, I., Trenkner, T., Zhou, J., Mao, Y., Verzini, S., Qian, S-B., & Ignatova, Z. (2018). Dynamic m^6A methylation facilitates mRNA triaging to stress granules. *Life Science Alliance*, 1(4), e201800113. DOI: 10.26508/lsa.201800113

My contribution included data curation, formal analysis, investigation, and writing - original draft.

Chapter 4 reports on the collaborative research project with a group from Universidade Nova de Lisboa, Oeiras, Portugal. This study was related to the bacterial RNA chaperone Hfq, where the RNA-seq and Ribo-seq provided a new insight on the function of the protein. The results have been published with my contribution comprised of the data analysis, writing the computational part of the "Materials and methods" section and overall editing of the manuscript.

> Andrade, J. M., dos Santos, R. F., Chelysheva, I., Ignatova, Z., & Arraiano, C. M. (2018). The RNA-binding protein Hfq is important for ribosome biogenesis and affects translation fidelity. *The EMBO Journal*, e97631. DOI: 10.15252/embj.201797631

Chapter 5 is related to another collaborative study, which mainly involved the collaboration with Dr. Thomas Gorochowski, University of Bristol, Bristol, UK. In this study, we were using Ribo-seq and RNA-seq data to quantify the translational processes and translational burden induced by stable RNA pseudoknot construct. This work has been recently published.

> Gorochowski, T. E., Chelysheva, I., Eriksen, M., Nair, P., Pedersen, S., & Ignatova, Z. (2019). Absolute quantification of translational regulation and burden using combined sequencing approaches. *Molecular Systems Biology*, 15(5), e8719. DOI: 10.15252/msb.20188719

I processed the sequencing datasets, contributed to the data analysis, writing, editing of the manuscript, producing the figures, tables and organization of the data.

Each of the Chapters from 3 to 5 has its own substructure, which includes the background, materials and methods, results and conclusion sections. After exploring the different aspects of translation, gene expression, approaches, prospects and limitations of data analysis throughout the thesis, the overall conclusion is shaped in Chapter 6.

Chapter 2

Bioinformatics tools and methods

2.1 Overview of the sequencing approaches and library preparation

The deep sequencing approaches, such as RNA-seq, Ribo-seq and others, used in the current thesis, are based on the Illumina sequencing technology workflows. The studies followed the standardised protocols, which are typically used in the field, unless mentioned specifically in the Materials and Methods section of each chapter from 3 to 5.

The laboratory part includes the preparation of the library, which will be further sent for sequencing. For RNA-seq, the total RNA is serves as an input, in Ribo-seq - ribosome-bound RNA first undergoes digestion. When the RNA is extracted, it is important to enrich the poll of RNA of interest out of all the RNA species present in the library, which is especially crucial for RNA-seq, where the total RNA is used. Considering that the rRNA is the predominant form RNA in the cell (up to 90%) in order to enrich the mRNAs instead, two main strategies are used: the rRNA depletion and polyA selection. The extracted RNA is fragmented and reverse-transcribed to cDNA, then the specific Illumina barcodes and adapters are ligated. The last step of the library preparation is an amplification of the cDNA by PCR. The resulting library is purified and sent for the sequencing.

Deep sequencing of the cDNA, in the case of RNA-seq, provides the sequences of all the RNAs in the cell, e.g. transcriptome; in Ribo-seq - only those RNAs, which are

bound by ribosomes during translation, e.g. translatome; if PAR-CLIP-seq is used, only those RNA species, which interact with a particular RNA-binding protein, are present in the library (Spitzer et al, 2014).

In the studies described in the current thesis, the libraries were sequenced on a HiSeq2000 Illumina machine. The typical output of the sequencing is a raw data-file containing millions of sequences, the amount reads is variable across the samples and called sequencing depth, which depends on the multiple factors, including the quality of input material, the concentration of RNA, the exact model of the sequencing system, used in the study.

Generally, the length of the sequencing fragments obtained from the sequencing may vary between 20 to 200 nucleotides depending on the protocol. In the current studies, unless mentioned, all the sequencing reads obtained during the sequencing have a read length corresponding to the size of the ribosome (~20-35 nucleotides). The libraries were generated following the smallRNA-seq protocol, which has been selected considering that the Ribo-seq served as a central method for the studies. Choosing the same range of read lengths for all the other approaches made them comparable and allowed to perform additional meta-analysis including various types of datasets (for example, to evaluate the translation efficiency via comparing the reads from the Ribo-seq to those from the corresponding RNA-seq sample).

The DNA in the genome has two strands, both of which encode different proteins. For each gene, one DNA strand always serves as coding strand and contains the genomic sequence of this gene, while another strand is "antisense" and is complementary to the "sense" strand. Knowing from which strand each sequencing fragment in the library has been delivered is important for the correct identification of its location in the genome (when aligning the reads as a part of the downstream analysis). Depending on the protocol, the sequencing can be either non-strand specific or strand specific. In our studies, the strand-specific (or stranded) protocol was clearly a method of choice, since it allowed to increase the precision of the alignment, which is crucially important for the library containing the fragments of the short length.

2.2 Pipeline - processing of the sequencing data

The raw samples obtained directly from the sequencing are stored in the cloud and received via the link from the sequencing facility. This is a starting point for the bioinformatic pipeline of processing and analysing the datasets. The current chapter describes the major steps of processing the data prior to any sample- or study-specific downstream analysis is performed.

2.2.1 Preprocessing

First, the raw sequencing samples are received in the FASTQ format, which is a text-based format, storing the sequences and their corresponding quality (Phred) scores for each base. It is a standard format for NGS data from Illumina sequencing, which serves as input for the further analysis with a variety of available tools (Cock et al, 2010).

The reasonable initial step of the pipeline is a preprocessing the FASTQ files before mapping the sequences to the genome, since it allows to obtain better mapping results - higher percentage of the aligned sequences and higher accuracy of the mapping if the low-quality sequences are removed or trimmed.

Therefore, initial quality control (QC) is performed on the preprocessing step and is based on the quality scores obtained from the FASTQ files, which are represented as ASCII characters. In this thesis, the FASTQ Quality Trimmer as a part of Fastx-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) is used for this purpose. While various tools and scripts are available to perform the QC, the selected method has been developed explicitly to trim the short reads and is the most suitable for our type of libraries. Besides the quality scores of each base within the sequence, which are considered and the threshold for accepted quality can be specified.

Along with a QC, the sequencing adapters, which tail the reads from both ends (3' and 5') in order to fill the read-length required to run the sequencing at the machine

(in our case - 50 bases), have to be trimmed from the reads prior to the mapping to the reference genome. While multiple existing programs contain this function, one of the most common tools - Cutadapt (Martin, 2011), which is developed exclusively for trimming the adapters, has been used in this thesis. Cutadapt allows to specify and cut the adapters from both sides of the sequences (3' and 5') at the same moment.

The minimal read length can be added into the algorithm in order to exclude the reads, which appeared to be too short after the adapter trimming. The short read length does not allow to precisely align these reads to the reference genome and a single mismatch or sequencing error can lead to the miss-alignment. Therefore, unnaturally short reads should be excluded from the downstream analysis (in our case the threshold is set to < 20 nucleotides).

Home-made shell script (.sh), combining the quality filtering with Fastx-toolkit and adapter cutting with Cutadapt into one preprocessing step, is applied to each sequencing FASTQ file.

The script generates another FASTQ file containing only the reads, which passed the quality control and in parallel were cleaned up from the adapter content at the ends of the sequence.

At the next step, the preprocessed FASTQ files are checked with the quality control tool - FASTQC. It generates a detailed report on each sequencing file, which is saved as several independent text files along with an intuitive representation in html format. FASTQC html report includes a general information regarding the sequencing quality per base along the read length, sequencing depth, read length distribution, percentage of remaining adapters in the sample and possible contamination through the statistics of overrepresented and duplicated sequences. In the case of several FASTQ files being processed in parallel, multiple reports can be afterwards combined into one .html file using another quality control tool - multiQC (Ewels et al, 2016).

If some of the issues with quality of the preprocessed file are revealed on this QC step, the detailed investigation and repetition of the previous preprocessing steps are

required. In the case if the problem of low-depth or low-quality library remains, the sample should not be considered for the further processing and has to be excluded from the study.

2.2.2 Read mapping or alignment

The high-quality preprocessed FASTQ samples serve as an input for the next processing step - read mapping (also called alignment). The goal of alignment is to map short sequencing reads contained in the FASTQ sample to a large reference genome via identification of the correct genomic location for each of the reads.

Multiple bioinformatics tools are available for the alignment of the short reads obtained from RNA-seq and Ribo-seq. Each of the well known mapping tools, such as Bowtie, Bowtie2, BWA, SOAP2, has its own strengths and weaknesses revealed by benchmarking (Hatem et al, 2011). However, the first one - Bowtie - being released in 2009 (Langmead, 2010), has the best throughput and still remains the most efficient for the particularly short reads (25-50 nt) and, therefore, is widely used in respective studies and serves a basis for the multiple other aligning tools, namely TopHat, Cufflinks and others.

In the current thesis Bowtie has been generally a method of choice for the alignment of all the sequencing datasets (unless explicitly specified).

The reference genomes (in this work - Homo Sapience and E. coli) are downloaded from the open source databases in a FASTA format, which is a text-based format for representing the nucleotide sequences with sequence names included (for example, the number of the chromosome) (http://zhanglab.ccmb.med.umich.edu/FASTA/). Then the genomes are indexed with Bowtie, which builds the Burrows-Wheeler index to keep its memory footprint small. Along with the genomes, the corresponding rRNA sequences are downloaded and indexed.

The alignment typically includes 2 steps:

1. mapping of the FASTQ sequencing file to the rRNA sequences in order to exclude the left-over rRNA in the sample;

2. mapping of the unmapped reads to the reference genome.

Both of the steps can be included into one mapping command, while the alignment options can be specified for each of the mapping steps separately. The typical setup includes the amount of mismatches allowed per read (-v parameter) and the maximum number of the locations in the reference genome where the particular read can be reliably mapped (-m parameter). Through the unique mapping allowing only those reads, which have one best matching location in the genome, to be reported (-m 1) is generelly preferable, under some specific conditions the parameter has to be adjusted and changed accordingly.

The main output of the 2-step mapping is a standard SAM (Sequence Alignment/Map format) file (Li et al, 2009). This TAB-delimited text format consists of a header section (optional) and an alignment section, reporting on all the reads and their genomic alignments. SAM includes the detailed information on each particular read, such as a presence of alignment, the direction of alignment, the mapping positions within the reference genome, quality of mapping and presence of mismatches.

Additionally, if specified and requested by user, the basic text files reporting on the mapping statistics, can be produced by Bowtie. The mapping statistics provides an information on the total number and percentage of the reads, which were successfully aligned to the genome, those reads which are aligned to rRNA and excluded as well as those reads which remained unmapped or were aligned but suppressed based on the mapping parameters. This overview allows to estimate the quality of the input RNA library - whether rRNA has been globally depleted or majorly remained in the sample, while a high percentage of the unmapped reads may indicate on the contamination in the sample if the reference genome has been chosen correctly.

Important to note that the initial SAM output contains both types of reads - those, which were aligned to the reference genome, along with those which remained unmapped. Special flags in the alignment section of SAM file reports whether the particular read has been aligned.

The next processing step aims to clean-up the alignment file to prepare it for the downstream analysis, where the number and positions of the mapped reads can be accessed and counted. A command-line tool - SAMtools (Li et al, 2009) - is used to manipulate with SAM files.

SAMtools allows to filter only the reads, which were mapped to the reference, based on the flag, thereby reducing the size of the alignment file and shortening the running time for the further analysis. Then, the alignments within the SAM file are getting sorted by the names of the reference sequences (for example, chromosomes) and the starting positions of alignments within the reference genome. Finally, the a text-based SAM file is converted to another alignment format - BAM file, which is basically a compressed binary version of a SAM file that is used to represent aligned sequences.

The produced BAM file is serving as an input for the multiple tools for the visualisation and analysis of the data.

Another round of QC, including FASTQC and multiQC (See 2.2.1 for the details), should be introduced on this step. Accessing the parameters, such as quality, read length distribution, sequencing depth, is necessary to understand the quality of the data, which remains after the mapping and will undergo the analysis. The overall FASTQC output of the sample is expected to show the better trends for the BAM file compared to the preprocessed FASTQ file.

2.2.3 Counting and normalisation of the reads

Once the reads in the sequencing file have been successfully mapped to the reference genome and the output has been converted to the BAM format, the reads in the sample can be quantified. The first general quantification issue is to count the number of reads, which are mapped to each of the genes or features within the genome. Therefore, in addition to the alignment file, this step requires an annotation file, containing the information on all the features of interest within the reference genome including the starting and ending positions for each feature, coding strand, etc.

Multiple annotation formats exist in the field; among the most common ones - BED and GTF formats, both of them have been used in the current thesis. BED is the most simple and easy-to-follow annotation format, which represents a tab-delimited text file that defines a feature track (Kent et al, 2002). GTF format has a more complex structure allowing to define sub-features within the existing ones. This is especially relevant for the Human genome, where each gene is represented by multiple exons having the introns in the between, which are usually should not be considered for the counting of the reads.

Many tools manipulating with sequencing files have the function for this purpose. BEDtools (Quinlan, 2010) is a one of the common toolsets for the genomics analysis, which includes CoverageBed (or bedtools coverage). This command computes the coverage of the sequencing alignments in BAM file across the annotated features in the reference genome. As the tool is named after the BED format, the required annotation file is supposed to be formatted as BED file. The output of the CoverageBed depends on the selected options and parameters and may contain one count per feature or one count for each position within the feature (when -d option is added). The second option is particularly useful when the distribution of the reads along the gene is questionable. Also, the strand-specificity of the reads - relevant for all of our datasets - will be considered only when -s option is added.

Another widely used counting tool is a part of Htseq - Python-based framework for the high-throughput sequencing data (Anders, 2015). Htseq count is a command-line function allowing to count the reads, which are mapped to each gene or feature of interest. The input requires an annotation file in GTF format along with an alignment file in BAM format. The output is a simple tab-delimited file containing the features and their corresponding read-counts. Multiple additional parameters can be included in order to provide more information about the input files, especially the sequencing alignment file. Different ways of counting the reads, which got mapped to the exon-intron junctions, can be added as well.

Both of the described tools have been used in the current thesis: coverageBed was generally more suitable option for E.coli genome, because of its small size, simple

structure and high percentage of overlapping genes; Htseq count has been a method of choice for counting the reads in the Human sequencing samples, because of the complexity and vast size of the human genome.

Once the output file containing the raw read counts for each feature is created, the following step is the normalisation of the obtained counts. Since the downstream analysis involves multiple sequencing samples, the main purpose of the normalisation is to make them comparable. Normalised expression units are necessary to remove technical biases in sequenced data such as sequencing depth.

The first type of normalisation is RPM (Reads per million mapped reads), which is calculated as follows, for each gene:

$$RPM\ of\ a\ gene = \frac{(Number\ of\ reads\ mapped\ to\ a\ gene) * 10^6}{Total\ number\ of\ mapped\ reads\ in\ a\ sample} \tag{2.1}$$

However, this normalisation does not take into account the length of the transcripts, which has to be considered in order to compare the abundance of different transcripts within the sample. RPKM stands for Reads Per Kilobase of transcript per Million mapped reads and is calculated using the formula:

$$RPKM\ of\ a\ gene = \frac{(Number\ of\ reads\ mapped\ to\ a\ gene) * 10^3 * 10^6}{(Total\ number\ of\ mapped\ reads\ in\ a\ sample) * (gene\ length\ in\ bp)} \tag{2.2}$$

In the current thesis, the second way of normalisation is preferably used, unless specified.

2.2.4 Downstream analysis

The correct selection of the most suitable tools for the each processing step, discussed in this chapter, is important for obtaining the reliable information from each

particular sequencing dataset. However, the crucial and most challenging part is the downstream analysis, where the ability to reveal the hidden features and mechanisms is highly dependent on the methods and approaches, which are applied to the data. This is the demanding part of the current thesis and the major task for the bioinformatician in general. Therefore, the exact tools and steps used for the further processing of the data are discussed in details in the correspondent chapters for each of the studies (See Materials and Methods sections, Chapter 3-5).

Chapter 3

Dynamic methylation facilitates mRNA triaging to stress granules

3.1 Background

Nucleotide modifications are one of the most evolutionarily conserved properties of RNAs, which occur to a newly transcribed RNA transcript and therefore also called post-transcriptional modifications. The structural diversity of modified nucleotides allow them to play a key role in regulation of gene expression and cellular functions. RNA modifications are present in all three phylogenetic domains (Archaea, Bacteria, and Eukaryotes) and widely spread across all the RNA species, such as mRNAs, tRNAs, rRNAs and non-coding RNAs (including lncRNAs, miRNAs, snRNAs, snoRNAs). To date, 172 various modification types have been documented; they are listed in RNA Modification Database - Modomics database (http://modomics.genesilico.pl). tRNAs are the most heavily modified RNA specie with an average of 13 different modifications per molecule (Pan, 2018) meaning that every 4^{th} base is modified.

Post-transcriptional modifications occur on mRNAs prior to their translation into the protein products, they are present in varying levels in most of the protein coding genes being one of the key regulatory mechanisms of RNA functions, where each modification type plays a different role.

Among the variety of known modifications found in along the mRNA molecule (Fig.3.1, from Fig.1 in Zaccara et al, 2019). N^6-methyladenosine (m^6A) is the most

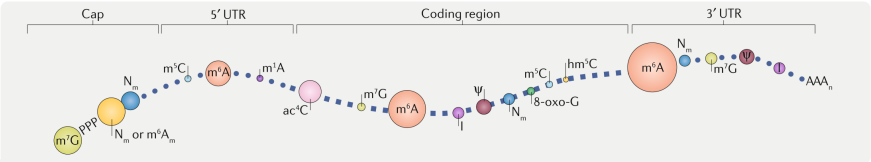

Figure 3.1: Overview of nucleotide modifications on mRNA (from Fig.1 in Zaccara et al., 2019)

abundant post-transcriptional modification in mammalian mRNA, which affects fundamental aspects of RNA metabolism. The methylation of adenosine is reversible, and its discovery revealed a new branch of post-transcriptional gene regulation.

The methylation is a complex reversible process, which involves three major groups of the proteins: methylating enzymes - "writers" - install methylation, demethylases - "erasers" - remove it, and binding proteins - "readers" - recognize the existing m^6A on the RNA and regulate the downstream molecular mechanisms (Fig.3.2, Fig.1 in Shi et al, 2019). Knowledge about the proteins involved in the m^6A life-cycle at each step is constantly expanding, the increasing resolution of the cutting-edge technologies, such as NGS and mass spectrometry, allow discovering the new effectors of m^6A.

Up to date, the most recent an complete scheme of m^6A cycle and the proteins involved in it, has been published in September this year (Shi et al, 2019). The effectors include two types of "writer" proteins: complex of METTL3 and METTL14 with additional adaptor proteins WTAP, VIRMA, ZC3H13, HAKAI (Fig.3.2, top-left - (1)); and independent "writer" protein - METTL16 (Fig.3.2, top-left - (2)). Three classes of "readers" differently recognize and bind m^6A: YTH-domain containing proteins (YTHDF1-3, YTHDC1-2) directly recognize methylation by YTH-domain (Fig.3.2, right - (1)); HNRNPC/G and HNRNPA2B1 can bind the methylated RNA in presence of the local structure (Fig.3.2, right - (2)); other common RNA-binding proteins have a potential to bind m^6A, but the exact mechanism of this binding remains unexplored (Fig.3.2, right - (3)). Two "erasers" - FTO and ALKBH5 (Fig.3.2, bottom-left - (1) and (2)) - make the methylation process reversible and demethylate the RNA.

Methylation had been shown to affect the stability of RNA (Wang et al, 2014), to

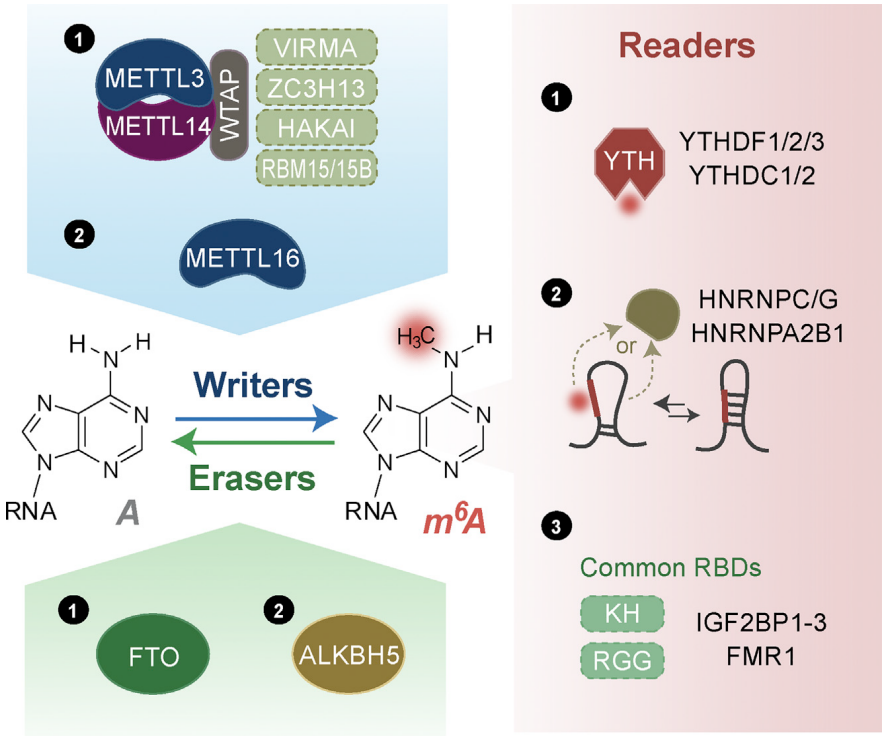

Figure 3.2: m^{6}A Effectors: Writers, Erasers, and Readers (Fig.1 in Shi et al, 2019)

play a crucial role in translation activity and efficiency (Wang et al, 2015), and to be involved in other biological processes such as microRNA biogenesis and splicing. Importantly, the exact function of each particular methylation site is dependent of its localization along the mRNA transcript. In particular, m^{6}A residues located within the 5'UTR region initiate translation of mRNAs in a cap-independent manner (Meyer et al, 2015), which is a crucial mechanism of translation under the stress conditions, such as heat shock (Meyer et al, 2015; Wang et al, 2015; Zhou et al, 2015). From the other side, the methylation sites localized in 3' UTR are involved in the stability of mRNA (Wang et al, 2014). Considering the variety of the functions of methylation, the distribution of m^{6}A-modified residues along mRNAs becomes one of the fundamental areas of investigation.

Novel sequencing approaches have been developed in order to study the m^{6}A and its functions, to access the methylation patterns and particular methylation sites. The map of the methylation sites across the transcriptome is referred as epitranscriptome. For instance, MeRIP-Seq, which maps m^{6}A-methylated RNA, allows the researchers

to successfully identify the location of the adenosines, which were methylated (Meyer et al, 2012). In this approach, m^6A-specific antibodies are used to immunoprecipitate ~100 nt-long RNA fragments, followed by NGS sequencing of the those fragments. Being a widely used method to access the methylation patterns, it is also called m^6A-seq. This sequencing technic has been used in the current study as well. This method generates m^6A peaks, but does not identify the exact position of the methylation sites. The other approach, which maps m^6A locations with a single-nucleotide resolution and precisely identifies their positions, is a miCLIP-seq, where anti-m^6A antibodies are crosslinked to mRNA sequences (Grozhik et al, 2017).

The first methylation profiles from mouse brain cells and human cells (HEK293T) were published in 2012 and assessed using MeRIP-Seq technology. Under the normal conditions, the enrichment of methylation has been detected around the STOP-codon region and the following up 3' UTR region. Even so, newly developed miCLIP-seq allowed to increase the resolution of m^6A mapping along the transcriptome, the first findings remain relevant and consistent with the most recent research.

The utilization of sequencing technologies allowed to specify the exact sequence motif, which is associated with m^6A. First, RRACH (Wang & Zhao, 2016) and later DRACH motif (Zhang et al, 2019) has been discovered. The current studies suggest that conserved DRACH motif, where D = A/G/U; R = A/G; and H = U/A/C, is strongly referred to the occurrence of methylation.

While some functions of methylation have been well annotated, the effect of the stress conditions on the methylation patterns as well as the involvement of the m^6A in the stress response remain majorly unexplored. While several studies have been conducted on the involvement of m^6A in the heat stress response (Zhou et al, 2015), the role of methylation in oxidative stress has not been yet reported.

The stress causes global reprogramming of the cell activity, shuts down the translation and leads to the formation of the stress granules (SGs). SGs are dense cytoplasmic aggregations that contain RNA-binding proteins, translation initiation factors,

large and small ribosomal subunit protein components, and mRNAs stalled in translation initiation. As the SGs are formed under the stress, they suppose to protect mRNAs from harmful conditions or serve as a decision point for untranslated mRNAs, from where those can under-go degradation or re-initiation of translation. However, not all of the mRNAs are sorted into the SGs, few of them remain translated under the stress. The mechanism behind this specificity is currently unknown.

In this study we addressed the following questions:

From one side, what serves as a sorting mechanism for the mRNAs to be recruited into the SGs? From the other side, whether the stress alters the methylation pattern and which mechanism states behind the potential changes if they occur. Overall, if these two features are interconnected then how exactly this is happening.

3.2 Materials and methods

3.2.1 Deep-sequencing: PAR-CLIP, RNA-Seq, Ribo-Seq, m^{6}A-Seq

As the main focus of this dissertation is a computational part of the project, therefore some experimental details omitted. Step-by-step sequencing protocols can be found in the Materials and methods section of the original publication (Anders et al, 2018).

In brief, HEK293 cells expressing N-terminally FLAG-tagged TIA1 under doxycycline-dependent promoter (Damgaard & Lykke-Andersen, 2011) were used in the study. For the simplicity these cells are called HEK-TIA1. Oxidative stress was elicited by adding sodium arsenite (AS) for 30 minutes at 37°C. Sequencing libraries were generated under one of the following conditions: unstressed HEK-TIA1 cells or stressed with 200 or 500 µM AS. RNA-seq, Ribo-seq and PAR-CLIP libraries were obtained from both, control and 200 µM AS; while following the exposure to 500 µM AS stress no translation occurred after 30 minutes, therefore no Ribo-seq samples were generated under the harsh stress condition.

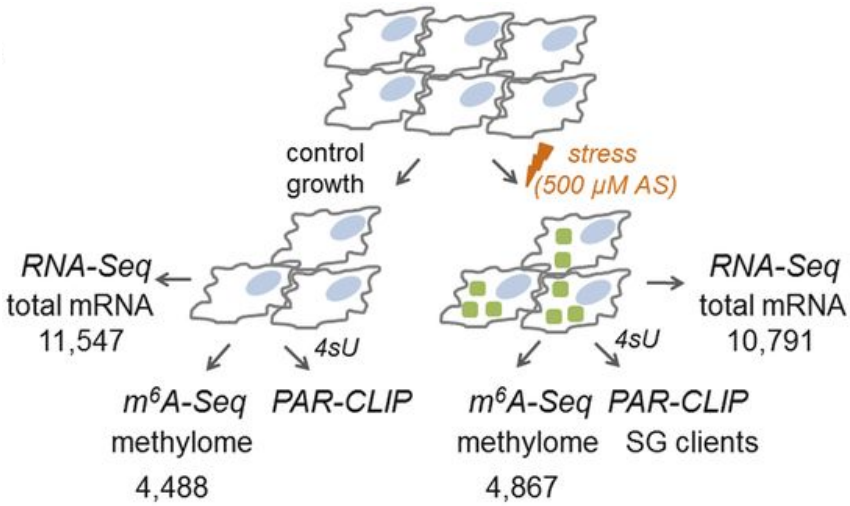

Figure 3.3: Overview of the experimental setup. Numbers denote mRNAs identified in each deep sequencing approach.

m^{6}A-Seq libraries were generated under the control condition and after the exposure to 500 µM AS for 30 minutes. Two samples have been produced for each experiment: an m^{6}A-Seq sample, where m^{6}A-specific antibodies were used to immunoprecipitate RNA, and a corresponding input RNA-seq sample, created with a standard RNA-seq protocol (Fig.3.3).

m^{6}A-Seq experiments and initial preprocessing of m^{6}A-Seq datasets have been performed by Jun Zhou and Yuanhui Mao, Division of Nutritional Science, Cornell University, Ithaca, NY, USA.

All the libraries in this study were sequenced on a HiSeq2000 (Illumina) machine.

3.2.2 Preprocessing and mapping of sequencing data

Sequencing reads from RNA-seq, Ribo-seq and PAR-CLIP experiments were preprocessed following a common pipeline. First, sequenced reads were quality trimmed using fastx-toolkit version 0.0.13.2 (quality threshold: 20), sequencing adapters were cut using cutadapt version 1.8.3 (minimal overlap: 1 nt), and processed reads were

mapped to the human genome (version GRCh37, Ensembl) using Bowtie version 1.1.2 either uniquely or allowing multimapping with a maximum of two mismatches.

Parameter settings for unique mapping, used in most of the cases, have been set as following -l 16 -n 1 -e 50 -m 1 –strata –best y. Parameter settings for multimapping, used for processing of Ribo-seq data under stress, have been defined slightly different: -l 16 -n 1 -e 50 -m 10 –strata –best y, where -m is referred to the number of locations in the genome where particular read can be aligned with the same - highest - probability. These settings were required to be applied to Ribo-seq samples under the stress in order to receive a comparable coverage of the genome, since the translation was majorly shut down under thr stress compared to the control condition. Reads aligning to rRNA and tRNA genes were excluded prior to the genome mapping, rRNA mapping has been done separately allowing no mismatches to only one copy of the rRNA reference sequences.

Uniquely mapped RPF reads (Ribo-Seq), fragmented RNA reads (RNA-Seq) and reads originated from PAR-CLIP were used to generate gene read counts with HT-Seq 0.11.1 (htseq-count) or bedtools 2.28.0 (coverageBed function with -s parameter, strand dependent counting of the reads). The annotation file containing the longest transcripts corresponding to each protein coding gene has been created based on the Ensembl GRCh37 annotation. Ambiguous reads were excluded by counting only the number of reads whose middle nucleotide (or the 5′ nt of the middle position for even read length) fell within the annotated feature. The reads were normalized as reads per kilobase per million mapped reads (RPKM units) and the total mapped reads per million (RPM units) (Mortazavi et al, 2008).

All sequencing experiments were performed in two biological replicates. Based on the high correlation between the replicates for RNA-seq and Ribo-seq data ($R^2 > 0.9$ for all datasets, Person correlation coefficient), reads from biological replicates were merged together into metagene sets following the standard algorithm as described earlier (Ingolia et al, 2009).

m^6A-Seq reads and the corresponding input RNA-Seq reads (20–40 nt), which served as a basal signal, were aligned to NCBI RefSeq mRNA sequences and UCSC genome

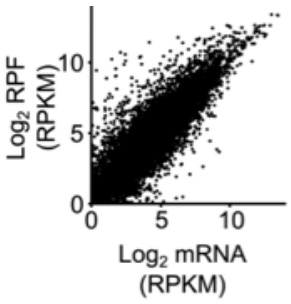

Figure 3.4: Correlation between the total mRNA detected in the RNA-Seq and translated genes generating RPFs in the ribosome profiling under control growth. R^2= 0.838, Pearson correlation coefficient.

sequences (version GRCh37, Ensembl) using Tophat version 2.1.1 (using the parameters –bowtie 1 –no-novel-juncs -G) as described previously (Trapnell et al, 2009).

3.2.3 Processing pipeline and downstream statistical analysis

Under control condition most of the transcribed mRNAs (detected in RNA-seq) were also translated (detected in Ribo-seq) (Fig.3.4).

The ribosomal density (RD) for each transcript (also known as "translation efficiency" - TE), was determined by the density of ribosomes from Ribo-seq per mRNA from RNA-seq dataset (Ingolia et al, 2009) and computed as follows:

$$RD = \frac{RPF[RPM]}{mRNA[RPM]} \tag{3.1}$$

RD values of all protein coding genes were normalized to the RD of mitochondrial genes as described (Iwasaki et al, 2016). Expression of the mitochondrially encoded genes remained unchanged under stress and therefore their RD values served as baseline for normalization of RD values of the nuclearly encoded genes (Fig.3.5).

Cumulative (also known as "metagene") profiles of the read density for RPFs and mRNA have been computed following the published algorithm (Gerashchenko et al, 2012). High ribosome occupancy at the start of the CDS under the oxidative stress

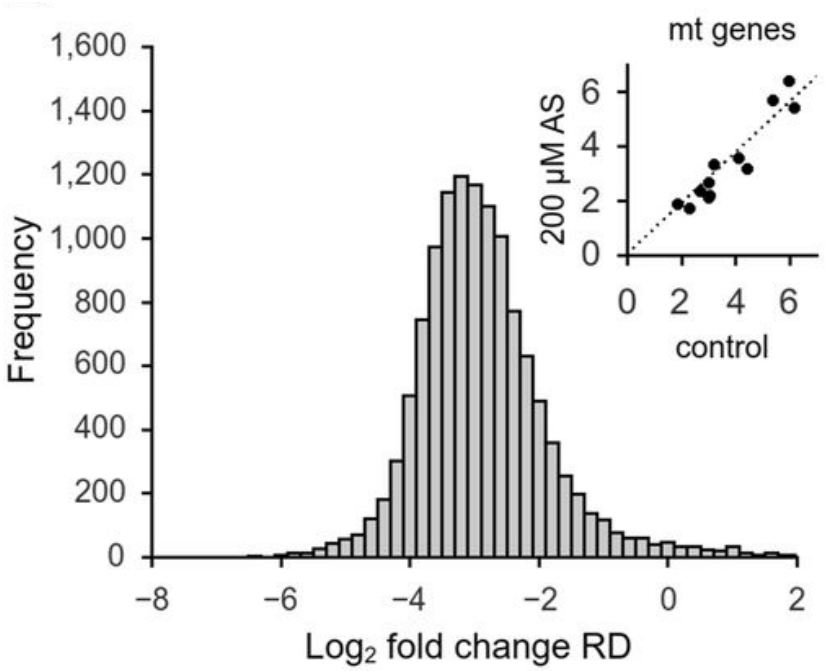

Figure 3.5: Log-changes of the RD values between control cells and following 200 µM AS. Inset: RD values of the mitochondrially encoded genes, which remained unaffected by stress and used for the normalization.

indicated that the ribosomes accumulate towards the 5′ end of the CDS instead of being uniformly distributed over the CDS length (Fig.3.6). Therefore, not all RPFs obtained in the sample reported on translation under stress. To distinguish between genuinely translated transcripts and those whose translation was inhibited by stress, the following ratio has been introduced:

$$Rt = \frac{Total\ RPF\ reads\ of\ initial\ stalled\ peak\ (first\ 100\ nt)\ [RPKM]}{Total\ RPF\ reads\ over\ the\ full\ gene\ length\ [RPKM]} \tag{3.2}$$

A threshold of Rt =0.5 has been defined, which allowed grouping all the transcripts obtained from Ribo-seq sample under 200 µM AS. 108 mRNAs exhibited $Rt \leq 0.5$ and were considered as actively translated, whereas for the others 2 104 genes detected in the sample, the majority of the RPFs were stalled at initiation, which led to $Rt > 0.5$. Those transcripts were not translated under stress, so they were designated as triaged for SGs.

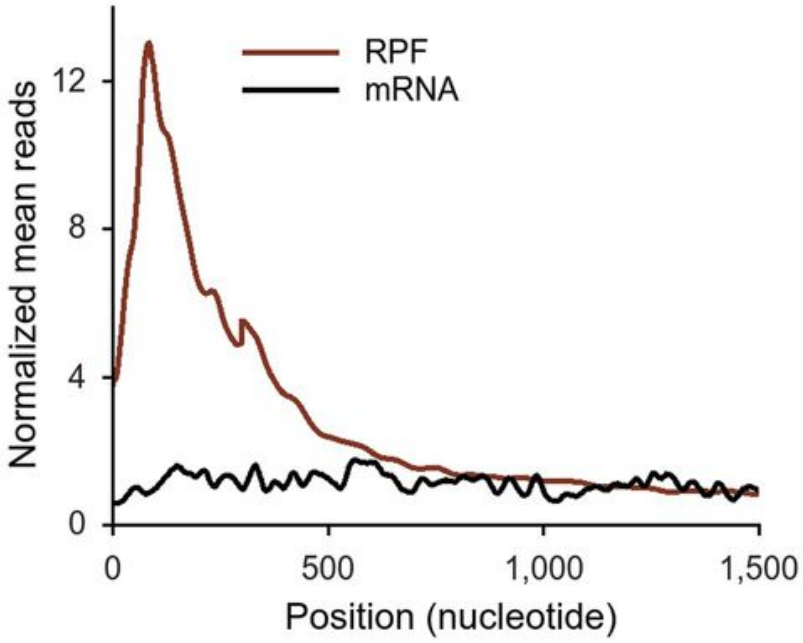

Figure 3.6: Cumulative ("metagene") profile of the read density as a function of position for RPFs (from Ribo-Seq) and mRNAs (from RNA-Seq) under 200 µM AS stress. The expressed genes were individually normalized, aligned at their start codons and averaged independently of their expression levels.

In PAR-CLIP experiments, SG clients in cells stressed with 200 µM AS or 500 µM AS were selected based on a threshold of $\log_2 = 2$ enrichment over the control (unstressed) growth condition. The variability between biological replicates in PAR-CLIP experiments (Pearson correlation coefficient) from cells exposed to mild or harsh stress were $R^2 = 0.695$ and $R^2=0.735$, respectively. The replicates were merged together for the downstream analysis. Furthermore, the correlation between the selected SG clients at both stress conditions was very high (Fig.3.7).

Taking into account Ribo-seq dataset at 200 µM AS, where 2 104 transcripts have been considered as triaged for SG, the list of selected SG clients under mild stress (200 µM AS) included both types of genes - detected in PAR-CLIP and triaged in corresponding Ribo-Seq (Fig.3.8). Most of the transcripts identified in Ribo-Seq at 200 µM AS with halted translation and designated as triaged for SG were also found among the SG clients from PAR-CLIP at harsh stress (500 µM AS). This comparison was relevant, as no Ribo-seq experiment could be performed under 500 µM AS

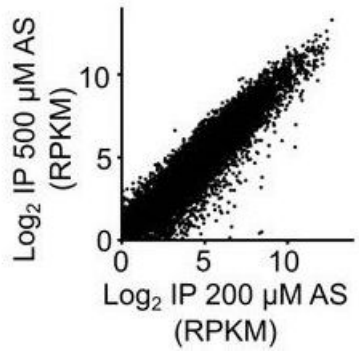

Figure 3.7: Correlation of the SG transcripts detected under 200 and 500 µM AS stress in the PAR-CLIP experiments (two merged biological replicates). R^2 = 0.883, Pearson correlation coefficient.

because of the absence of translation.

Thus, all selected transcripts (either enriched in PAR-CLIP datasets under the mild or harsh stress or designated as triaged in Ribo-Seq under the mild stress) were merged together into a metagene set of SG clients containing 6 020 transcripts in total. These mRNAs found in SGs comprised a large range of expression in RNA-seq datasets, which remained unchanged between the control condition and mild stress (Fig.3.9).

Gene function analysis (GO enrichment) among the mRNAs translated under stress (108) was performed with the DAVID tool version 6.8.

Statistical analysis was mainly performed in R version 3.3.3 using RStudio environment (version 1.1.4) with a partial usage of the relevant Bioconductor (https://www.bioconductor.org/) software packages.

3.2.4 Motif analysis

De novo search for DRACH motifs was performed using a command-line FIMO version 5.0.5 (FIMO-MEME suite; http://meme-suite.org/doc/fimo.html), which allows to scan a set of sequences for individual matches to the motif. The transcript groups of interest and their corresponding sequences (e.g. 5′ UTRs, CDSs, 3′ UTRs) were prepared with Ensembl Biomart. The threshold of the motif matches has been set at

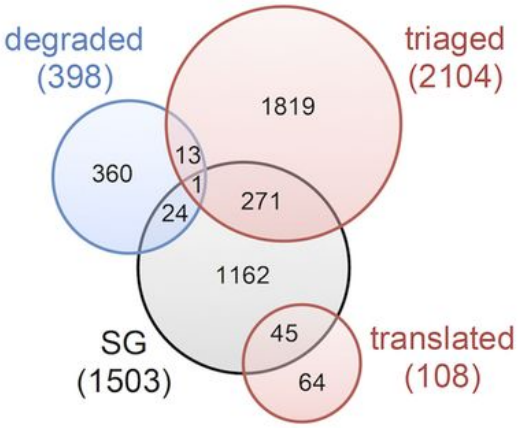

Figure 3.8: Venn diagram of the distribution of various transcript groups detected under the mild (200 µM AS) stress. SG - mRNAs in SGs, detected in the PAR-CLIP; degraded - mRNAs, identified from the RNA-Seq under stress degraded compared to the control RNA-Seq; red circles - triaged and translated - two groups of mRNAs with RPFs in the Ribo-Seq.

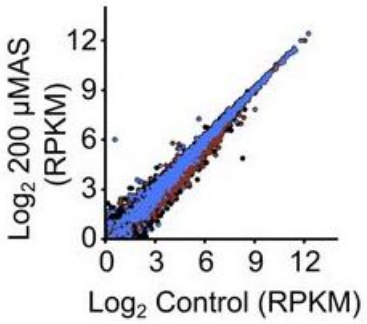

Figure 3.9: Identified SG clients spread large expression span. Total mRNAs – black, mRNAs in SGs – blue, mRNAs generating RPFs under 200 µM AS – red.

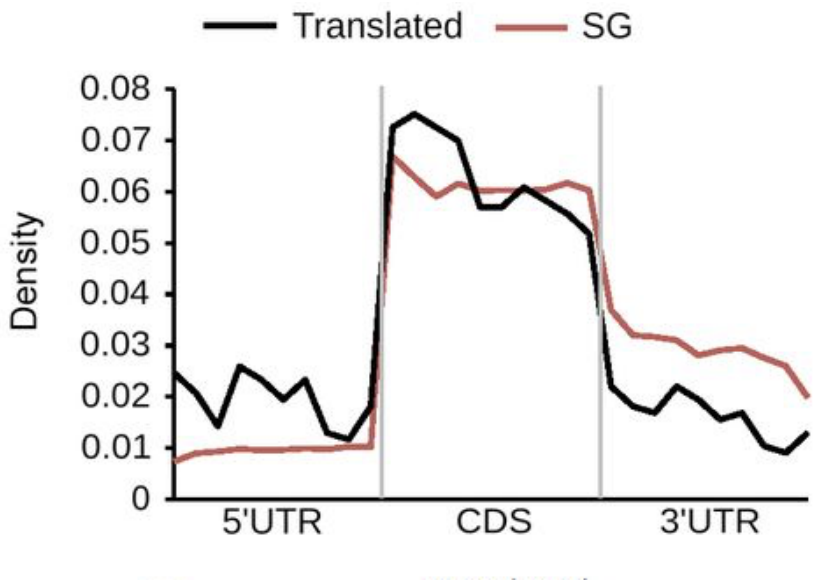

Figure 3.10: Distribution of the predicted DRACH motifs in different transcript segments of the SG clients and translated genes. Genes translated under the mild stress (200 µM AS) contain more DRACH motifs in their 5´ UTRs compared with the 5´ UTRs of the SG clients, P = 1.4×10^{-3}, Mann–Whitney test.

p-value < 0.001, only the coding strands of the given mRNAs have been considered. For comparing the number of DRACH motifs in each transcript region, 5′ UTRs, CDSs and 3′ UTRs were divided into equal bins of comparable length (10 equal segments each). The amount of motifs in each segment was averaged over the whole set of genes in the selected group (Fig.3.10). Mann-Whitney test has been used for comparison of two independent groups (variables) - subsets of population of total mRNAs in the cell - translated genes vs. SG clients; the dependent variable was represented by density of motifs in each of 10 segments along 5′ UTR (Fig.3.10).

A general search to discover novel motifs among the sequences of SG clients was performed using a command-line MEME suite version 5.0.5. Any number of motifs per sequence was allowed, the potential motif length till 10 nt has been considered. Typical motifs scoring for various RNA-binding proteins have been identified (Fig.3.11).

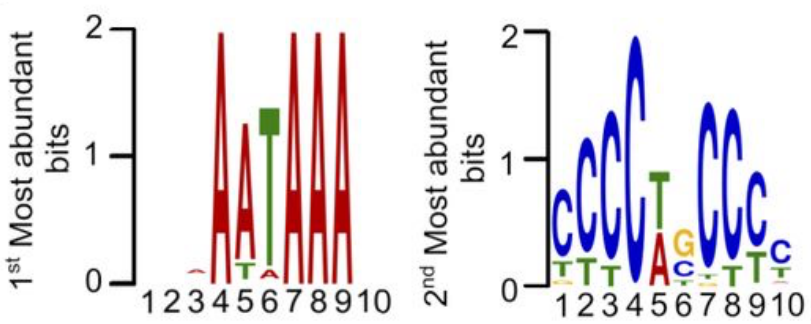

Figure 3.11: The top-two abundant motifs among the SG clients found by MEME motif search (rest was insignificant).

3.2.5 Identification of methylated sites

All full-length mapped reads were used to generate an m^{6}A-Seq coverage profile for individual protein coding genes. To compare metagene m^{6}A profiles between control and stress (500 µM AS) conditions, the raw coverage values were first internally normalized by the mean coverage of each individual gene. The genes with maximal coverage less than 15 reads were excluded from the further consideration. Next, the corresponding RNA-Seq profiles for each sample were subtracted as a basal coverage level from the normalized m^{6}A-Seq profiles of the individual genes. This resulted in generation of adjusted m^{6}A-Seq profiles, where methylated regions of the transcripts were detected as peaks in coverage from immunoprecipitated RNA relative to the input RNA-Seq sample.

Since the m^{6}A peaks in the m^{6}A-Seq vary in length (with a median ~100 nt), in order to increase the precision of the localization and exclude the potential false-positives, the peaks were assigned to the predicted DRACH motifs. Peaks occurring in regions covering at least one DRACH motif predicted by MEME (see above) were selected for further analysis. If more than one DRACH motif was found within a given m^{6}A peak, in this case all of them have been considered as methylated.

Metagene profiles of m^{6}A distribution used for the comparison between control versus stress condition (500 µM AS) were created following the similar strategy as described above for the predicted DRACH motifs and were derived by averaging all

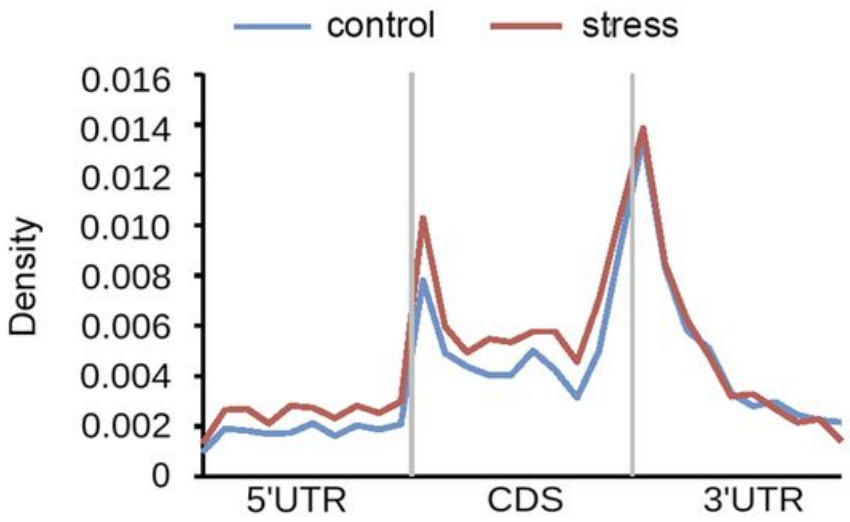

Figure 3.12: Metagene profiles of distribution of m^6A sites along different transcript regions of SG mRNAs from control condition or under 500 µM AS (stress). P = 1.4×10^{-3} for 5′ UTRs and P = 1.6×10^{-2} for 5′ vicinity of the CDSs; Mann–Whitney test between stress vs. control.

adjusted m^6A-Seq profiles of individually detected genes. Transcript regions were binned for comparable lengths (10 segments within each region, such as 5′ UTR, CDS, 3′ UTR). Next, the ratio between m^6A-modified DRACH motifs detected in a given m^6A-Seq sample and total number of predicted DRACH motifs has been determined for each transcript segment. Finally, all these ratios in each segment were averaged for the given set of genes within the sample (Fig.3.12). Mann-Whitney test has been used for comparison of two independent groups (variables) - subsets of population of total mRNAs in the cell - genes detected under the control vs. genes detected under the stress; the dependent variable was represented by density of motifs in each of 10 segments along 5′ UTR or in each of the first 3 segments along CDSs - 5′ vicinity of the CDSs (Fig.3.12).

Box-plots of methylated sites represent the similar ratios with the difference that the length was not taken into account and the whole transcripts were considered (Fig.3.21 - left; Fig.3.22, Fig.3.25).

Figure 3.13: Venn diagram of m^{6}A peaks identified in HEK-TIA1 (HEK) cells in this study compared to those in U2OS-G3BP1 (U2OS) cells from the previously published study (quote), both at permissive control growth.

3.2.6 Connection with previous studies using publicly available data

For the comparison of methylation profiles between the different cell lines, m^{6}A peaks in HEK-TIA1 from this study have been compared to those of U2OS-cells from a previously published m^{6}A-Seq dataset (Xiang et al, 2017) (Fig.3.13). The comparison has been performed on the transcript level and on the level of single methylation sites, which have been detected. The m^{6}A-Seq data for U2OS cells has been downloaded from GEO database (https://www.ncbi.nlm.nih.gov/geo/), accession number GSE92867.

SG clients identified in this study using combined sequencing approaches have been compared to the mRNA clients of the each of three m^{6}A readers (YTHDF1, YTHDF2 and YTHDF3) previously identified by PAR-CLIP (Wang et al, 2015) (Fig.3.14 and Fig. 3.15). The PAR-CLIP data has been downloaded from GEO database (https://www.ncbi.nlm.nih.gov/geo/), accession number GSE63591. The lists of the target genes have been taken from Supplementary table 1 of the manuscript.

3.2.7 Data access

As part of the publishing process, deep-sequencing data from RNA-Seq, Ribo-Seq, PAR-CLIP and m^{6}A-Seq experiments were deposited in the BioSample database (https://www.ncbi.nim.nih.gov/biosample/) under accession number SRP121376.

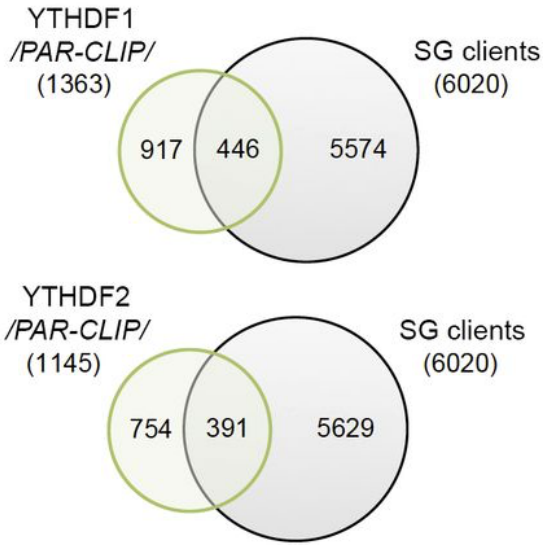

Figure 3.14: Venn diagrams of mRNA clients of YTHDF1 (top) and YTHDF2 (bottom) identified by PAR-CLIP in Wang et al (2015) compared with the SG transcripts identified in this study. P = 0.006 (YTHDF1), P = 3.9×10^{-4} (YTHDF2), hypergeometric test.

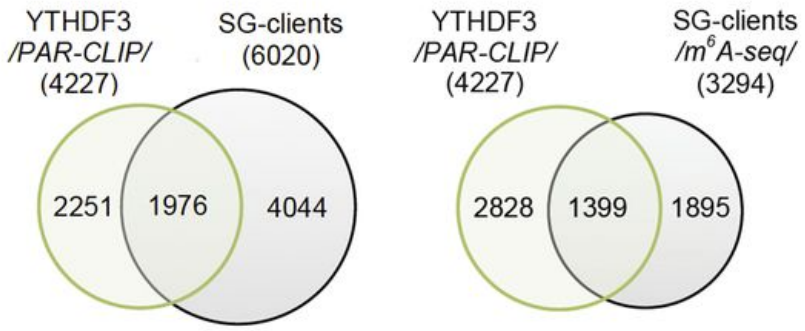

Figure 3.15: Venn diagram of the common clients between the YTHDF3 PAR-CLIP target genes (4 227) and total SG clients (6 020 mRNAs) - left; and the methylated SG clients detected with in m^6A-Seq (3 294 mRNAs) - right. P = 1.07×10^{-155} (for PAR-CLIP, left) and P = 3.78×10^{-214} (for m^6A-Seq, right), hypergeometric test.

3.3 Results

3.3.1 Additional methylation in mRNAs under oxidative stress

In order to access the dynamics of methylation under stress, we used HEK-TIA1 (Damgaard & Lykke-Andersen, 2011) cell line, which expresses SG marker protein TIA1. Being FLAG-tagged, this protein allows immunofluorescent detection of SGs. Arsenite (AS) has been used to induce the mild (200 µM AS) or harsh (500 µM AS) oxidative stress in cells. SGs were formed in a dose-dependent manner, which has been detected by Maximilian Anders via fluorescent microscopy (Fig S1A in the publication). Further, using m^6A-antibodies to highlight the methylation, m^6A-modified RNAs have been detected co-localized with SGs under mild and harsh oxidative stress as well as under the heat stress, which has been also tested (Fig. 1A in the publication).

RNA-Seq did not reveal any global changes in the total mRNA levels even under the maximal stress dose (500 µM AS) we used in the study (Fig.3.16). When comparing with the total mRNAs detected under permissive growth (control), only a 6.5% decrease in the total mRNAs under stress has been observed. The oxidative stress has been exposed for 30 mins in our case and lacking of RNA degradation is consistent with previous study, which reported that short AS does not trigger a global transcriptional response while only a few specific mRNAs are affected (Andreev et al, 2015).

We extracted a set of mRNAs, which were expressed under the control condition, but were missing and therefore degraded under the harsh stress. Gene Ontology (GO) enrichment analysis revealed the categories related to transcription (fold enrichment: 1.94; $P = 7.88 \times 10^{-8}$) being enriched (enrichment score: 7.67) in degraded gene-set. The subcategories included "regulation of transcription" (fold enrichment: 2.01; $P = 8.28 \times 10^{-7}$), "transcription factor activity" (fold enrichment: 2.11; $P = 3.12 \times 10^{-5}$), and "DNA binding" (fold enrichment: 1.64; $P = 5.56 \times 10^{-4}$).

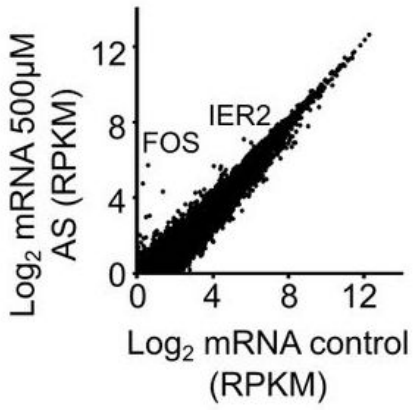

Figure 3.16: Comparison between total mRNA from control and 500 µM AS stress cells determined by RNA-Seq. Genes with significantly increased expression under stress are designated. R^2 = 0.978, Pearson correlation coefficient.

Besides a minor set of degraded mRNAs, the RNA expression of two other genes was significantly up-regulated under stress: immediate early response protein 2 (IER2) and FOS transcription factor. Both of these mRNAs are clearly associated with an oxidative stress response, as they are usually up-regulated under the environmental conditions, which increase intracellular levels of reactive oxygen species (Cekaite et al, 2007) (Fig.3.16).

m^6A has been shown to modulate mRNA stability (Wang et al, 2014; Mauer et al, 2017), therefore we used RNA-Seq to determine the effect of the silencing of "writer" complex on the total mRNA abundance. Overall, comparing with the total mRNA levels detected under control condition to those following the knockdown of "writer" complex, which would lead to the absence of methylation, we did not detect significant changes in the global mRNA abundance (Fig.3.17), therefore the presence of methylation itself does not affect the mRNA levels.

Then, we compared the level of the m^6A under control and stress conditions. When testing the total RNA, we observed an increased m^6A signal under stress compared to the permissive growth suggesting the methylation increasing under stress. It followed a stress-dose-dependent manner, in analogy to the SG formation (dot-blot Fig 1D in the publication). However, as it has been shown previously, a large fraction of non-coding RNAs (e.g. rRNAs) are also methylated (Pan, 2013) and therefore

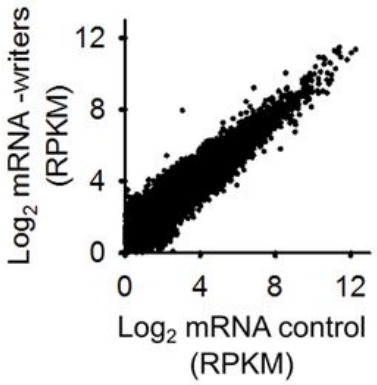

Figure 3.17: Comparison of total mRNA expression in control growth condition and following a knockdown of the "writer" complex (-writers) determined by RNA-Seq. R^2 = 0.928, Pearson correlation coefficient.

can be recognized by the m^{6}A antibodies. Thus, a large portion of the m^{6}A signal from the total RNA may correspond to the non-coding RNA species, which are more abundant in the cell than mRNAs.

To extract the methylation pattern of mRNAs only, we performed m^{6}A-sequencing (Meyer et al, 2012; Zhou et al, 2015) under harsh oxidative stress (500 µM) and permissive growth. Under control condition, 8 046 m^{6}A peaks have been detected in total at consensus DRACH motifs. Those peaks appeared within 4 488 unique mRNAs. So, from 11 547 mRNAs identified in the RNA-Seq, 38.9% contained at least one m^{6}A peak. The number of m^{6}A peaks increased significantly under oxidative stress: from 8 046 under control condition to 9 142 under stress (P =2.8 × 10^{-6}; Fig.3.18, bottom diagram). The number of mRNAs, where m^{6}A peaks have been detected, increased as well (44.2% of 10 791 detected total mRNAs in the RNA-Seq, P =2.8 × 10^{-6}; Fig.3.18, top diagram). This findings support the previous observation of the increased m^{6}A levels under stress and suggest that mRNAs and not (only) other RNA species exhibit stress-induced additional methylation.

Importantly, these additional m^{6}A peaks appeared not only in mRNAs, which were not modified under the control condition, but also on transcripts that were already partly methylated under the control permissive growth (Fig.3.18).

All the sequencing experiments have been performed on HEK293 cells, but when

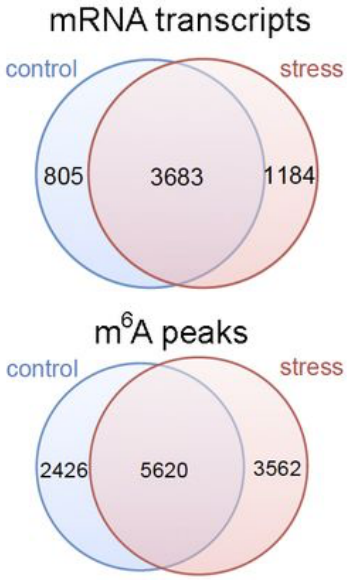

Figure 3.18: Venn diagram of mRNAs with at least one m^{6}A peak detected (top) and venn diagram of all unique methylation sites identified in mRNAs (bottom) under the control growth and following 500 µM AS stress.

comparing the response from the different human-derived cell lines, m^{6}A modifications appeared largely overlapping between HEK293 and U2OS cells (Xiang et al, 2017) (Fig.3.19), which suggests a conserved methylation pattern.

3.3.2 Distinct m^{6}A pattern of mRNAs in SGs

As a next step, we asked, whether the enrichment of m^{6}A under stress was associated with mRNAs, which were recruited into SGs. We used another sequencing approach in addition to RNA-Seq, Ribo-Seq and m^{6}A-Seq. We isolated the mRNAs from SGs using photo-activatable ribonucleoside cross-linking and immunoprecipitation approach (PAR-CLIP) (Hafner et al, 2010) (Fig.3.3). Briefly, SGs were stabilized with 4sU-mediated cross-linking of mRNAs to RNA-binding proteins, and intact SGs were isolated using previously described protocol (Khong et al, 2017).

To extract the set of mRNA clients segregated in the SGs in response to harsh AS stress (500 µM), I defined a threshold of two-fold enrichment over PAR-CLIP control (all sequencing reads were preliminary normalized to RPKM). 6 020 unique mRNAs

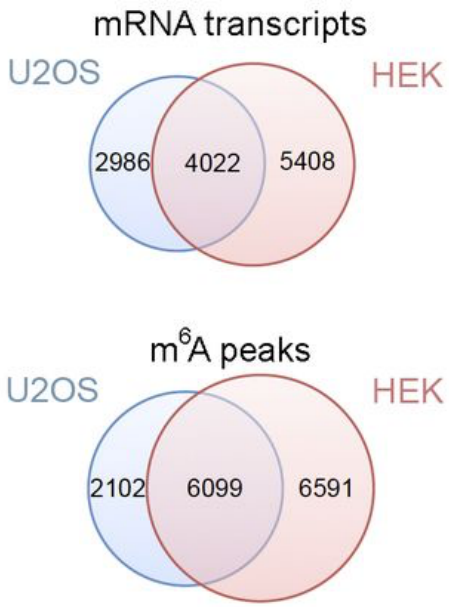

Figure 3.19: Venn diagram of mRNAs containing at least one m^{6}A modification (top) and unique m^{6}A peaks detected in mRNAs (bottom) identified in HEK-TIA1 (HEK) cells in this study compared with those in U2OS cells from Xiang et al (2017).

have been identified as associated with the SGs (Fig.3.20). The number of unique mRNAs detected in SGs was much larger (6 020 of 10 791 total mRNAs from RNA-Seq) than has been previously found in the SG cores (Khong et al, 2017), suggesting that our approach allowed to capture the full-size SGs: not only the cores, but also peripheries, which contained the other mRNAs.

Although we used specific anti-TIA1 antibodies in our PAR-CLIP experiments to pull down the SGs, the motif search in our selected SG mRNA clients revealed the

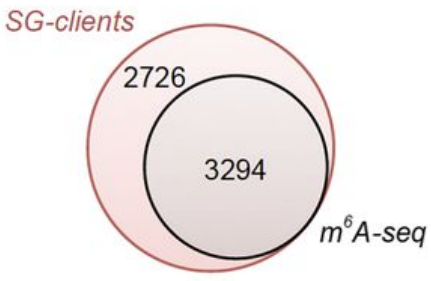

Figure 3.20: Overlap of the SG clients from the PAR-CLIP and m^{6}A-Seq experiments.

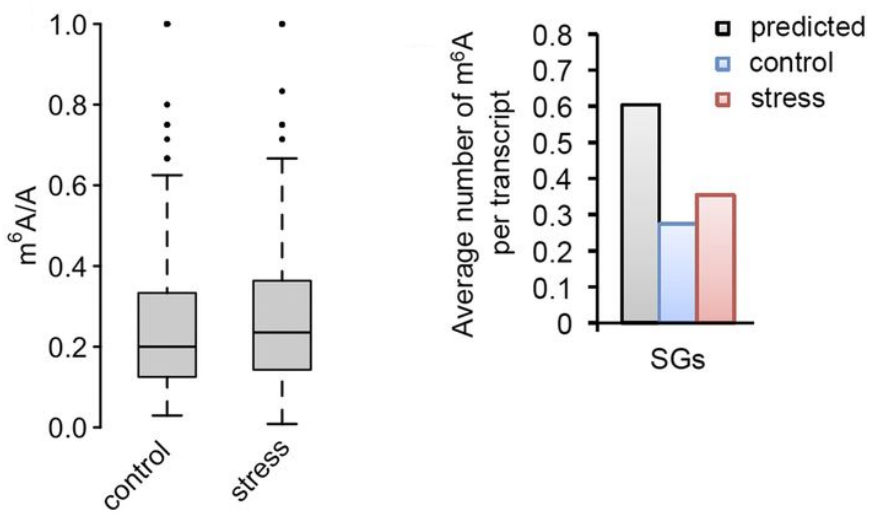

Figure 3.21: Increased methylation of SG mRNAs under oxidative stress. Left - Boxplot of m^6A sites detected in SG transcripts of untreated condition (control) or under the stress (500 µM AS) and presented as a ratio of the total m^6A sites - predicted DRACH motifs designated as A in the ratio m^6A/A. $P = 5.1 \times 10^{-4}$ control vs. stress, Mann–Whitney test. Right - Average number of m^6A-modified DRACH motifs detected in the SG mRNAs under stress compared with their methylation level under control growth. $P = 1.49 \times 10^{-5}$ control vs. stress, Mann–Whitney test. The average number of all predicted DRACH motifs per mRNA is included for comparison.

typical RNA-binding motifs, but not only the TIA1-binding motifs (Fig.3.10). This suggests that through the approach we used - unspecific 4sU-mediated cross-linking - we captured diverse mRNAs binding to different RNA-binding proteins.

Cross-comparing the mRNA clients in SGs extracted from PAR-CLIP with a set of mRNAs with detected methylation sites from m^6A-Seq under stress, we have found that 54.7% of mRNAs in SGs were methylated (Fig.3.20). Those mRNAs had significantly higher proportion of m^6A peaks (Fig.3.21 - left) and higher number of methylation sites per transcript (Fig.3.21 - right), both compared to the control condition. Mann-Whitney test has been used for comparison of two independent groups (variables) - subsets of population of all SG-clients - SG-clients detected under the control vs. SG-clients detected under the stress; the dependent variable was represented by m^6A/A ratio for each gene within the sets (Fig.3.21 - left) or by number of detected m^6A sites for each gene within the sets (Fig.3.22 - left).

Importantly, that even under the stress condition, when more methylation sites have

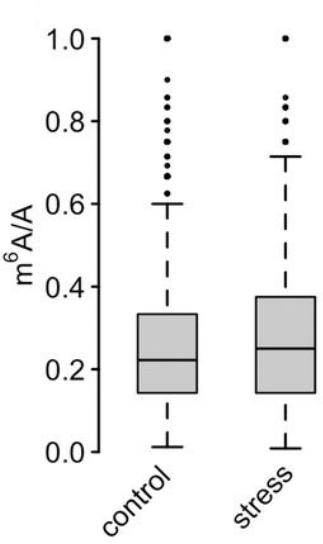

Figure 3.22: Box-plot of m^6A sites detected across all mRNAs of untreated condition (control) or under the stress (500 µM AS) and presented as a ratio of the total m^6A sites - predicted DRACH motifs designated as A in the ratio m^6A/A. P =2.8×10^{-6} control vs. stress, Mann–Whitney test.

been detected, not all of the predicted m^6A sites (falling into consensus motif - DRACH), were methylated (Fig.3.21 - right).

Moreover, the stress-induced m^6A peaks, detected in mRNAs associated with SGs (Fig.3.21 - left), displayed 96% of all mRNAs, which m^6A signals increased in response to stress (Fig.3.22). This observation suggests that most m^6A-modified mRNAs were sorted into SGs. Mann-Whitney test has been used for comparison of two independent groups (variables) - subsets of population of total mRNAs in the cell - all mRNAs detected under the control vs. all mRNAs detected under the stress, where the dependent variable was represented by m^6A/A ratio for each gene within the sets (Fig.3.22).

Next, we analyzed the distribution of m^6A peaks along the transcripts and their localization within the different segments, such as 3′ UTRs, CDSs and 5′ UTRs. The transcripts corresponding to the set of mRNAs in SGs were binned to equal lengths for comparison (Fig.3.12). On a global scale, we observed that following the stress exposure, the number of m^6A sites increased in the 5′ UTRs and 5′ vicinity of CDSs compared to the distribution of methylation sites under the control condition (Fig.3.12).

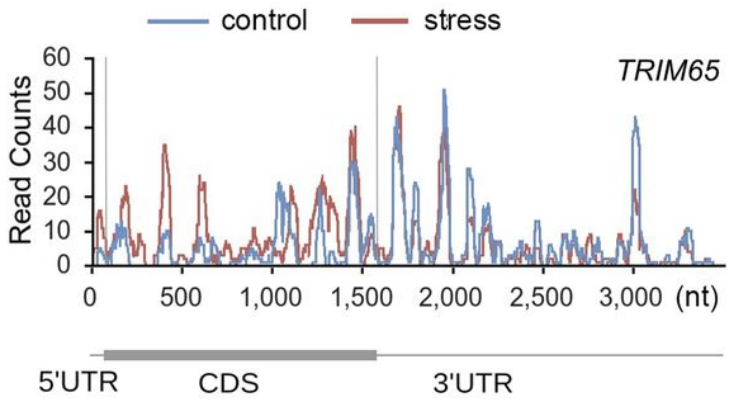

Figure 3.23: An example of stress-induced increase in methylation in the SG mRNA (*TRIM65*).

Same methylation trend remained on a single-gene level (Fig.3.23) Conversely, the m^6A pattern around the stop codons and the 3′ UTRs (Fig.3.12 and Fig.3.23), which has been previously shown to control mRNA stability (Meyer et al, 2012; Wang et al, 2014), remained unaffected by stress. Our data suggests a region-specific methylation pattern for the mRNAs, which were sorted into SGs under stress.

3.3.3 Translationally active mRNAs are methylated in the 5′UTRs under control and stress conditions

As has been reported previously, m^6A in the 5′ UTR is responsible for the cap-independent translation of mRNAs under heat stress (Meyer et al, 2015; Wang et al, 2015; Zhou et al, 2015). The typical translation initiation in eukaryotes involves 5′ cap structure, which is required for efficient binding of translation initiation factors and this mode of translation is called cap-dependent. However, under the various stress conditions cap-independent mode of translation is becoming prevalent (Sonenberg & Hinnebusch, 2009).

We have also detected a greater m^6A level in the 5′ UTR of mRNAs under the oxidative stress, thus, we conducted a separate analysis of the methylation pattern of

transcripts remained translationally active under stress in order to reveal the differences with those sorted into SGs. Under harsh stress (500 µM AS) translation was almost completely repressed. We observed it on the polysome profiling - there was no apparent polysomal fraction, which was supposed to report on ribosomes undergo translation (Fig. S1A of the original publication, performed by Maximillian Anders).

Taking this observation into account, we selected a mild stress (200 µM) - the condition when all three pools of mRNAs existed in the cytosol: actively translated transcripts, mRNAs stalled at translation initiation (intermediate step - see below), and mRNAs already sorted into SGs. To identify the mRNAs in each of these states we combined various sequencing approaches - PAR-CLIP, RNA-Seq and Ribo-Seq (Fig.3.3).

Based on the Ribo-seq data, under the mild stress (200 µM AS), some translation activity still remained in the cell. However, a significant global reduction of translation has been observed when comparing with control growth condition - median reduction of the ribosome density [RD] of $\log_2 = 2.9$ (See Materials and Methods - 3.2.3, Fig.3.5). RNA-seq data perfectly correlated between the control and stress conditions suggesting that transcription was unaffected (Fig.3.9 and Fig.3.16).

In Ribo-Seq under the mild stress, 2 212 unique mRNA transcripts generated sequencing reads - ribosome-protected fragments (RPFs); mRNAs with various expression levels were present in this gene-set (Fig.3.9). However, instead of being distributed along the whole length of the transcripts, which would report on the translation (Ingolia et al, 2009) (Fig.3.24 - top), under stress, most RPFs accumulated at the beginning of the CDS (Fig.3.6 and Fig.3.24 - bottom). These mRNAs were stalled at initiation and early elongation (~first 100 nt of CDS) and were not translated under stress despite being detected in Ribo-Seq. In analogy, similar observation has been made in a previous study, but in relation to thermal stress (Liu et al, 2013).

To distinguish the mRNAs, genuinely translated under stress, and those stalled at initiation, we introduced the translation ratio (Rt) allowing us to select mRNAs

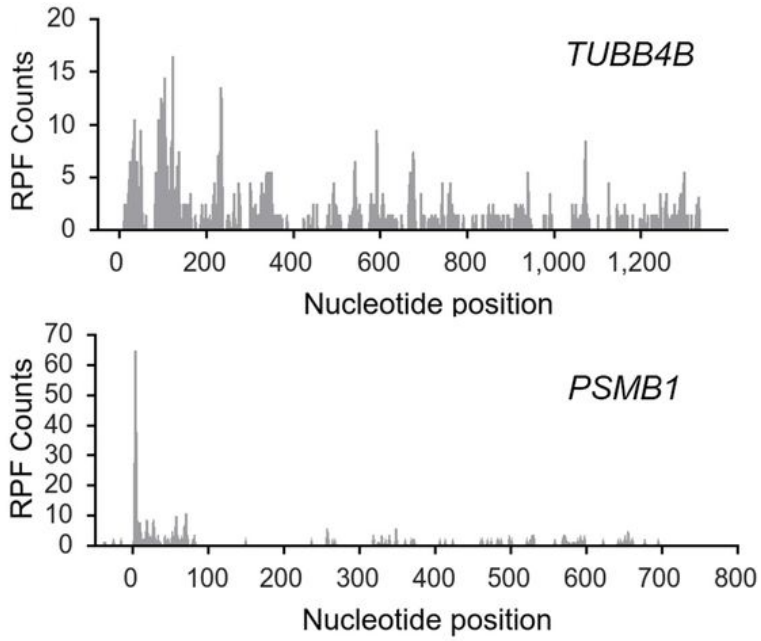

Figure 3.24: Representative example of a transcript (*TUBB4B*) genuinely translated under stress and a transcript with stalled translation (*PSMB1*). The first nt of the start codon is designated as 0.

with a uniform RPF distribution and define them as translated (See Matherial and Methods - 3.2.3). Following the selection criteria, 108 mRNAs fallen into "translated" category (Fig.3.3 and Fig.3.8). Gene Ontology analysis of those genes revealed the enriched terms (enrichment score 12.2) such as "translation" (fold enrichment 10.26; P =1.73×10^{-10}), "nonsense-mediated mRNA decay" (fold enrichment 20.37; P =1.43×10^{-13}), and "rRNA processing" (fold enrichment 11.33; P =2.58×10^{-10}).

Transcripts translated under stress were richer in DRACH motifs - predicted methylation sites - in their 5′ UTRs compared to the set of transcripts sorted into SGs (Fig.3.10). Most of these mRNAs were methylated under control condition (Fig.3.25), which is in line with previous observations (Meyer et al, 2015; Zhou et al, 2015), then, under stress condition, the m^6A level in 5′ UTR of a set of translated mRNAs did not change (Fig.3.25). Mann-Whitney test has been used for comparison of two independent groups (variables) - subsets of population of genuinely translated transcripts (108), which were detected under the control vs. those detected under the stress; the dependent variable was represented by m^6A/A ratio for each gene within the sets.

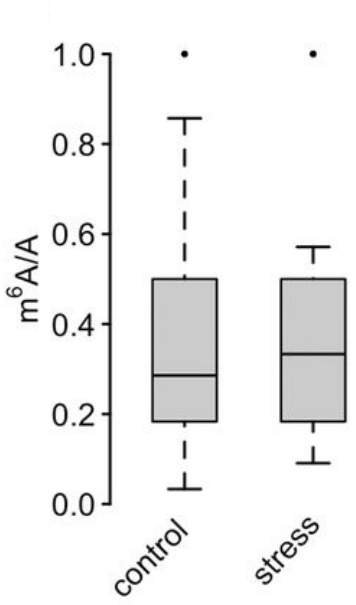

Figure 3.25: Box-plot of m^6A sites detected in the genuinely translated 108 transcripts under the control growth or under stress, presented as a ratio of the total m^6A sites - predicted DRACH motifs designated as A in the ratio m^6A/A. P = 0.97 control versus stress, Mann–Whitney test.

The rest of the transcripts detected in Ribo-Seq under stress (2 104) showed the translation, which was stalled at initiation and/or early elongation, so the majority of the RPFs raised from the beginning of CDS and their Rt > 0.5. Considering the PAR-CLIP data, many of these mRNAs were already sequestered into SGs under the mild stress (200 µM AS) or completely segregated in the SGs under harsh stress (500 µM AS; Fig.3.7). 69.7% of these mRNAs showed an increase in m^6A level in their 5′ termini under harsh stress, which is consistent with our previous observations of enriched methylation pattern sorting the mRNAs into SGs. However, the remaining 30.3% did not have any m^6A modifications, suggesting that their sequestering into SGs is most likely driven by stalling at initiation/elongation itself, without the methylation involved. Latter mechanism has been observed and described earlier (Sonenberg & Hinnebusch, 2009; Kedersha et al, 2013).

Taking together, our data showed that mRNAs remained translationally active under stress were highly methylated in their 5′ UTRs under the control condition al-

ready, but stress did not induce additional modifications. Conversely, the majority of mRNAs with stalled translation, which were triaged to SGs, represented a stress-induced methylation in their 5′ UTRs and 5′ vicinity of the CDS, while few of those were sorted into SGs via stalling of translation only.

3.3.4 Triaging of methylated mRNAs to SGs is mediated by "reader" - YTHDF3

To unveil the mechanistic aspect of mRNA recruitment to SGs, we analyzed the "reader" proteins, which selectively recognize m^6A sites and mediate its' functions. YTH domain–containing proteins "reader" proteins are evolutionary conserved cell-type–independent proteins (Edupuganti et al, 2017), which bind the m^6A moiety with their YTH domain (Dominissini et al, 2012). Thus, we first analyzed the localization of three YTH domain–containing proteins (YTHDF1-3) under the harsh stress (500 µM AS). YTHDF3 co-localized exclusively with the SGs, whereas YTHDF1 only marginally co-localized with the SGs, and YTHDF2 remained in cytosol and did not move to SGs (fluorescent microscopy by Max Anders - Fig 4A in the original publication). Similarly to the previous observations (Meyer et al, 2015; Zhou et al, 2015; Li et al, 2017), under the control growth conditions, YTHDF1 and YTHDF2 have been detected in cytosol and nucleus, whileYTHDF3 resided exclusively in the cytosol (fluorescent microscopy by Maximilian Anders - Fig S4A in the original publication). Then, we used a knockdown of the "writer" complex to prevent the occurrence of *de novo* methylation, we detected that localization of the YTHDF3 in SGs was completely abrogated while had no effect on YTHDF1 (fluorescent microscopy by Maximilian Anders - Fig 4A in the original publication). Altogether, these observations proposed a new role of YTHDF3 "reader" in recruiting m^6A-modified mRNAs into SGs, while the localization of the YTHDF1 in SGs was passive and has no mechanistic insight.

Importantly, performing the knockdown of the "writer" complex or YTHDF3 itself led to the decreased amount of methylated mRNAs in SGs, while the amount of non-methylated mRNAs was almost not affected (dot blot by Maximilian Anders -

Fig 4C in the original publication) suggesting another recruiting mechanism taking place in the case of absence of methylation in particular mRNA. This correlates with the observation from the sequencing data, where we found 45.3% of all mRNAs in SGs were not methylated (Fig.3.20).

Next, we analyzed our PAR-CLIP and m^6A-Seq data in combination with previously published PAR-CLIP data, where the specific mRNA clients for YTHDF1-3 have been identified (Shi et al, 2017). We aimed to identify the specificity of YTHDF3 towards mRNAs in SGS and to cross-compare those with the other two "readers".

We detected a considerable overlap of YTHDF3 clients with mRNAs sequestered in SG (Fig.3.15, left venn diagram - total PAR-CLIP and m^6A-seq clients were included; right venn diagram - only those mRNAs detected with m^6A-seq were shown). While the other readers, YTHDF1 and YTHDF2, had some clients in common with SG clients, the overlap was much smaller compared to YTHDF3 (Fig.3.14).

Altogether, our results from experimental data and sequencing data analysis allowed to conclude that YTHDF3 mediates the triaging of mRNAs, which were m^6A modified in their 5′ termini, to SGs under oxidative stress.

3.4 Conclusions

In this study we revealed two modes of sequestering of mRNAs into SGs under oxidative stress. The first, larger fraction of mRNAs (~55%), carry DRACH motifs in the 5′ vicinity of the transcripts, allowing the stress to induce position-specific m^6A modifications, which then serve as a mechanism for triaging them into SGs (Fig.3.26). The second fraction of mRNAs (~45%) are not methylated and most likely triaged to the SGs via stress-induced stalling at initiation (Fig.3.26), which correlated with previously suggested mechanisms (Sonenberg & Hinnebusch, 2009; Kedersha et al, 2013).

We found that m^6A in the 5′ UTR and 5′ vicinity of CDSs are dynamic and induced by oxidative stress, which is recognized by the YTHDF3 reader and allows it to relocate those mRNAs to SGs (Fig.3.26).

It has been shown previously that SGs are enriched with proteins, which contain IDRs, allowing them to self-aggregate through hetero- and homotypic interactions (Gilks et al, 2004; Lin et al, 2015; Jain et al, 2016). Structural predictions of the YTH-domain "reader" proteins revealed Gln/Asn-rich IDRs in all three of them (YTHDF1-3). Therefore, in our case, when YTHDF3 binds the stress-induced m^6A on mRNAs, it apparently relocates them to SGs through protein-protein interactions with its IDR (Fig.3.26).

Considering the earlier studies (Decker & Parker, 2012; Kedersha et al, 2013) and our observations of SGs formation in absence of methylation, the primary nucleation and assembly of SGs happens in an m^6A-independent manner involving the fraction of translationally stalled mRNAs together with initiation factors. This is supported by the detection of the fraction of non-methylated mRNAs in SGs (Fig.3.26). While none-translating mRNAs stalled at initiation and/or early elongation form the core of the SGs, we propose that the methylated mRNAs might be mostly located in the more dynamic SGs peripheries (Fig 3.26), however, this feature is yet unexplored.

mRNAs remained genuinely translated under stress were enriched in methylation in their 5' UTRs, and their methylation pattern was stable under control and stress conditions, in contract with SG-clients, which were additionally methylated under the stress. It is still unknown what allows the YTHDF3 reader to discriminate those two types of mRNAs while recruiting the second pool into SGs, however, the difference in dynamics of methylation may play a crucial role in recognition. It is clear that YTHDF3 relocate m^6A-modified mRNA in SGs.

Overall, our study revealed an unexpected feature of YTHDF3 reader protein in sequestering mRNAs into SGs under oxidative stress along with a specific methylation pattern and stress-induced increase in m^6A signal of the mRNAs, which are sorted into SGs following the methylation-dependent mechanism (Fig.3.26).

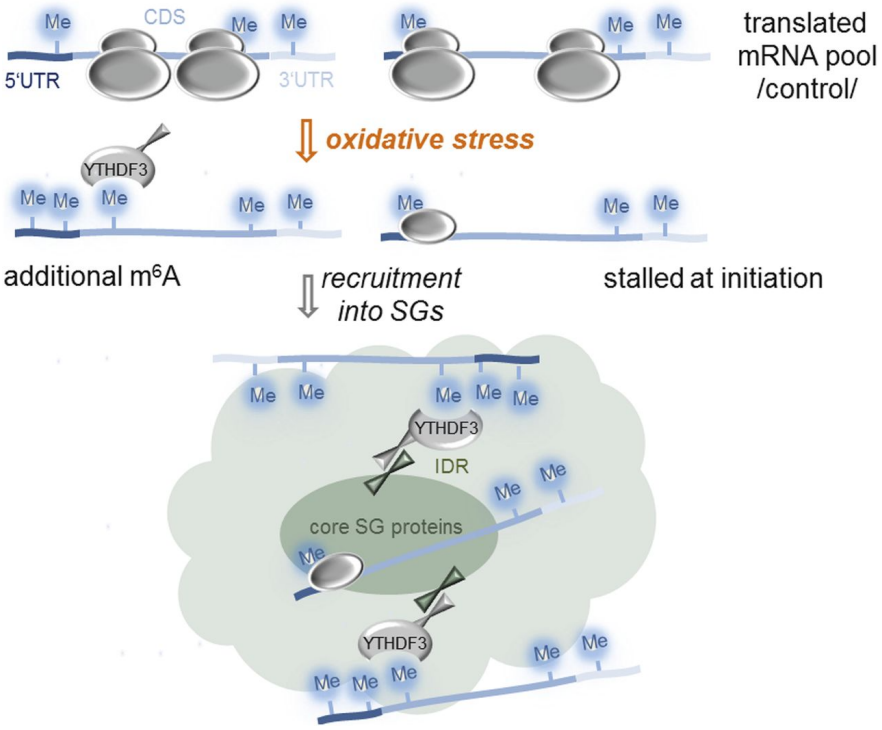

Figure 3.26: Proposed model of mRNA triaging into SGs: mRNAs are recruited into SGs via stress-induced methylation in an YTHDF3-dependent manner (left side) or via stress-induced translational stalling at initiation (right side).

Chapter 4

RNA-chaperone Hfq is crucial for ribosome biogenesis

4.1 Background

Ribosomes are ribonucleoprotein complexes, which are responsible for the protein synthesis in the cell and where the translation of the mRNA into protein occurs, therefore the ribosome are also called biosynthetic or translational machineries. The ribosomes are composed of two major components: the small ribosomal subunit and the large subunit. Each of the subunits contains of one or several ribosomal RNA (rRNA) molecules and a variety of ribosomal proteins (r-proteins). The structure and function of the ribosome during translation is relatively well described in the literature (Mura, 2013), while the molecular events underlying the assembly of the ribosomes is still under investigation. Ribosome biogenesis is a highly coordinated cellular process of making ribosomes, which takes a major fraction of the cell's energy in bacteria and consists of multiple steps involving variety of factors.

As has been discussed in the previous chapters of this thesis, the rRNA is the most abundant type of the RNA across all the domains of life, while the "sizes" of rRNA molecules and the ribosomes (both measured in Svedberg units=S) are different between pro- and eukaryotes. Prokaryotes have 70S ribosomes while eukaryotic ribosomes are larger - 80S.

Since the study described in this chapter involves Escherichia coli (also known as E. coli) - gram-negative bacteria, more attention and details will be given to the structure and biogenesis of the prokaryotic ribosome.

Prokaryotic ribosome is made up of a 30S small subunit, which contains 16S (1542 nt) rRNA, and a large 50S subunit containing 23S (2904 nt) and 5S (120 nt) rRNA. Both subunits include numerous r-proteins (54 in total described in E. coli, Chen, 2012) and altogether compose a functional 70S ribosome (Shajani et al, 2011).

Ribosome biogenesis factors are proteins that transiently bind to assembling ribosomal particles and increase the efficiency of subunit maturation. Mutations affecting many of these proteins cause dysfunctional ribosomes. Over 60 of such factors are present in E.coli, they include GTPases, rRNA modification enzymes, helicases, and other maturation factors, which assist rRNA folding and r-protein assembly pathway (Davis & Williamson, 2017).

The bacterial Hfq protein is encoded by the hfq gene that has been initially discovered as a host factor, essential for replication of the bacteriophage QB in E.coli and is called accordingly. Up to date, it is clear that an Hfq is an abundant bacterial RNA-binding protein from the Sm/ Lsm family of proteins (Wilusz & Wilusz, 2013) with multiple important biological functions, which has homologues in all domains of life.

The RNA chaperone Hfq binds small noncoding RNAs (sRNAs) and facilitates interactions their mRNA targets. By this, Hfq controls the expression of mRNAs and can lead to their up- or down-regulation, which is a crucial mechanism for the formation of the response to the changing environmental conditions and various stresses (Vogel & Luisi, 2011; Hajnsdorf & Boni, 2012; Updegrove et al, 2016). However, in many bacterial species, Hfq is not an essential component of sRNA-dependent pathway (Christiansen et al, 2006; Rochat et al, 2015), which suggests there are other yet undefined functions of Hfq beyond its already described role in sRNA activity.

It has been shown already decades ago that Hfq interacts *in vitro* with the 16S rRNA (de Haseth & Uhlenbeck, 1980), but the functional role of this interaction remained

unknown. In E. coli, a cross-linking-based study identified an interaction of Hfq with rRNA *in vivo* as well (Tree et al, 2014). Another study reported an interaction between Hfq and S12 ribosomal protein (Strader et al, 2013). However, all of these interactions have been described without any functional insights behind. The question remained: whether this interactions are redundant or Hfq may be involved in rRNA processing and formation of the ribosome?

In the current study, we have identified a novel role of Hfq in ribosome biogenesis and proposed Hfq as a novel ribosome biogenesis factor, which is required for the formation of the functional ribosomes in the cell.

4.2 Materials and methods

4.2.1 Deep-sequencing: RNA-seq and Ribo-seq

As in the previous chapter, the experimental details are described in brief, and the detailed protocols, strain information and growth conditions can be found in the Materials and methods section of the original publication (Andrade et al, 2018).

The study has been conducted in collaboration with the laboratory of Prof. Cecilia M Arraiano (Universidade Nova de Lisboa), while all the sequencing-related experiments and bioinformatics analysis have been performed in the laboratory of Prof. Zoya Ignatova (Inst. of Biochemistry and Molecular Biology, University of Hamburg).

All E. coli strains used in this study were derived from strains MG1693 or MC1061 following up with the respective mutation for Hfq depletion mutant. Strains were grown in LB medium and the cultures were collected at exponential (OD_{600} ~ 0.5) or stationary phase (after ~ 14 h growth). Ribosome-protected fragments for Ribo-seq and randomly fragmented mRNA for RNA-Seq were isolated following the previously described protocol (Del Campo et al, 2015) and sequenced on a HiSeq2000 (Illumina) machine.

4.2.2 Preprocessing and mapping of sequencing data

Sequenced reads from Ribo-Seq and RNA-Seq experiments were processed with a common pipeline. Preprocessing has been performed as described in previous chapters: the reads were quality trimmed using fastx-toolkit (0.0.13.2; quality threshold: 20), and sequencing adapters were cut using cutadapt (1.8.3); minimal overlap: 1 nt). Pre-processed samples were uniquely mapped to the E. coli genome (strain MG1655, version U00096.3, downloaded from NCBI) using Bowtie (1.1.2) allowing a maximum of two mismatches.

The number of raw reads was used to generate read counts for each gene, by counting the number of reads whose middle nucleotide fell in the coding sequence (for even read lengths the 5′ nucleotide from the middle position was used). CoverageBed function from bedtools (version 2.28.0) has been run twice: to produce the gene-wise read counts (-s parameter) as well as single-nucleotide resolution counts (-s -d parameters), which have been used to create the profiles for each gene (Fig.4.1). E.coli genome annotation files in BED format have been created and customized based on the annotation version U00096.3. Gene read counts were normalized to RPKM (on gene-level) or RPM (relevant for single nucleotide resolution). Additionally, total RNA was spiked in with RNA standards (ERCC, Thermo, Germany): the mixture containing the spike-ins in the known concentrations were added to the RNA-Seq samples and used to set the detection threshold in each sequencing set. Similar threshold has been applied to the corresponding Ribo-Seq samples.

4.2.3 Processing pipeline and downstream statistical analysis

In order to access the effect of Hfq depletion on translational efficiency, the density of the ribosomes from Ribo-seq samples per mRNA from RNA-seq has been computed. First, on a single-gene level, following formula 3 from the chapter 3 of this thesis, which refers to the method described by Ingolia, 2009. Then, to compare the global translational efficiency between the mutant and the wild-type strains, the density

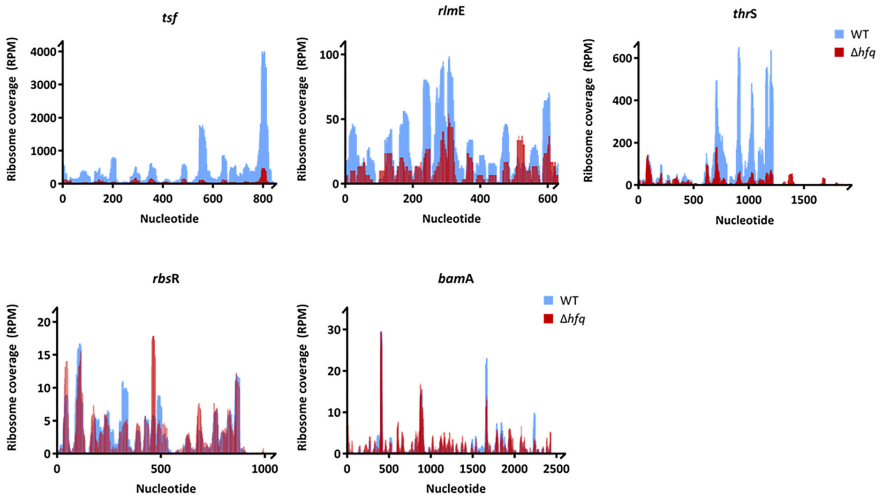

Figure 4.1: Representative examples of coverage profiles of down-regulated genes affected by Hfq deletion (top) and genes, whose expression remained unchanged upon Hfq deactivation (bottom).

plots of $\log_{10}$TE (also called RD) values have been created (Fig.4.2). The observed reduction of the efficiency in Hfq-depleted strain was significant (Mann–Whitney U-test or Wilcoxon rank-sum test, P = 0.0001996).

Considering the organisation E.coli genome, which contains a high number of overlapping genes, the initial filter was applied, so the overlapping genes were excluded from this analysis. In these genes, the initiation of the downstream gene and termination of the upstream gene fall into the one genomic region, therefore the reads in this region cannot be assigned to either gene, which would bias the RD analysis, since each of the genes have different mRNA abundance.

The downstream analysis and visualisation have been performed in R (version 3.3.3) using RStudio environment (version 1.1.4) partially with a help of the relevant packages from GitHub, R-cran and Bioconductor repositories.

Cumulative profiles of read density for RPFs have been computed as described (Ingolia et al, 2009); the overlapping genes have been excluded in order not to bias the amount of the reads aligned the translation initiation region (Fig.4.3, 1 075 and 1 231 genes from wild-type and Hfq-depleted strains, respectively, were considered).

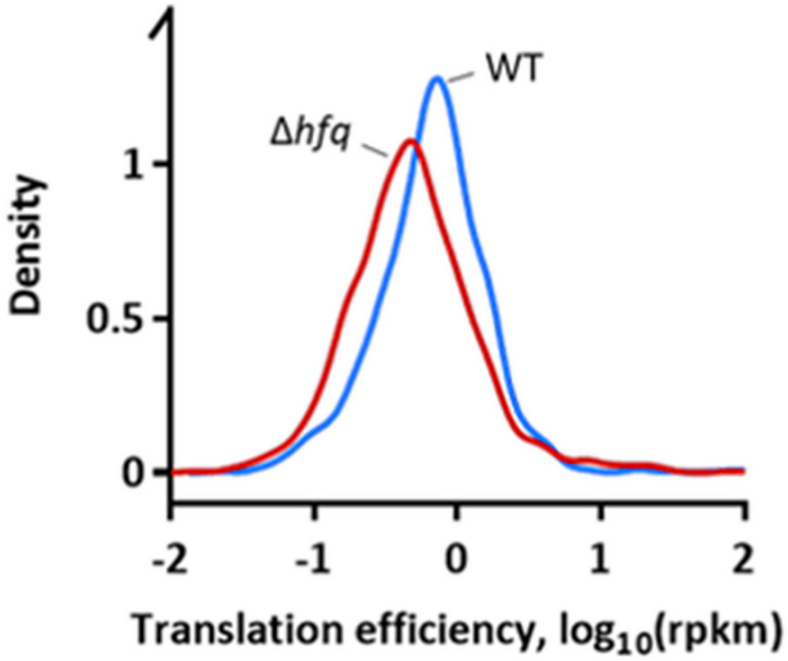

Figure 4.2: Translation efficiency (Ribosomal density) of wild-type and Hfq-depleted cells obtained by Ribo-Seq.

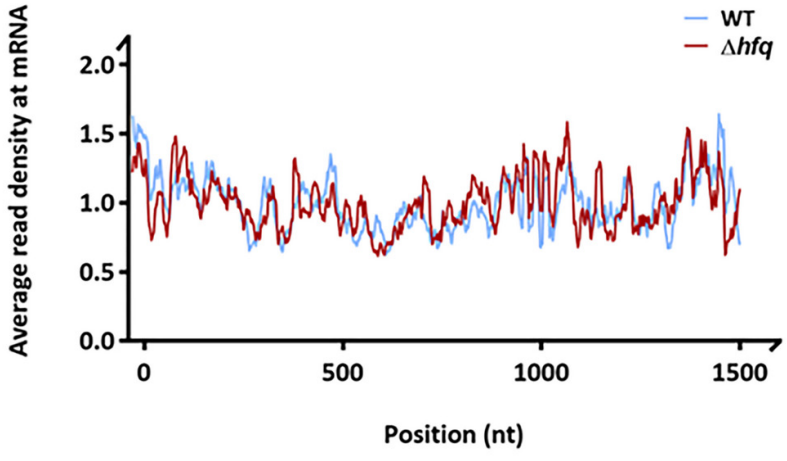

Figure 4.3: Cumulative (metagene) profile of the read density as a function of position for RPFs from wild-type and depletion mutant. The genes were individually normalized, aligned at their start codons, and averaged.

Differential gene expression analysis was based on fold-changes of RPKM values between Ribo-seq datasets. The threshold of 2 has been used to define the set of differently expressed genes, which have been assigned to the categories based on gene ontology (GO) terms. GO enrichment including statistical analysis was performed using the bioinformatics tools and gene lists (E.coli K12 genome) from Gene Ontology Consortium (http://geneontology.org/).

4.2.4 Proving reproducibility using publicly available data

To prove the reproducibility of the obtained results, independent control RNA-Seq and Ribo-Seq samples were taken from the previously published E. coli dataset (Hwang & Buskirk, 2017). The strain and growing conditions were similar to those used for the wild-type libraries in this study. Both RNA-seq and Ribo-seq datasets from wild-type and Hfq depletion mutant were compared to the independent wild-type samples, downloaded from GEO database, accession number GSE85540). The mapping and normalization of the datasets were performed similar to those from the current study (see 3.2.1-3.2.3). The reproducibility was very high, with R^2 = 0.865 and R^2 = 0.816 (Spearman correlation coefficient) for the RNA-Seq and Ribo-seq datasets, respectively. Overall, published dataset, being a truly independent biological replicate, correlated well with the wild-type from this study, which served as an additional prove of the results and observations.

4.2.5 Data access

As a part of the publishing process, deep-sequencing data from RNA-seq and Ribo-seq were deposited in the GEO database (https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE100373.

4.3 Results

While most of the laboratory experiments for this publication have been performed by José Andrade and Ricardo dos Santos at Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, Portugal, the combination of deep-sequencing approaches - RNA-Seq and Ribo-Seq and their downstream bioinformatics analysis conducted by me, were crucially important to shape the conclusions and to provide a full picture of the overall outcome of the study. In the current thesis, I have concentrated on that part of the results, which have been raised from my own analysis on the deep-sequencing data, and its integration in the overall picture. I am just briefly going through the results from the other experiments instead of describing the whole story, which can be found in the publication itself.

4.3.1 Hfq is required for maturation of 16S rRNA

We observed that the correct processing and folding of the 16S rRNA is affected by Hfq inactivation, resulting in the structural occlusion of the residues. Our observations indicate that Hfq interacts with 17S rRNA and this interaction is necessary for the processing and folding of the mature 16S rRNA, and the absence of Hfq affects the formation of the central pseudoknot of 16S rRNA. More details on the experiments behind these conclusions can be found in the publication.

4.3.2 Inactivation of Hfq leads to defects in ribosome biogenesis

Our data clearly demonstrated that the inactivation of Hfq leads to a reduction in the pool of 70S ribosomes, which is a result of defects in the 70S assembly, since both of the sub-unites 30S and 50S were present in the cell in appropriate amounts irrespective of the mutation in Hfq.

rRNA synthesis implies the synthesis of r-proteins (Scott et al, 2014). We performed RNA-Seq and Ribo-Seq to assess the expression of the r-proteins at transcriptional

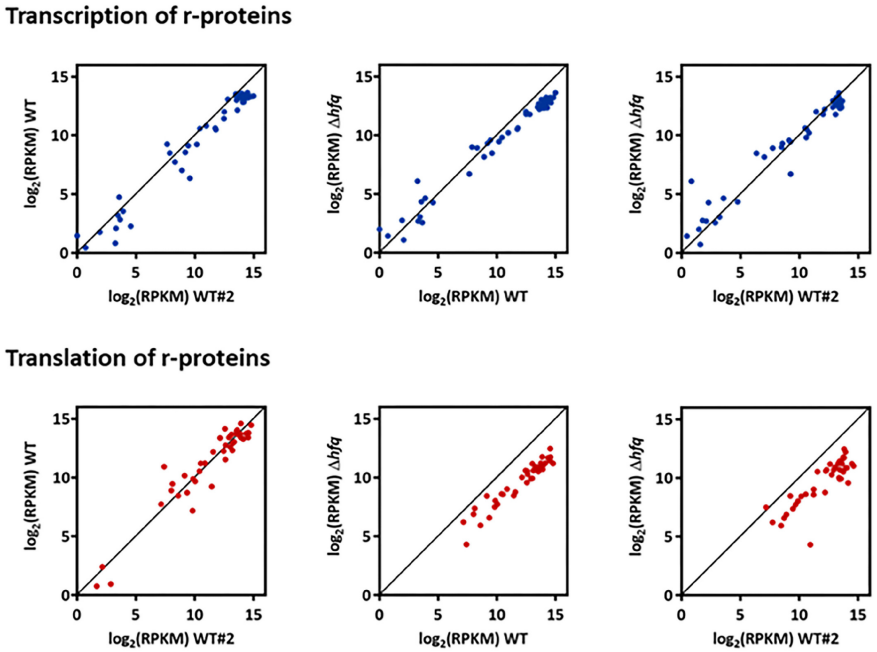

Figure 4.4: Translational down-regulation of r-proteins upon inactivation of Hfq. Comparison of mRNA expression (top) from RNA-Seq and protein production (bottom) from Ribo-Seq between the wild-type and Hfq-depleted strains used in this study and another wild-type dataset (WT#2; Hwang & Buskirk, 2017).

and translational levels, respectively. Strikingly, we observed that all of the ribosomal proteins were significantly translationally down-regulated in the Hfq depletion mutant compared to the wild-type, while the transcriptional level remained unaffected by the depletion and the expression was similar in both strains (Fig.4.4). The second wild-type control has been obtained from the previously published dataset (See Materials and Methods, 4.2.4) and had similar trends when compared to the Hfq depletion mutant. This observation supports the previous studies showing the translational coupling of the expression of the r-proteins and rRNA synthesis (Jinks-Robertson & Nomura, 1981).

Importantly, that even so we observe a translational response from r-proteins upon mutation, the initiation of translation is not affected by the depletion of Hfq, since the cumulative profiles of all expressed genes do not show any differences between the wild-type and depletion mutant (Fig.4.3).

Altogether, our results showed that the depletion of Hfq leads to defect in ribosome

biogenesis, particularly reducing the pool of mature 70S ribosomes, which proposes Hfq as an auxiliary factor regulating ribosome biogenesis.

4.3.3 Hfq copurifies with precursor 30S ribosomes

Hypothesising that Hfq would bind to immature 30S ribosomal subunits, since they are enriched in 17S RNA, we used a knockout mutant of RbfA, which is a late assembly factor that accumulates pre-30S particles (Jones & Inouye, 1996; Thurlow et al, 2016). Strikingly, Hfq was found to co-purify only with immature 30S isolated from the rbfA depletion mutant but not with the mature 30S isolated from the wild-type (Fig. 3B from the original publication). This result supports the idea of Hfq being a novel ribosome assembly factor.

4.3.4 Translation efficiency is affected by Hfq depletion

Defects in ribosome biogenesis can lead to major deficiency in translation. First, when assessed the translation of the depletion mutant by polysome profiling, we observed a reduced polysome fraction compared to that one in the wild-type (Fig.4A in the original publication). Then, we measured a global translation efficiency, which was determined by the density of ribosomes from the Ribo-Seq per mRNA from the RNA-Seq (See Materials and Methods, 4.2.3). It was was significantly reduced in the mutant compared to the wild-type (Mann–Whitney U-test or Wilcoxon rank-sum test, P = 0.0001996; Fig 4.2). Taking together, these results suggest that deactivation of Hfq led to the defects in processing of rRNA precursor and ribosome biogenesis, which resulted in the decreased translation volume and efficiency in comparison to the wild-type E.coli.

Next we asked whether there were any genes, which did not follow the global reduction in translation efficiency. In order to assess it, we performed a fold-change analysis based on the translational changes in Ribo-Seq dataset and considered only those with stable mRNA expression from RNA-Seq. Two-fold enrichment threshold

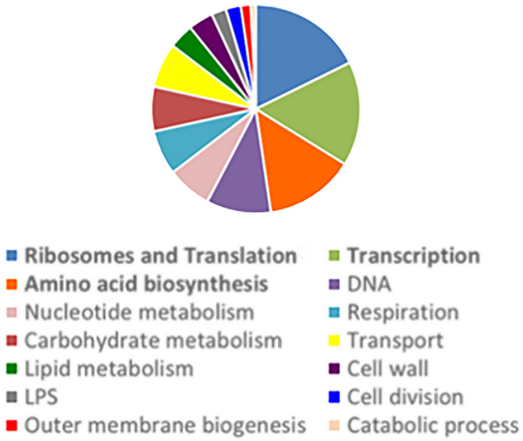

Figure 4.5: GO enrichment analysis of genes, translationally down-regulated upon depletion of Hfq. The top three affected categories are in bold.

has been applied on both up-regulated and down-regulated gene-sets. Then, we performed the Gene Ontology (GO) analysis of the genes, which were translationally down-regulated upon the Hfq depletion. Several pathways showed significant GO term enrichment, especially, the genes participating in ribosome biogenesis, translation, and amino acid metabolism, were mostly affected (Fig.4.5). Translational profiles of representative examples on single-gene level are included: those genes, which were down-regulated in the Hfq depletion mutant (Fig.4.1, top) those, which remained unaffected (Fig.4.1, bottom).

4.3.5 Translation fidelity is affected by Hfq depletion

We compared the quality and accuracy of translation of the Hfq depletion mutant to those the wild-type. The mutant showed an increase in all the kinds of translational errors, such as frameshifting, aberrant initiation from alternative start codon(s) and read-through of a stop codon (Fig.5B in the original publication). This clearly indicated that the accuracy of translation was severely affected by the depletion.

Our experimental data suggest that inactivation of Hfq enhances misreading of mRNA and links together the effect of Hfq depletion on rRNA processing, ribosome biogenesis and translation fidelity.

4.3.6 The distal face of Hfq is crucial for ribosome biogenesis

Hfq assembles into a hexamer of a ring-like shape and therefore has at least three RNA-binding surfaces. As shown in previous studies, sRNAs are preferably bound to the proximal face of the hexamer, while the distal face binds to target mRNAs (Mikulecky et al, 2004; Link et al, 2009; Sauer & Weichenrieder, 2011; Sauer et al, 2012; Zhang et al, 2013). We tested multiple Hfq mutations at the difference surfaces (Zhang et al, 2013) in order to reveal the surface, responsible for the newly discovered Hfq-dependent regulation of ribosome biogenesis. Strikingly, only mutations in the distal face caused reduction in the levels of mature 70S ribosomes, similar to those observed for the Hfq deletion mutant (Fig 2A and B in the original publication), wehreas two other mutants (in the proximal or rim surface) showed the profile similar to the wild-type.

This data allows to conclude that the distal face of Hfq is crucial for the rRNA maturation and is responsible for the role of Hfq in ribosome biogenesis, which suggests that the novel function of Hfq in the ribosome biogenesis might be independent of its' sRNA activity, since another surface is involved.

4.4 Conclusions

Our work unveils a novel role of Hfq in bacterial ribosome biogenesis with important consequences for translation (Fig.4.6).

We demonstrated that Hfq is a new regulator of 16S rRNA maturation, which is required for the correct processing of rRNA. We found that Hfq directly interacts with the 17S rRNA, and its depletion results in the accumulation of unprocessed 17S rRNA precursors. RNA-structure mapping showed that the formation of the central pseudoknot of 16S rRNA is altered in cells lacking Hfq.

Following the inactivation of Hfq, we observed a significant reduction of r-proteins synthesis on the translational level, along with reduction in the levels of mature 70S ribosomes, and accumulation of immature 30S and 50S ribosomal subunits.

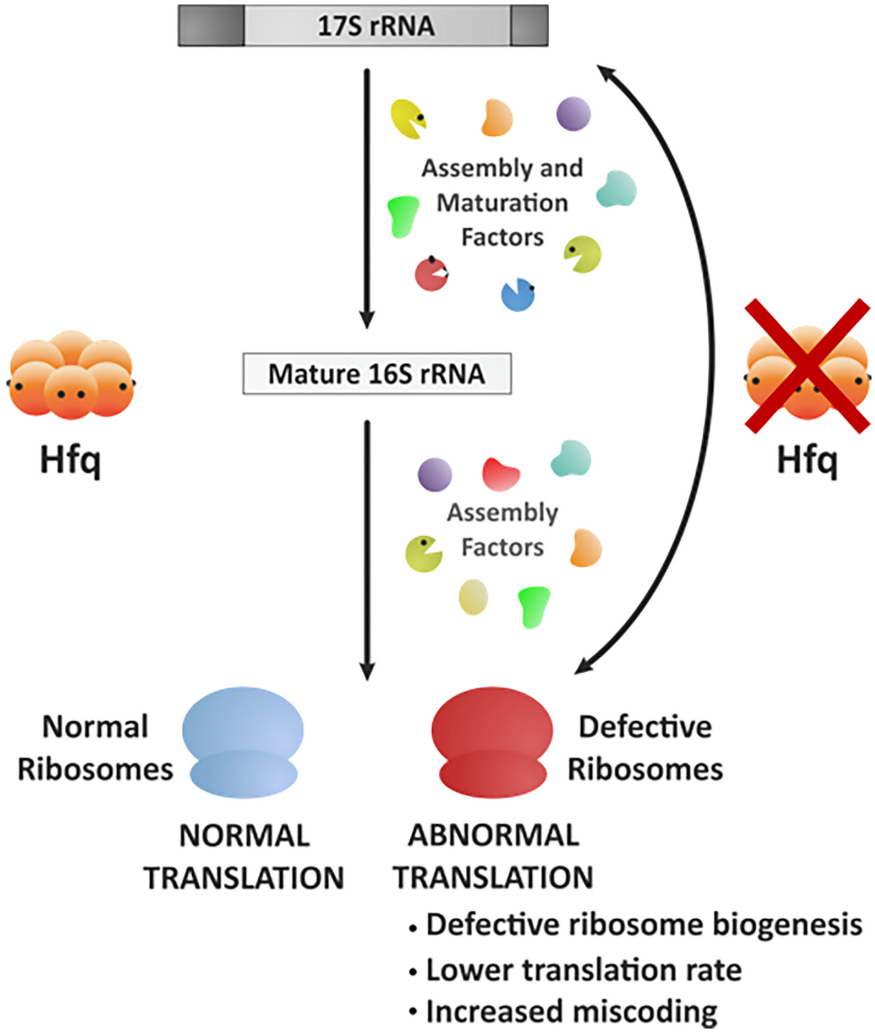

Figure 4.6: Model for the Hfq regulation of ribosome biogenesis

Altogether, the major defects in ribosome biogenesis of Hfq depletion mutant led to global translation deficiency with both, translation efficiency and fidelity being compromised.

In summary, our study showed that Hfq is a new ribosome assembly factor (Fig.4.6). This new role expands the functions of Hfq beyond the the mostly known activity of this protein in the regulation of small non-coding RNA.

Chapter 5

Absolute quantification of translational regulation and burden using combined sequencing approaches

5.1 Background

The central dogma of biology explained in the first chapter of the current thesis is not just a multistage process of protein expression, but also an important mechanism, which allows the cell to function and adapt to the changing environmental conditions. In order to fully understand the response on each of the steps, quantitative methods to monitor the processes of transcription and translation are required (Belliveau et al, 2018). Gene regulatory networks (or genetic circuits) regulate these processes and control where and when they take place.

During the last years, the idea of using the synthetic genetic circuits for understanding the functions of natural gene regulatory networks is spreading across the field (Smanski et al, 2016; Wang et al, 2016). However, the construction of a genetic circuit requires an assembly of multiple DNA-encoded parts controlling the initiation and termination of transcription and translation. This brings in an additional challenge

- necessity to predict how each part of large genetic circuits will function when assembled with others (Cardinale et al, 2013) and how all of the parts will behave in concert. Up to date, there are no approaches to simultaneously measure the performance of multiple parts within a circuit, which prevents the method by itself from reaching the initial goal of the study.

In order to overcome the problem of ambiguous interaction of different parts within the circuit, fluorescent probes and proteins have been used for characterization of the functions of genetic parts (Jones et al, 2014; Hecht et al, 2017). However, another potential bias have been introduced by this approach, as there is a possibility that the fluorescent tag itself affects a part's function (Baens et al, 2006; Margolin, 2012).

The power of the deep sequencing technologies rapidly developing during the last decades is applicable for characterization and debugging genetic parts and circuits. It has clear advantages over fluorescent probes, since it does not require any modification of the circuit DNA and provides a more direct measurement of the processes, such as monitoring of transcription of specific RNAs. Finally, deep sequencing allows to capture the information on the host response and the indirect effects on a part's function.

In 2017, Dr. Thomas Gorochowski, - the collaborator who conceived the study described in the current chapter as well, - used RNA-Seq to characterize every transcriptional component in a large logic circuit composed of 46 genetic parts (Gorochowski et al, 2017). While that study was successful and demonstrated the ability to characterize genetic part function and much more, the RNA-Seq alone as a method of choice limited the approach to purely transcriptional elements, which did not allow to move the quantification towards the physically meaningful units.

In the most recent study, to which this chapter is devoted, we used a combination of deep sequencing methods to surpass previously obtained results. We developed an approach combining Ribo-Seq with quantitative RNA-Seq, which enabled us to characterize endogenous sequences and synthetic genetic parts controlling both levels - transcription and translation - as well as to quantify them in absolute units. Since

Ribo-seq provides position-specific information on translating ribosomes (See Materials and Methods for the precise allocation of ribosome-protected fragments 5.2.3), this allows to calculate the genome-wide protein synthesis rates with a high degree of accuracy, which is comparable to that in quantitative proteomics (Li et al, 2014). By the addition of other experimentally measured cell parameters (such as total protein concentration snd cell numbers), we generated transcription and translation profiles that capture the flux of RNA polymerases (RNAPs) and ribosomes governing these processes.

In this study we applied our method to E. coli and were able to demonstrate how local changes in the profiles can be interpreted using mathematical models and to represent the performance of genetic parts in absolute units. Our study illustrated the genome-wide shifts in transcription and translation, which mark the burden that synthetic genetic constructs place on the host cell.

My contribution to this work included a complete pre- and post-processing, detailed analysis of all the sequencing data along with my participation developing of a method itself and introducing the novel calculations to the datasets, which became particularly challenging. The data analysis was a part of a bigger collaborative work, where the modeling part of the study has been performed by Thomas Gorochowski, and was based on the results obtained from deep sequencing in combination with cellular features measured from the other experiments.

In the current thesis, I highlight that part of the study, which has been performed by me, and concentrate on a characterization of synthetic pseudoknot, which induced translational recoding/frameshifting, and cellular response to the induction of pseudoknot expression. The details on the model itself including the genome-wide calculations of the translation initiation and termination rates can be found in the original publication (Gorochowski et al, 2019).

5.2 Materials and methods

5.2.1 Deep-sequencing: RNA-seq and Ribo-seq

In analogy to the Chapters 3 and 4, the experimental details are partially omitted and can be found in the corresponding section of the publication (Gorochowski et al, 2019).

Two E. coli strains have been used in the current study. Both derived from K12 strain, [K-12, recA1 D(pro-lac) thi ara F':lacIq1 lacZ:: Tn5 proAB+], by the addition of one of the plasmids:

(1) a pBR322-derived plasmid containing lacZ with a fragment insert that contains a truncated lac operon with the Ptac promoter and the wild-type lacZ under lacI control - the strain and the corresponding samples latter called LacZ;

(2) a pBR322-derived plasmid containing a pseudoknot-lacZ (PK-lacZ) consisting of gene10, a virus-derived RNA pseudoknot (Tholstrup et al, 2012), 22/6a, fused upstream of the lacZ - the strain and the corresponding samples latter called PK-LacZ.

Bacteria were grown in MOPS minimal medium for 10 generations or more at 37°C to ensure stable exponential growth.

Libraries were generated under the normal growth - before the induction - and following the induction of the LacZ and PK-lacZ expression. The expression was induced with isopropyl b-D-1-thiogalactopyranoside (IPTG). cDNA library for Ribo-seq and total RNA library for RNA-seq have been generated following the previously described protocol Bartholomaus et al (2016). In addition, total RNA was spiked in with RNA standards in predefined concentrations (ERCC RNA Spike-In Mix; Ambion), which were used to determine the threshold of detection for each sample and latter to calculate the copy numbers of transcripts per cell (see below). All the libraries were sequenced on a HiSeq2000 (Illumina) machine.

5.2.2 Preprocessing and mapping of sequencing data

RNA-seq and Ribo-seq datasets were preprocessed and aligned to the reference genome following the algorithm described above. In brief, reads were quality trimmed using fastx-toolkit version 0.0.13.2 (quality threshold: 20), sequencing adapters were removed with cutadapt version 1.8.3 (minimal overlap: 1 nt).

The reference genome used for the mapping (E. coli K-12 MG1655 strain) has been modified by masking the genomic regions that were similar to those in the plasmids (such as *LacZ* gene and other parts of *LacZ* operon) and adding the plasmid sequences and ERCC spike-in sequences to the annotation. The preprocessed reads were uniquely (-m 1) mapped to the modified version of genome using Bowtie (version 1.1.2), allowing up to 2 mismatches. Reads aligning to more than one region including tRNA and rRNA were excluded from the data.

The raw reads were used to generate read counts per gene using CoverageBed function from samtools, the initial E.coli annotation file has been modified by adding plasmids and ERCC spike-ins. All the reads have been normalized to RPM and RPKM (See Chapter 2-4).

Detection threshold in RPKM has been set using custom script written in R at values with a linear correlation between the reads aligned from the spike-in controls and their concentration in the mixture for each RNA-seq dataset separately (Fig.5.1). The same detection limits have been applied to the corresponding Ribo-seq samples.

All the sequencing reactions have been performed in 2 replicates. Based on the high correlation between the replicates (Fig.5.2), reads from both biological replicates were merged together to form metagene sets as described Ingolia et al (2009), which have been used for the downstream analysis.

5.2.3 Processing pipeline and downstream statistical analysis

The downstream data analysis was performed using custom scripts in R (version 3.4.4) using RStudio environment (version 1.1.4). Core R functions and statistical packages have been used to perform the statistical tests.

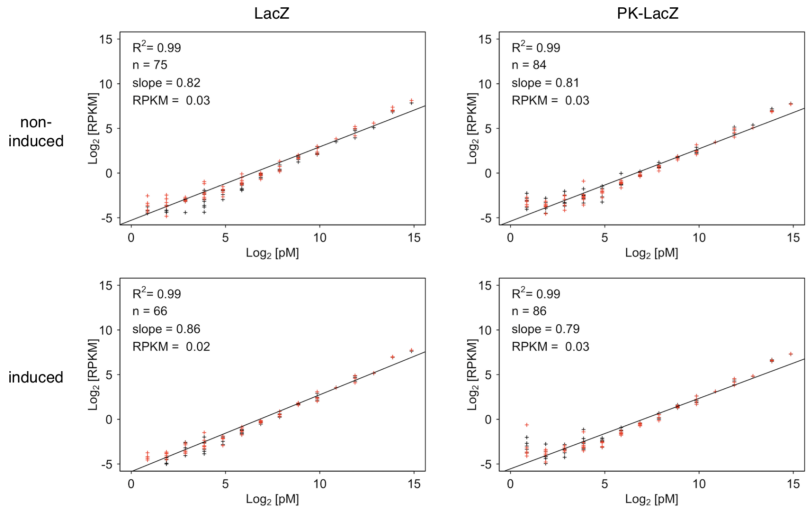

Figure 5.1: Expression of the RNA spike-in standards in the RNA-Seq libraries. Each point represents a single RNA from the spike-in mix. Each of the biological replicates are shown in red and black, respectively. Expression of each spike-in RNA is given in RPKM; "n" denotes the number of RNA standards with linear dependence of their concentration in the spike-in mixture (slope); R^2, Pearson correlation coefficient.

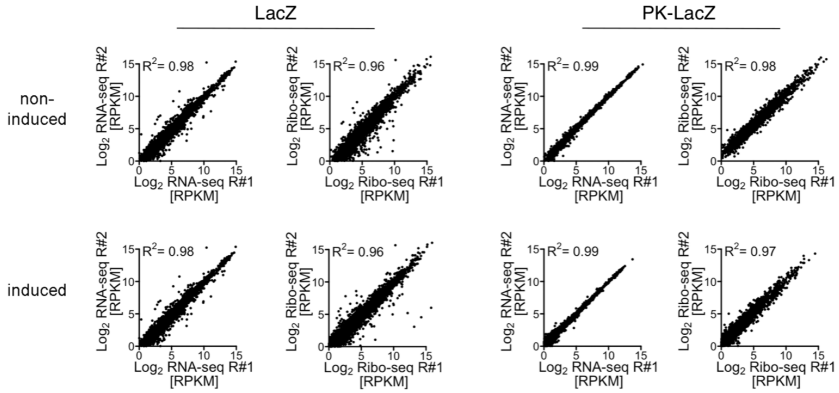

Figure 5.2: Correlation of the RNA-Seq and Ribo-Seq data of two biological replicates from induced and non-induced cells expressing LacZ or LacZ-PK; R^2, Pearson correlation coefficient.

Differential gene expression on transcriptional (RNA-seq) and translational levels was performed using DESeq2 package (version 1.20). While the transcriptional changes have been accessed using RNA-seq data solely, for the translational differences, the translation efficiency values have been compared following the guidelines from DESeq2 manual, which suggests to consider both Ribo-seq and corresponding RNA-seq sample together.

First, the level of significance P = 0.01 have been considered for both comparisons - in translational efficiency and mRNA expression. The genes with P< 0.01 have been selected as differently expressed. Then, P-values were adjusted for multiple testing using false-discovery rate (FDR) according to Benjamini and Hochberg. Since all of the RNA-seq datasets had very high correlation and reproducibility - R^2>0.99, the more restrictive threshold has been applied - P < 0.001 and then we additionally selected the 25^{th} percentile of the most differently expressed genes from the resulting gene-list.

GO enrichment analysis was performed on the differentially expressed gene lists using the bioinformatics tools and reference E.coli genome from Gene Ontology Consortium (http://geneontology.org/). GO-terms with significant enrichment (P < 0.01) have been selected.

In this study, we also used RNA-seq data and the spike-in controls within each RNA-seq sample for the calculation of the absolute transcript copy numbers per cell. We applied a method previously described by Bartholomaus et al (2016) and Mortazavi et al (2008). In brief, the mapped reads for each transcript were related to the total reads per sample (sequencing depth) and the length of the transcriptome, which has been determined using the molecules of all the spike-in standards above the detection threshold, and finally the value was normalized by the number of cells.

To obtain the precise single-nucleotide resolution from the Ribo-seq data, which is especially crucial for the assessment of the initiation and termination rates, the position of each read have to be adjusted in order to correspond to either P-site or A-site of the ribosome. The central nt of the read cannot be taken blindly into the further

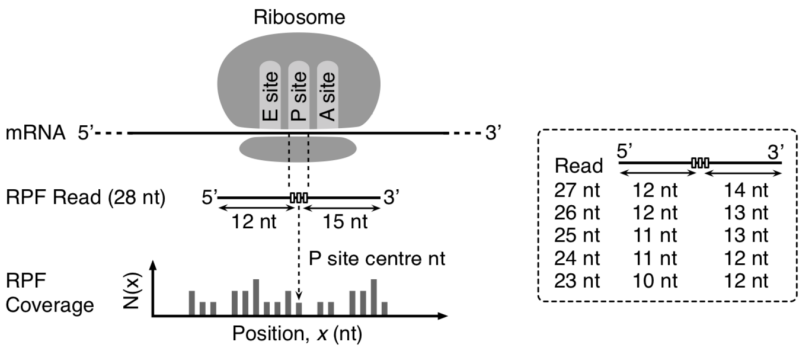

Figure 5.3: Algorithm of estimating the ribosome P-site position from an RPF read. Box shows different lengths used from 5′ and 3′-end of various RPF read length used to calculate position of central nt in the P-site codon.

processing, since each of the samples contain the broad range of the read lengths (~22-30nt) and as it has been shown previously, the offsets from the 5′ and 3′ ends of each read vary greatly.

To overcome this variability and precisely allocate the reads, the following algorithm have been applied:

First, RPFs were binned in the groups of equal read length, and each group was aligned at the stop codons as previously described by Mohammad et al (2016). For each read length, the distance between the point where the transcript leaves the ribosome and the middle nucleotide at the P-site had been calculated. This distance was used to determine a center of each P-site codon along each mRNA (Fig.5.3) and calibrate the reads. As expected, the majority of sequencing reads were 23–28 nt, which corresponds to the size of the prokaryotic ribosome, so these read lengths were selected for the further analysis.

The ribosome occupancy per each codon over the whole transcriptome was calculated as described by Lareau et al (2014). In order to overcome the bias raised by differences in expression levels between the genes, the reads for each position within a gene were normalised to the average number of reads for this gene.

Metagene analysis of the ribosome occupancies of the START and STOP codon regions was performed as described by Baggett et al (2017). Overlapping genes were

excluded from this analysis, as the reads in the overlapping regions cannot be unambiguously referred to one of the genes. Only the genes, which had at least 5 reads in the chosen window have been considered.

5.2.4 Data access

As a part of the publishing process, deep-sequencing data from RNA-seq and Ribo-seq were deposited in the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra/) under accession number SRP144594.

5.3 Results

5.3.1 Characterizing a synthetic pseudoknot that induces frameshifting

Proteins are translated by reading tri-nucleotides (=codons) on the mRNA strand, from 5′ end of the mRNA to the 3′ end, when each codon is translated into amino acid. A shift of nucleotides not divisible by 3 in the reading frame will result in the different codons to be read and another protein product to be produced. This change in the reading frame is known as translational recoding, also called translational frameshifting.

Pseudoknots (PKs) are among the most prevalent folding motifs of RNA. These stable nucleic acid secondary structures regulate gene expression. In combination with slippery sequences they stimulate translational recoding/frameshifting in viral genomes making them compact and allowing to produce several protein products from a single gene (Tsuchihashi & Kornberg, 1990; Sharma et al, 2014).

PKs are the most common type of structure used to induce frameshifting, typically to -1 frame (Atkins et al, 2016), but in rare cases they cause +1 frameshifting (Ivanov et al, 2004). PK typically consists of a hairpin with an extra loop that folds back to stabilize the hairpin through additional base pairing (Fig.5.4). In addition to inducing frameshifting, PKs can also regulate translation initiation, via obstructing a ribosome-binding site (RBS) with antisense sequences that base pair with it (Bordeau & Felden, 2014).

In order for the recoding event to occur, two elements are required. The first is a slippery site consisting of a sequence in a form of XXXYYYZ (where X, Y, Z are nucleotides). This site enables base paring in the A or P site of the ribosome outside of the normal reading frame (zero frame), stimulating recoding events. The second element is a PK, located 6–8 nt downstream of the slippery site. This distance between the slippery site and the PK provides an extended time for frameshifting to happen (Giedroc & Cornish, 2009).

In this study, to assess the process of frameshifting and the affect of the PK on the host, we created an inducible genetic construct (further referred to as PK-LacZ). We incorporated a virus-inspired PK structure within the natural context (gene10 of bacteriophage T7) integrated to bacterial lacZ gene in a -1 frame (Fig.5.4) (Tholstrup et al, 2012). Gene10 produces two proteins of bacteriophage capsid: first as a result of translation in the natural zero-frame and another one - through a -1 frameshift. Gene10 ends with a stop codon in a way that translation of a downstream-located lacZ gene requires frameshifting event to occur at the PK. A slippery site (UUUAAAG) anticipated the PK.

We have chosen a PK variant (22/6a) with a lower frameshifting efficiency (~3%) (Tholstrup et al, 2012) compared to the wild-type PK (~10% frameshifting) in the natural context, as the latter one is known to heavily stall ribosomes in the cell and induce a significant stress response of a host (Tholstrup et al, 2012). We were seeking to perform a quantification of the frameshifting efficiency in our construct, but more importantly to discover the reason of such a significant cellular stress observed.

Using the Ribo-Seq data before and after the induction of the PK expression, we generated translation profiles to assess ribosome flux along the entire construct (Fig.5.5). We observed high translational levels until the position of the PK with a major drop of 80–90% at the PK itself to the end of the gene10, and a further drop of ~97% downstream of this region (Fig.5.5).

Next, we analyzed the frameshifting of gene10. The construct has been divided into three regions:

(1) the gene10 segment up to the slippery site;

(2) the middle region (including the slippery site, the PK and until the gene10 stop codon);

(3) the downstream lacZ gene in a -1 frame.

Although, Ribo-Seq is a precise method reporting on the positions of the translating ribosomes, the single-codon or single-nucleotide resolution, which is required to assess the frameshifting, becomes challenging in a context when the expression levels

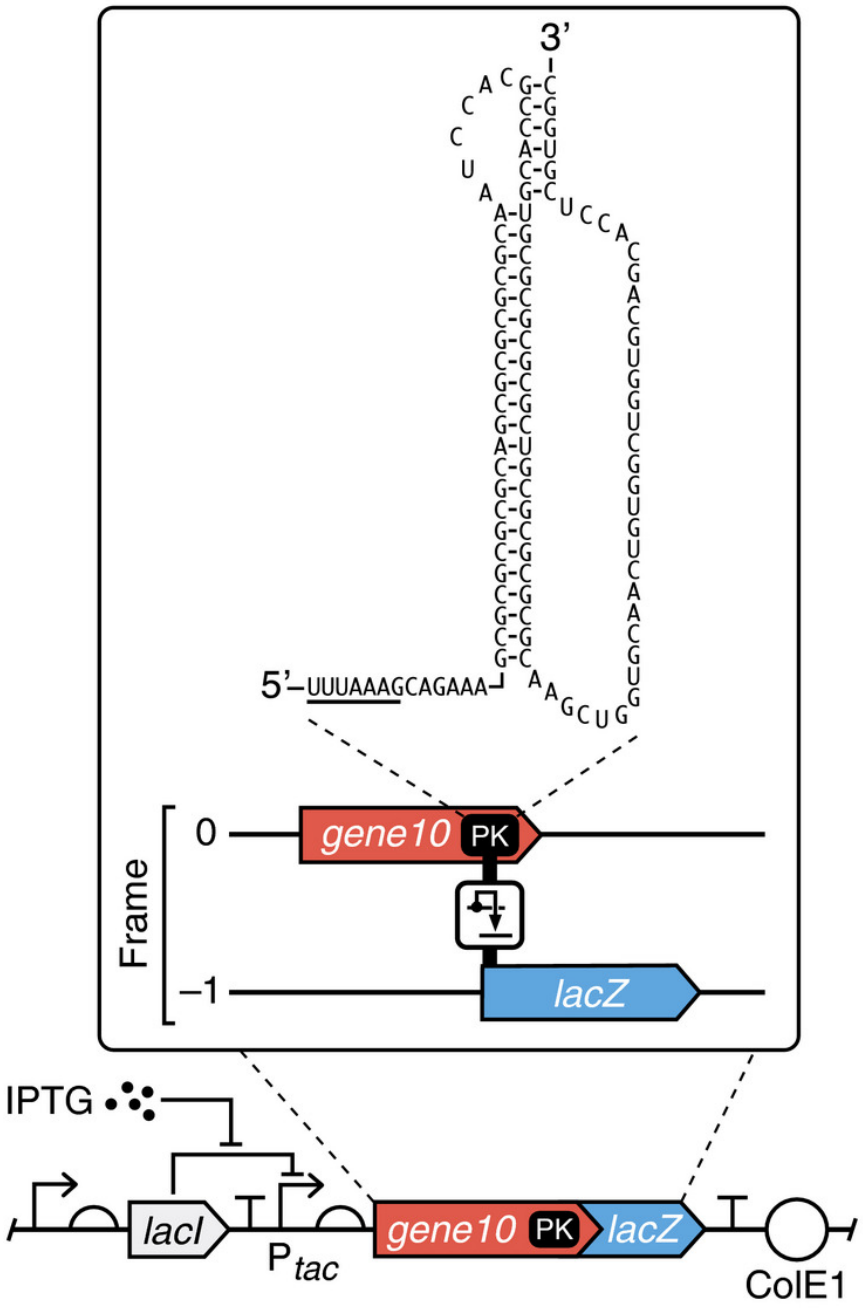

Figure 5.4: Genetic design of the PK-LacZ construct: the PK secondary structure, the slippery site (underlined), gene10 and lacZ, which are in the differing reading frames.

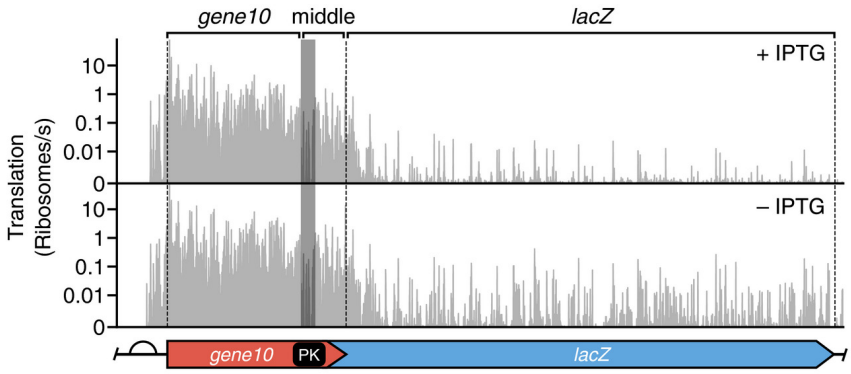

Figure 5.5: Translation profiles for the PK-LacZ construct before (bottom) and after the induction (top) with IPTG (1 mM). The gene10 (1), middle (2), and lacZ (3) regions are labeled; shaded region denotes the PK, dashed lines denote the start and stop codons of gene10 and LacZ.

are low because of the noise. In our case, significant drops in translation at the PK and gene10 stop codon led to the low numbers of ribosome-footprints (RPFs) along the last (3) region - lacZ gene. Therefore, the direct comparison of frame-specific expression patterns was impossible.

To overcome this problem, we pooled together all the RPFs within each of the regions (1)-(3) and calculated the fraction of RPFs in each frame (-1, 0, +1) from a total of three possible frames. We detected the zero and -1 frames dominating in the gene10 (1) and lacZ (3) regions, respectively, with > 46% of all RPFs being found in these frames (Fig.5.6, top). The middle region (2) showed a mixture of all three frames, while the zero-frame further dropped in the lacZ region. Most likely, this illustrated a combination of ribosomes that have successfully passed the PK and terminated in zero-frame at the stop codon of gene10 and those ribosomes that have frameshifted. We obtained similar distribution of the frames for both conditions - before and after the induction by IPTG (Fig.5.6). To verify that the reading frames observed in Ribo-Seq were reliable and no sequencing bias was introduced, we performed the same analysis in RNA-Seq datasets (Fig.5.6, bottom). As expected, no specific frames were prevalent in any of the regions, RNA-Seq before and after induction showed equal fractions for each of the frames.

Next, we examined the possibility to recover the major translation frame from the

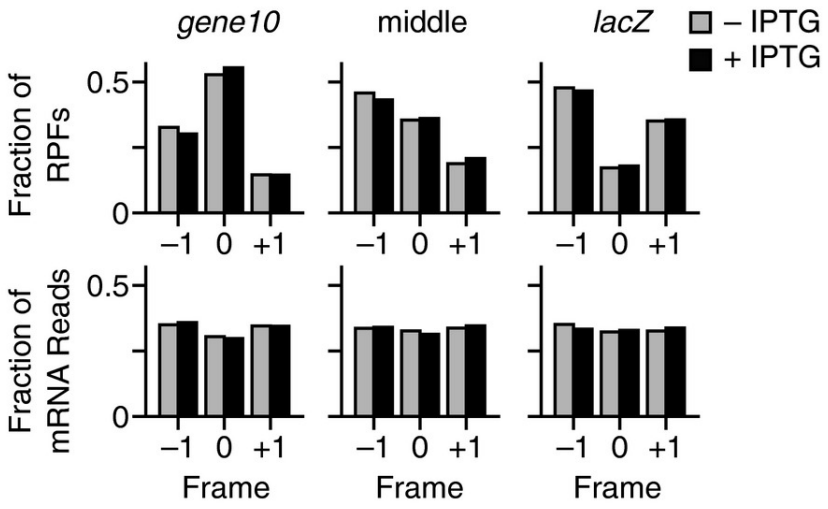

Figure 5.6: Fractions of the total RPF (top) and mRNA (bottom) reads in each reading frame for the gene10 (1), middle (2), and lacZ (3) regions, before and after the induction of expression with IPTG.

entire genome based on the Ribo-Seq data. In order to assess this, we measured the fraction of each of three frames for every gene. The natural zero-frame dominated over the other two frames (Fig.5.7, top), while all the frames were equally distributed when similar analysis was performed on RNA-Seq samples (Fig.5.7, bottom).

Lastly, we calculated the efficiency of frameshifting induced by the PK, by comparing the density of RPFs per nucleotide for the middle (2) and lacZ (3) regions, before and after frameshift, respectively. Using an equation described in the original publication (See the original manuscript, equation (7)), we found that, in our case, PK caused 2–3% of frameshifting. This result matched the previous measurement of 3% frameshift for the same PK construct (22/6a) (Tholstrup et al, 2012).

5.3.2 Cellular response to a strong synthetic pseudoknot

It has been shown previously, that expression of strong PKs severely impact the cell growth, however, the reason behind this remained unexplored (Tholstrup et al, 2012). In our Ribo-Seq data, we observed a large number of RPFs in the gene10 region (Fig.5.5). These reads could be raised from premature termination of ribosomes or

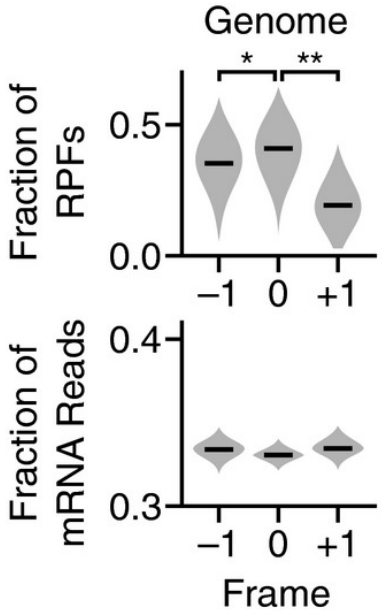

Figure 5.7: Violin plots of the distributions of fractions of total RPFs and mRNA reads in each of the reading frames for all transcripts in E. coli genome. Median values shown by horizontal bars. *P = 0.049; **P = 1.6×10^{-9} (Mann–Whitney U test).

stalled translation at the PK itself. Previous characterization of the PK (22/6a) used in this study showed that it sequestered ribosomes (Tholstrup et al, 2012), therefore, it is likely that many of reads, we observed on the profile (Fig.5.5), illustrate stalled ribosomes. Besides of the increased amount of partially synthesized protein products, stalling also limits the availability of translational resources in the cell.

Considering these observation, we asked whether the expression of the PK-LacZ construct causes a cellular stress by sequestering ribosomes. Along with PK-LacZ construct used in the study, we introduced another dataset, carrying a LacZ plasmid, which does not contain a PK, and, therefore, lacZ itself does not induce any of the specific PK effects.

Next, we compared the burden that expression of lacZ and PK-lacZ caused in the host cell in each of the cases. In order to do that, we compared the shifts in transcription - based on RNA-Seq data - and translation efficiency, e.g. density of the ribosomes per mRNA, - based on both, Ribo-Seq and RNA-Seq data (See the Materials and Methods for the precise calculation). All genes in the genome have been used to calculate the

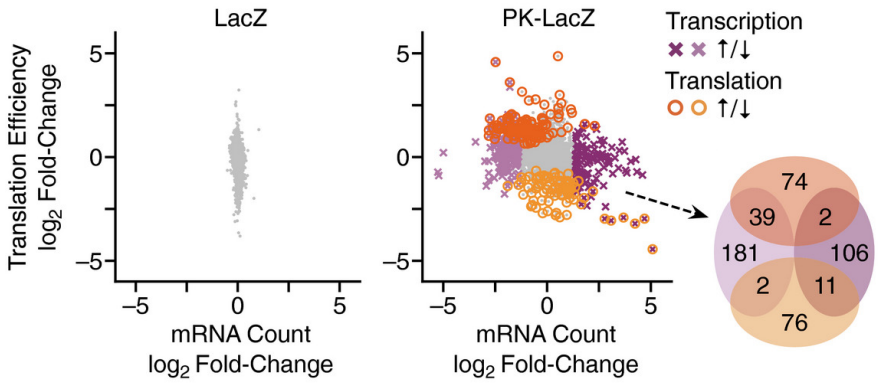

Figure 5.8: Change in expression of E. coli genes following induction of PK-lacZ expression. Each point denotes a transcript. Differentially expressed genes are highlighted in color and by an alternative point shape (transcriptional regulation: purple cross; translational regulation: orange open circle). Right - Venn diagram of genes significantly regulated transcriptionally and translationally after induction of the PK-LacZ expression.

fold-changes of RPKM values between the non-induced and IPTG-induced samples (Fig.5.8).

We have not observed any significant changes before and after the induction of the LacZ construct (Fig.5.8). In opposition to that, the expression of PK-LacZ construct led to significant shifts in the expression of 491 genes. From these genes, 341 were transcriptionally (i.e., having significant changes in RNA-Seq) and 204 translationally regulated (i.e., with significant changes in translational efficiency), containing a little overlap of 54 genes between these two types of regulation (Fig.5.8, right).

Most of the genes being differentially expressed on the transcriptional level, showed a drop in their mRNA counts (e.g. transcriptionally down-regulated). On the other hand, translationally regulated genes were split between two groups showing either an increased or a decreased translational efficiency. Gene ontology (GO) enrichment analysis revealed that transcriptionally down-regulated genes fall into categories mostly associated with translation, such as ribosomal proteins, amino acid biosynthesis, amino acid activation, also containing some genes involved in respiration and catabolism. Transcriptionally up-regulated genes were linked to ATP binding, also included chaperones (ftsH, lon, clpB, dnaJK, groLS, htpG), ion binding, proteolytic

activities (ftsH, prlC, htpX), and an endoribonuclease (ybeY).

The expression of all of these fell under r32 regulation, which is the most common regulatory mode of response to the heat stress. r32 up-regulation is frequently observed by expressing synthetic constructs, although the precise mechanism of r32 activation remains unclear (Ceroni et al, 2018). In the case of synthetic PK we introduced, the peptides, which were incomplete, because of the stalled ribosomes on the PK-LacZ, were most likely mis-folded and therefore generated mis-folding stress similar to that in the heat shock response. Binding of the major E. coli chaperones (DnaK/DnaJ and GroEL/S) to the mis-folded proteins negatively regulates the expression of r32. The shift of these chaperones to mis-folded proteins releases r32, which induces the expression of heat shock genes (Guisbert et al, 2004). This idea is supported by the fact that dnaJ, groL/S, and grpE were transcriptionally up-regulated following the PK induction as well as ftsH gene, which encodes the protease that is responsible for the degradation of r32.

Next, we examined whether the expression of PK-LacZ construct caused changes in translation dynamics, such as ribosome pausing at specific codons. Using Ribo-Seq datasets we computed the ribosomal occupancy at each codon (also called codon occupancy) across the genome and compared it between the two conditions - before and after the induction of PK-lacZ expression (Lareau et al, 2014). We observed an increased occupancies of the few codons: AGA, CTA, CCC, and TCC, encoding Arginine, Leucine, Proline, and Serine, respectively (Fig.5.9). All of these codons are rarely used in the genome, e.g. having a low genomic frequency in E.Coli, but all of them were found in higher proportions across gene10. Along with a high expression level of gene10, the stress, which is induced by this atypical demand on resources, would be additionally boosted. Altogether, the broad shifts in regulation at a cellular level and the changes in codon occupancies propose that PK-LacZ expression significantly limits the availability of cellular resources.

We further compared the transcriptome composition (from RNA-Seq) and distribution of the ribosomes across cellular transcripts and plasmids (from Ribo-Seq) before

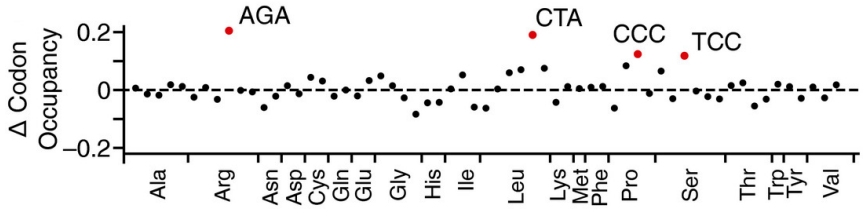

Figure 5.9: Change in codon occupancies for cells with PK-LacZ construct after induction, calculated from the Ribo-seq data. Each point corresponds to a codon, which are ordered by amino acid and by abundance in the genome. Dashed horizontal line denotes no changes. Outliers are labeled and highlighted in red (Tukey test: 1.5 times the interquartile range below the first quartile or above the third quartile).

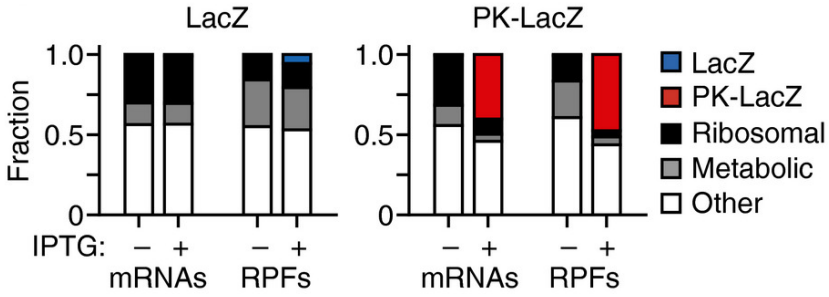

Figure 5.10: Fractions of mRNA and RPF reads mapped to each of the synthetic expression constructs (LacZ and PK-LacZ) and genomic E. coli transcripts (divided into three categories: ribosomal, metabolic, and other functions), before and after the induction with IPTG.

and after the induction. This analysis revealed that following the induction, the PK-LacZ construct represented 40% of all mRNA transcripts produced in the cell and captured 47% of the whole ribosome pool engaged in translation (Fig.5.10). This illustrates the global drop in translation and mis-folding stress induced by the partially translated proteins from gene10 transcripts and explains the r32-mediated response in the case of PK-LacZ induction.

From the other side, no notable changes in the distribution of the mRNA or ribosome pools have been observed after the induction of LacZ plasmid itself. In addition, we also noted a large difference in the number of transcripts for each construct after induction: the number of LacZ transcripts was 43-fold lower than those for PK-LacZ (81 versus 3,504 transcripts/cell, respectively). This difference cannot be caused solely by the increased transcription initiation at the promoter in the PK-LacZ construct.

One of the potential explanations has been described in the previous studies: the decay rate of the lacZ is highly dependent on the proportion of its transcription and translation rates (Yarchuk et al, 1992). RNase E sites within the coding region of LacZ become accessible to cleavage by RNase E in the case of the low translation initiation rate, since fewer translating ribosomes are present to prevent degradation (Yarchuk et al, 1992). This mechanism could reason the low lacZ transcript numbers, which would lead to the reduced number of ribosomes sequestered for translating lacZ in this construct. Altogether, this explain the lack of a stress response for this construct compared to the PK-LacZ.

5.4 Conclusions

In this study, we assessed the behavior of a genetic construct that contains a strong virus-inspired PK structure that induces a translational frameshift (Fig.5.4). We demonstrated the ability to quantitatively assess various transcriptional and translational processes using a combination of deep sequencing technologies.

Following the induction of expression of PK-LacZ construct, the main reading frame shifts with the same efficiency as measured in previous studies for the same PK using another method (Tholstrup et al, 2012). PK-lacZ also causes a major burden to the cell, sequestering a large portion of the shared ribosome pool from the host cell, in contrast to the LacZ construct, which does not cause this effect (Fig.5.10).

Also, we observed transcriptome-wide increase in ribosome occupancies of the codons, which are rarely present in endogenous E. coli genes, but more frequently occur in the synthetic PK-LacZ construct. This suggests that the strong expression of the construct leads to significant demands on the translational resources of the cell. This burden also resulted in significant changes in gene regulation on both transcriptional and translational levels. This response was mediated by the alternative polymerase subunit - r32 - that has been shown to remodel the bacterial protein synthesis under the thermal stress (Guo & Gross, 2014). In our case, r32 activation is most likely

caused by a combination of strong over-expression of gene10 and mis-folding stress raised from incompletely synthesized peptides (Guo & Gross, 2014).

The stress response induced by a strong pseudoknot has not been reported and described previously, which suggests the novelty of our study where we revealed a new branch of research for the future exploration.

Chapter 6

Conclusions

The chapters 3 to 5 of the current thesis are based on the three different studies. All of them used a combination of the various deep sequencing approaches, which served as a crucial component for the successful research outcome. Those results would be impossible to achieve without a specific bioinformatics analysis pipeline developed uniquely for each of these studies.

I have shown that the correct interplay of the different types of sequencing datasets, such as RNA-Seq, Ribo-Seq, PAR-CLIP-Seq and even more specific MeRIP-Seq, can reveal multiple features of the cell biology and uncover the hidden aspects of the processes taking place in the cell under the changing conditions.

I demonstrated how the integration of the analysis of the deep sequencing data into the other experiments can serve as an additional evidence for the findings and unveil the new hidden details on a single-gene level due to the high depth, specificity and precision of the method compared to the other experimental technics.

Most importantly, the bioinformatics component in each of the studies included an appropriate selection, usage and adaptation of variety of existing tools in combination with self-written scripts and pipelines. This part should not be underestimated, since the depth of the information hidden in the sequencing data can be revealed only through the algorithms, which in majority of the cases cannot be unified.

To conclude, the big data analysis is a necessary part of a cutting-edge research process and not only a technical service provided by bioinformaticians to the other scientists as often assumed.

Acknowledgements

First, I am extremely grateful to Prof. Andrew Torda, who, initially being my second supervisor, did not refuse me and supported my desire to complete PhD.

I would like to thank to my current research group - Oxford Vaccine Group, Department of Paediatrics, University of Oxford, for their warm welcome, valuable discussions and appreciation of my achievements and skills. Particularly, I am thankful to Dr. Daniel O'Connor, who believed in my abilities and promoted my candidature after the interview, so Prof. Andrew Pollard offered me this PostDoctoral research position in University of Oxford even before I finished PhD.

Besides, I would like to thank my previous academic supervisor, Prof. Zoya Ignatova, with whom I have been working since my Master thesis and majorly during my PhD. With her, I went through the challenging life and professional circumstances, and left them with dignity. Thanks to her I grew up as a scientist and as a person during these years, having gained a lot of lessons and experience, which would be crucially helpful in my further scientific career. I thank the lab members of AG Ignatova for the ability to remain sincere, supportive and attempt to be independent of the imposed opinion.

My special thank you to Dr. Alexander Laatsch; without his timely advices on the good scientific practice I would not be able to finish and submit my thesis. I would like to thank Mrs. Waltraud Wallenius from Studienburo of Chemistry Department for her kind support and assistance in my complicated situation. Also, I am thankful to Prof. Chris Meier, the Head of the Doctoral committee, for his understanding and fair assessment of my work.

I am grateful to all the scientific collaborators I had a chance to work with. Two chapters of this thesis are based on the results obtained via collaborations with Prof.

Cecília Arraiano and her lab members, from Universidade Nova de Lisboa, Portugal; and Dr. Thomas Gorochowski from University of Bristol, the United Kingdom. I was involved in many more collaborative projects during my PhD, and even so the results are not included to the final version of the thesis, I would like to particularly thank Prof. Alexander Mankin and the members of his lab for the productive discussions.

Furthermore, I would like to thank the University of Hamburg for the scholarship, which allowed me to start my PhD here. I am grateful to Studierendenwerk for the support during the tough periods, for the childcare for my son over the weekends and for the kindly provided flat in the dormitory. I would like to thank my recent funding program, SPP 2002 from DFG, which provided me with another year to finish my thesis.

I am grateful to the former Head of the Russian National Research Medical University, where I obtained my Diploma, Prof. Dr. Andrey Kamkin, who helped me to believe that science has no boundaries. I would like to express my gratitude to Assistant Professor Natalia Popova, who always made high demands on me and thus instilled a deep interest in biology.

I wish to thank my first primary school teacher, Natalia Solovieva, who was always especially strict but noticed my potential and proposed me an academic career when I was only 7 years old.

Importantly, I am immensely grateful to Dr. Ulla Döhnert from AKK, who made me believe that even incurable diagnosis is not a death sentence, but a new stage.

I would like to thank my ex-neighbours and wonderful friends, Yulia, Waldemar and Mamoun, for the late evening discussions over tea, coffee and vine.

Finally, very kind and deepest thank you to my family: to my parents who regularly helped me to take care of my old son, and especially to my parents in law, who raised my small son during the whole period of my Doctoral studies. Thank you for being my biggest supporters. I would like to thank my children for their patience and understanding.

References

Anders M, Chelysheva I, Goebel I, Trenkner T, Zhou J, Mao Y, Verzini S, Qian SB, & Ignatova Z. (2018). Dynamic m^6A methylation facilitates mRNA triaging to stress granules. Life Science Alliance, 1(4), e201800113

Anders S, Pyl PT, & Huber W. (2015). HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics (Oxford, England), 31(2), 166–169

Andrade JM, dos Santos RF, Chelysheva I, Ignatova Z, & Arraiano CM. (2018). The RNA-binding protein Hfq is important for ribosome biogenesis and affects translation fidelity. The EMBO Journal, e97631

Andreev DE, O'Connor PB, Fahey C, Kenny EM, Terenin IM, Dmitriev SE, Cormican P, Morris DW, Shatsky IN, Baranov PV. (2015). Translation of 5' leaders is pervasive in genes resistant to eIF2 repression. Elife 4: e03971

Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov PV. (2016). Ribosomal frameshifting and transcriptional slippage: from genetic steganography and cryptography to adventitious use. Nucleic Acids Res 44: 7007–7078

Baens M, Noels H, Broeckx V, Hagens S, Fevery S, Billiau AD, Vankelecom H, Marynen P. (2006). The dark side of EGFP: defective polyubiquitination. PLoS ONE 1:e54

Baggett NE, Zhang Y, & Gross C. (2017). Global analysis of translation termination in E. coli. PLoS Genetics, 13(3), e1006676

Bartholomäus A, Fedyunin I, Feist P, Sin C, Zhang G, Valleriani A, Ignatova Z. (2016). Bacteria differently regulate mRNA abundance to specifically respond to various stresses. Philos Trans R Soc A Math Phys Eng Sci 374: 20150069

Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, Moradian A, Hess S, Kinney JB, Phillips R. (2018). Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. Proc Natl Acad Sci USA 115:E4796 – E4805

Bordeau V, Felden B. (2014). Curli synthesis and biofilm formation in enteric bacteria are controlled by a dynamic small RNA module made up of a pseudoknot assisted by an RNA chaperone. Nucleic Acids Res 42: 4682–4696

Cardinale S, Joachimiak MP, Arkin AP. (2013). Effects of genetic variation on the E. coli host-circuit interface. Cell Rep 4: 231–237

Cekaite L, Peng Q, Reiner A, Shahzidi S, Tveito S, Furre IE, Hovig E. (2007). Mapping of oxidative stress responses of human tumor cells following photodynamic therapy using hexaminolevulinate. BMC Genomics 8: 273

Ceroni F, Furini S, Gorochowski TE, Boo A, Borkowski O, Ladak YN, Awan AR, Gilbert C, Stan G-B, Ellis T. (2018). Burden-driven feedback control of gene expression. Nat Methods 15: 387–393

Chen SS, Sperling E, Silverman JM, Davis JH, & Williamson JR. (2012). Measuring the dynamics of E. coli ribosome biogenesis using pulse-labeling and quantitative mass spectrometry. Molecular bioSystems, 8(12), 3325–3334

Cock PJ, Fields CJ, Goto N, Heuer ML, & Rice PM. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research, 38(6), 1767–1771

Crick FH. (1958). On protein synthesis. Symposia of the Society for Experimental Biology, 12, 138–163

Damgaard CK, Lykke-Andersen J. (2011). Translational coregulation of 5′ TOP mRNAs by TIA-1 and TIAR. Genes Dev 25: 2057–2068

Davis JH, Williamson JR (2017) Structure and dynamics of bacterial ribosome biogenesis. Philos Trans R Soc Lond B Biol Sci 372: 20160181

Del Campo C, Bartholomaus A, Fedyunin I, & Ignatova Z. (2015). Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. PLoS Genetics, 11(10), e1005613

Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, et al. (2012). Topology of the human and mouse m^6A RNA methylomes revealed by m^6A-seq. Nature 485: 201–206

Edupuganti RR, Geiger S, Lindeboom RGH, Shi H, Hsu PJ, Lu Z, Wang SY, Baltissen MPA, Jansen P, Rossa M, et al. (2017). N^6-methyladenosine (m^6A) recruits and repels proteins to regulate mRNA homeostasis. Nat Struct Mol Biol 24:870–878

Ewels P, Magnusson M, Lundin S, & Kaeller M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England), 32(19), 3047–3048

Gerashchenko MV, Lobanov AV, Gladyshev VN. (2012). Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc Natl Acad Sci 109: 17394–17399

Giedroc DP, Cornish PV. (2009). Frameshifting RNA pseudoknots: structure and mechanism. Virus Res 139: 193 – 208

Gilks N, Kedersha N, Ayodele M, Shen L, Stoecklin G, Dember LM, Anderson P. (2004). Stress granule assembly is mediated by prion-like aggregation of TIA-1. Mol Biol Cel 15: 5383–5398

Gorochowski TE, Chelysheva I, Eriksen M, Nair P, Pedersen S, & Ignatova Z. (2019). Absolute quantification of translational regulation and burden using combined sequencing approaches. Molecular Systems Biology, 15(5), e8719

Grozhik AV, Linder B, Olarerin-George AO, Jaffrey SR. (2017). Mapping m^6A at individual-nucleotide resolution using crosslinking and immunoprecipitation (miCLIP). Methods Mol Biol. 1562:55–78

Guisbert E, Herman C, Lu CZ, Gross CA. (2004). A chaperone network controls the heat shock response in E. coli. Genes Dev 18: 2812–2821

Guo MS, Gross CA. (2014). Stress-induced remodeling of the bacterial proteome. Curr Biol 24:R424 –R434

Hafner M, Landthaler M, Burger L, KhorshidM, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp AC, Munschauer M, et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 141: 129–141

Hatem A, Bozdag D, & Çatalyurek UV. (2011). Benchmarking short sequence mapping tools. Proceedings - 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2011, 109–113

Hecht A, Glasgow J, Jaschke PR, Bawazer LA, Munson MS, Cochran JR, Endy D, Salit M. (2017). Measurements of translation initiation from all 64 codons in E. coli. Nucleic Acids Res 45: 3615–3626

Hwang J-Y, Buskirk AR. (2017). A ribosome profiling study of mRNA cleavage by the endonuclease RelE. Nucleic Acids Res 45: 327–336

Ingolia NT, Ghaemmaghami S, Newman JR, & Weissman JS. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science (New York, N.Y.), 324(5924), 218–223

Ivanov IP, Anderson CB, Gesteland RF, Atkins JF. (2004). Identification of a new antizyme mRNA +1 frameshifting stimulatory pseudoknot in a subset of diverse invertebrates and its apparent absence in intermediate species. J Mol Biol 339: 495–504

Jinks-Robertson S, Nomura M. (1981). Regulation of ribosomal protein synthesis in an Escherichia coli mutant missing ribosomal protein L1. J Bacteriol 145: 1445–1447

Jones DL, Brewster RC, Phillips R. (2014). Promoter architecture dictates cell-to-cell variability in gene expression. Science 346: 1533–1536

Jones PG, Inouye M. (1996). RbfA, a 30S ribosomal binding factor, is a cold-shock protein whose absence triggers the cold-shock response. Mol Microbiol 21: 1207–1218

Kedersha N, Ivanov P, Anderson P. (2013). Stress granules and cell signaling: More than just a passing phase? Trends Biochem Sci 38: 494–506

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, & Haussler D. (2002). The human genome browser at UCSC. Genome research, 12(6), 996–1006

Khong A, Matheny T, Jain S, Mitchell SF, Wheeler JR, Parker R. (2017). The stress granule transcriptome reveals principles of mRNA accumulation in stress granules. Mol Cell 68: 808–820 e805

Koboldt DC, Steinberg KM, Larson DE, Wilson RK, & Mardis ER. (2013). The next-generation sequencing revolution and its impact on genomics. Cell, 155(1), 27–38

Langmead B. (2010). Aligning short sequencing reads with Bowtie. Current protocols in bioinformatics, Chapter 11, Unit–11.7

Lareau LF, Hite DH, Hogan GJ, Brown PO. (2014). Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. Elife 3:e01257

Lee S, Liu B, Lee S, Huang SX, Shen B, & Qian SB. (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(37), E2424–E2432

Li GW, Burkhardt D, Gross C, Weissman JS. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell 157: 624–635

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, ... 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England), 25(16), 2078–2079

Link TM, Valentin-Hansen P, Brennan RG. (2009). Structure of Escherichia coli Hfq bound to polyriboadenylate RNA. Proc Natl Acad Sci USA 106: 19292–19297

Liu B, Han Y, Qian SB. (2013). Cotranslational response to proteotoxic stress by elongation pausing of ribosomes. Mol Cell 49: 453–463

Mackowiak SD, Zauber H, Bielow C, Thiel D, Kutz K, Calviello L, ... Obermayer B. (2015). Extensive identification and analysis of conserved small ORFs in animals. Genome biology, 16, 179

Margolin W (2012) The price of tags in protein localization studies. J Bacteriol 194: 6369–6371

Martin M. (2011) Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. EMBnet Journal, 17, 10–12

Mauer J, Luo X, Blanjoie A, Jiao X, Grozhik AV, Patil DP, Linder B, Pickering BF, Vasseur JJ, Chen Q, et al. (2017). Reversible methylation of m^6Am in the 5′ cap controls mRNA stability. Nature 541: 371–375

McGlincy NJ, & Ingolia NT. (2017). Transcriptome-wide measurement of translation by ribosome profiling. Methods (San Diego, Calif.), 126, 112–129

Meyer KD, Patil DP, Zhou J, Zinoviev A, Skabkin MA, Elemento O, Pestova TV, Qian SB, Jaffrey SR. (2015). 5′UTR m^6A promotes cap-independent translation. Cell 163: 999–1010

Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3′UTRs and near stop codons. Cell 149: 1635–1646

Mikulecky PJ, Kaw MK, Brescia CC, Takach JC, Sledjeski DD, Feig AL. (2004). Escherichia coli Hfq has distinct interaction surfaces for DsrA, rpoS and poly(A) RNAs. Nat Struct Mol Biol 11: 1206–1214

Mohammad F, Woolstenhulme CJ, Green R, Buskirk AR. (2016). Clarifying the translational pausing landscape in bacteria by ribosome profiling. Cell Rep 14: 686–694

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628

Motameny S, Wolters S, Nürnberg P, & Schumacher B. (2010). Next Generation Sequencing of miRNAs - Strategies, Resources and Methods. Genes, 1(1), 70–84

Mura C, Randolph PS, Patterson J, & Cozen AE. (2013). Archaeal and eukaryotic homologs of Hfq: A structural and evolutionary perspective on Sm function. RNA biology, 10(4), 636–651

Pan T. (2013). N^6-methyl-adenosine modification in messenger and long non-coding RNA. Trends Biochem Sci 38: 204–209

Pan T. (2018). Modifications and functional genomics of human transfer RNA. Cell research, 28(4), 395–404

Quinlan AR, & Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England), 26(6), 841–842

Ryle AP, Sanger F, Smith LF, Kitai R. (1955). The disulphide bonds of insulin, Biochemical Journal, 60 (4): 541–556

Sauer E, Weichenrieder O. (2011). Structural basis for RNA 30-end recognition by Hfq. Proc Natl Acad Sci USA 108: 13065–13070

Sauer E, Schmidt S, Weichenrieder O. (2012). Small RNA binding to the lateral surface of Hfq hexamers and structural rearrangements upon mRNA target recognition. Proc Natl Acad Sci USA 109: 9396–9401

Scott M, Klumpp S, Mateescu EM, Hwa T. (2014). Emergence of robust growth laws from optimal regulation of ribosome synthesis. Mol Syst Biol 10: 747

Shajani Z, Sykes MT, Williamson JR. (2011). Assembly of bacterial ribosomes. Annu Rev Biochem 80: 501–526

Sharma V, Prère MF, Canal I, Firth AE, Atkins JF, Baranov PV, Fayet O. (2014). Analysis of tetra-and hepta-nucleotides motifs promoting-1 ribosomal frameshifting in Escherichia coli. Nucleic Acids Res 42: 7210–7225

Shi H, Wang X, Lu Z, Zhao BS, Ma H, Hsu PJ, Liu C, He C. (2017). YTHDF3 facilitates translation and decay of N^6-methyladenosine-modified RNA. Cell Res 27: 315–328

Shi H, Wei J, & He C. (2019). Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. Molecular Cell, 74(4), 640–650

Smanski MJ, Zhou H, Claesen J, Shen B, Fischbach MA, Voigt CA. (2016). Synthetic biology to access and expand nature's chemical diversity. Nat Rev Microbiol 14: 135–149

Sonenberg N, Hinnebusch AG. (2009). Regulation of translation initiation in eukaryotes: Mechanisms and biological targets. Cell 136: 731–745

Spitzer J, Hafner M, Landthaler M, Ascano M, Farazi T, Wardle G, ... Tuschl T. (2014). PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins. Methods in enzymology, 539, 113–161

Tholstrup J, Oddershede LB, Sørensen MA. (2012). MRNA pseudoknot structures can act as ribosomal roadblocks. Nucleic Acids Res 40: 303–313

Thurlow B, Davis JH, Leong VF, Moraes T, Williamson JR, Ortega J. (2016). Binding properties of YjeQ (RsgA), RbfA, RimM and Era to assembly intermediates of the 30S subunit. Nucleic Acids Res 44: 9918–9932

Trapnell C, Pachter L, Salzberg SL. (2009). TopHat: Discovering splice junctions with RNA-seq. Bioinformatics 25: 1105–1111

Tsuchihashi Z, Kornberg A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proc Natl Acad Sci USA 87: 2516–2520

Wang L-Z, Wu F, Flores K, Lai Y-C, Wang X. (2016). Build to understand: synthetic approaches to biology. Integr Biol 8: 394 – 408

Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, et al. (2014). N^6-methyladenosine-dependent regulation of messenger RNA stability. Nature 505: 117–120

Wang X, Zhao BS, Roundtree IA, Lu Z, Han D, Ma H, Weng X, Chen K, Shi H, He C. (2015). N^6-methyladenosine modulates messenger RNA translation efficiency. Cell 161: 1388–1399

Wang Y, & Zhao JC. (2016). Update: Mechanisms Underlying N^6-Methyladenosine Modification of Eukaryotic mRNA. Trends in genetics : TIG, 32(12), 763–773

Wilusz CJ, Wilusz J. (2013). Lsm proteins and Hfq: life at the 30 end. RNA Biol 10: 592–601

Xiang Y, Laurent B, Hsu CH, Nachtergaele S, Lu Z, Sheng W, Xu C, Chen H, Ouyang J, Wang S, et al. (2017). RNA m^6A methylation regulates the ultraviolet-induced DNA damage response. Nature 543: 573–576

Yarchuk O, Jacques N, Guillerez J, Dreyfus M. (1992). Interdependence of translation, transcription and mRNA degradation in the lacZ gene. J Mol Biol 226: 581–596

Zaccara S, Ries RJ, & Jaffrey SR. (2019). Reading, writing and erasing mRNA methylation. Nature Reviews Molecular Cell Biology, 20(10), 608–624

Zhang A, Schu DJ, Tjaden BC, Storz G, Gottesman S. (2013). Mutations in interaction surfaces differentially impact E. coli Hfq association with small RNAs and their mRNA targets. J Mol Biol 425: 3678–3697

Zhang SY, Zhang SW, Fan XN, Meng J, Chen Y, Gao SJ, & Huang Y. (2019). Global analysis of N^6-methyladenosine functions and its disease association using deep learning and network-based methods. PLoS computational biology, 15(1), e1006663

Zhou J, Wan J, Gao X, Zhang X, Jaffrey SR, Qian SB. (2015). Dynamic m^6A mRNA methylation directs translational control of heat shock response. Nature 526: 591–594