Understanding regulation of translation through RNA structure and investigating regulatory synthetic long non-coding RNA (slncRNAs)

Roni Cohen

Understanding regulation of translation through RNA structure and investigating regulatory synthetic long non-coding RNA (slncRNAs)

Research Thesis

In partial fulfillment of the requirements for the

Degree of Master of Science in Biotechnology & Food Engineering

Roni Cohen

Submitted to the Senate of the

Technion - Israel Institute of Technology

Adar bet 5779 Haifa April 2019

The research thesis was done under the supervision of Prof. Roee Amit in the Department of Biotechnology and Food Engineering.

The generous financial help of the Technion – Israel Institute of Technology (IIT) is gratefully acknowledged.

Contents

List of Figures

Abstract

For many years, gene expression manipulations were only possible with a handful of characterized promoters and transcription factors. However, recently, we have seen increasingly more RNA-based regulation inspired by natural RNA-based systems. Our genome is extensively transcribed into many species of long non-coding RNA (lncRNA), performing a variety of defined functions, tightly related to the structural versatility of RNA. And while this versatility makes RNA an appealing target for genomic regulation, it holds the biggest challenge of RNA engineering: design of functional synthetic lncRNA (slncRNA). Therefore, we need to further explore the relationship between sequence, structure and function of lncRNA molecules in a more systematically manner.

In this work, I studied RNA regulation from two different perspectives: understanding translation regulation of mRNA from a structural perspective and engineering synthetic lncRNA (slncRNA) for transcriptional activation.
In the first part of my research I employed Selective 2′Hydroxyl acylation Analyzed by Primer Extension followed by sequencing (SHAPE-Seq) to reveal the underlying structural changes lead to post-transcription down- or up-regulation phenomena previously observed in bacterial mRNA encoding for binding sites of RNA-binding proteins (RBP). I developed an extension to the SHAPE protocol by using a purified recombinant RBP added to *in vitro* RNA sample, to accomplish a complementary observation to the *in vivo* settings. By using the different SHAPE-Seq protocols, we established that the down-regulation effect is due to a transition from nonstrcutured translationally active state to repressed state exhibiting structured signature, which in turn inhibits translation. Additionally, the up-regulation effect apparently stems from highly closed structure that blocks translation, which is stabilized upon binding of the corresponding protein to facilitate translation.

In the second part, I describe the design of a slncRNA library and a screening system for functional variants. I successfully established a stable reporter cell-line based on an inducible mCherry gene, characterized by low basal levels and strong expression activation only in the presence of a transcription activator. Additionally, I took an innovative approach for oligo-pool study in mammalian cells by integrating it into an artificial chromosome of CHO cells. Although the overall goal of the second part of my research was not completed, I believe the work presented in this thesis may open the door to future work in the field of regulatory synthetic RNA.

List of Abbreviation

AMP - ampicillin

ATP - adenosine triphosphate

BA - bioassay media

°C - Celsius degree

C4-HSL - N-butanoyl-L-homoserine lactone

Cas9 - CRISPR associated protein 9

cDNA - circular DNA

CHO/K1 - chinese hamster ovary cell line K1

CIP - calf Intestinal

CMV - cytomegalovirus

CO2 - carbon dioxide

CRISPR - clustered regularly interspaced short palindromic repeats

2D - two-dimensional, 3D - three-dimensional

DMSO - dimethyl sulfoxide

DNA - deoxyribonucleic acid

dNTP - deoxynucleotide

DRBP - DNA-RNA-binding proteins

dsDNA - double-stranded DNA

DTT - dithiothretiol

EDTA - ethylenediaminetetraacetic acid

eYFP - enhanced yellow fluorescent protein

FACS - flow cytometry activated cell sorter

FBS - fetal bovine serum

GFP - green fluorescent protein

gr - gram

HAC - human artificial chromosome

HEK-293 - human embryonic kidney cells, clone 293

HOTAIR - HOX transcript antisense RNA

hr - hour

HSF1 - heat-shock factor 1

KAN - kanamycin

LB - lysogeny broth/Luria-Bertani

lncRNA - long non-coding RNA

M - molar

MBW - Molecular biology water

MCP - MS2 phage-coat protein

min - Minutes

M - molar

mL - milliliter, mM - millimolar, mg - milligram

mRNA - messenger RNA

NAI - 2-methylnicotinic acid imidazole

ncRNA - non-coding RNA

NGS - next-generation sequencing

NLS - nuclear localization signal

nM - nanomolar, ng - nanogram

nt - nucleotide

OD - Optical density

oligos - oligonucleotides

PBS - phosphate buffered saline

PCP - PP7 phage-coat protein

PCR - polymerase chain reaction

PEI - polyethylene imine

pmol - picomol

PRC2 - polycomb repressive complex 2

RBP -RNA-binding protein

RNA - ribonucleic acid

rpm - rounds per minute

rRNA - ribosomal RNA

RT - reverse transcriptase

sec - second

SHAPE - Selective 2'-hydroxyl acylation analyzed by primer extension

slncRNA - synthetic long non-coding RNA

ssDNA - single-stranded DNA

SV40 - simian virus 40

rTetR - reverse tetracycline repressor

U - Units

ubC - Ubiquitin C

UPW - Ultra pure water

UTR - untranslated region

v/v - Volume per volume

VPR - VP64-P65-Rta

XIST - X-inactive specific transcripts)

µg - microgram, µl - microliter

1. Introduction

1.1. Translation regulation of bacterial mRNA via RNA-binding protein

One of the main goals of synthetic biology is the construction of complex gene regulatory networks. The majority of engineered regulatory networks have been based on transcriptional regulation, with only a few examples based on post-transcriptional regulation[1–4], even-though RNA-based regulatory components have many advantages. Several RNA components have been shown to be functional in multiple organisms[5–9]. RNA can respond rapidly to stimuli, enabling a faster regulatory response as compared with transcriptional regulation[10–13]. From a structural perspective, RNA molecules can form a variety of biologically functional secondary and tertiary structures[2], which enables modularity. For example, distinct sequence domains within a molecule[13,14] may target different metabolites or nucleic acid molecules[15,16]. All of these characteristics make RNA an appealing target for engineered-based applications[2,3,17–22].

In bacteria, post-transcriptional regulation has been studied extensively in recent decades. There are well-documented examples of RBPs that either inhibit or directly compete with ribosome binding via a variety of mechanisms. These include direct competition with the 30S ribosomal subunit for binding via single stranded recognition[23], entrapment of the 30S subunit in an inactive complex via a nested pseudoknot structure[24] and ribosome assembly inhibition when the RBP is bound to a structured RBP binding site, or hairpin[25–28]. RNA hairpins have been studied in three distinct positions: either immediately downstream of the AUG[26], upstream of the Shine-Dalgarno sequence[27], or as structures that entrap Shine-Dalgarno motifs, as in the case for the PP7 and MS2 phage coat-protein binding sites. There is also a well-characterized example of translation stimulation: binding of the phage Com RBP was shown to destabilize a sequestered ribosome binding site (RBS) of the Mu phage *mom* gene, thereby facilitating translation[29,30]. While these studies indicate a richness of RBP-RNA-based regulatory mechanisms, a systematic understanding of the relationship between RBP binding, sequence specificity, the underlying secondary and tertiary RNA structure, and the resulting regulatory output is still lacking.

Synthetic biology approaches that simultaneously characterize large libraries of synthetic regulatory constructs have been increasingly used to complement the detailed study of single mRNA transcripts. While these synthetic approaches have been mostly applied to the transcriptional regulatory platforms[31–34], their potential for deciphering post-transcriptional regulatory mechanisms have been demonstrated in a recent study that interrogated IRES sequences in mammalian cells[35]. Building on this advancement and on a smaller-scale demonstration of translational repression by the RBP L7Ae in both bacteria and mammalian cells[12], we measured the regulatory output of a small library of synthetic constructs in which we systematically varied the position and type of RBP binding sites.

Our findings indicate that structure-binding RBPs (coat proteins from the bacteriophages GA[36], MS2[37], PP7[38], and Qβ[39]) can generate a range of translational responses, from previously-observed down-regulation[12] to, surprisingly, up-regulation. These results imply that RNA-RBP interactions can provide a platform for constructing gene regulatory networks that are based on translational, rather than transcriptional regulation.

1.2. RNA secondary structure interrogation using SHAPE-Seq

To further characterize the RBP-based regulatory effect from a structural perspective, we applied Selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)[15,22,40] to representative mRNA variants.

SHAPE-Seq is a relatively new method that aims to investigate secondary structures of RNA and its interaction with other molecules such as proteins or other nucleic acid. By combining chemical nucleotide labeling approaches[41–45] and next generation sequencing (NGS) we can obtain an insight into the structure of an mRNA molecule via selective modification of "unprotected" RNA segments. "Unprotected" segments mean single-stranded nucleotides that do not participate in any form of interaction, which include Watson-Crick base-pairing (secondary structure), tertiary interactions (*e.g.* Hoogsteen base-pairing, G-quadruplex formation, pseudoknots, *etc.*) and RBP-based interactions. These modifications cause the reverse transcriptase to stall and fall off the RNA strand, leading to a pool of cDNA molecules at varying lengths. Therefore, by counting the number of sequencing reads that end in positions along the molecule we can directly measure the number of molecules within this length and can estimate

the propensity of this RNA base to be un-bound (*i.e.* single-stranded). Subsequently, by applying bioinformatics analysis, we can generate a structural "footprint" of the chosen mRNA molecule *in vivo*, while in complex with ribosomes and/or other RBPs.

While other RNA probing methods are limited to *in vitro* analysis (*e.g.* PARS[46]) or suffers from nucleotides specifications (*e.g.* DMS[47]), SHAPE-Seq strength is in that if interrogates all four bases *in vivo,* allowing structure measurement at single-nucleotide resolution.

In this study we show that the mechanism for translation downregulation is most likely steric hindrance of the initiating ribosome by the RBP-mRNA complex which in turn leads to RNA-restructuring that spans a large segment of the RNA, including both the RBP binding site and the RBS. For the 5' UTR sequences that exhibit upregulation, RBP binding seems to facilitate a transition from an RNA structure with a low translation rate, into another RNA structure with a higher translation rate.

1.3. Long non-coding RNA (lncRNA)

The central dogma of biology posits that genomic DNA is transcribed into RNA, which is in turn translated into proteins that are responsible for most cellular functions. Approximately 10 years ago, it was discovered that while 90% of the human genome is being actively transcribed, only 1.5% of that RNA is translated into protein, thus providing the most serious challenge to date to the central dogma of biology. As a result of this discovery, research interest shifted to studying non-protein coding RNA (ncRNA) molecules and their role in cell biology. In the following year, ncRNA molecules were further classified to many different types of RNAs based mainly on their function (*e.g.* siRNA, miRNA, piwiRNA), and while the vast majority of ncRNA were short RNA molecules (<200 bp), a surprising class of long non-coding RNA molecules (lncRNA) whose length >200 bp was also discovered. The existence of these molecules was not thought to be possible due to the inherent instability of RNA inside the cells. As a result, lncRNAs were crudely defined as transcribed RNA molecules longer than 200 nucleotides, which are characterized by a conserved and stable 3D structures, despite rarely containing conserved sequence motifs[48–50]. In the last decade, tremendous increase in lncRNA publications (3966 results in PubMed in 2018

compared to 225 in 2008) have established that lncRNAs participate in various transcriptional regulatory roles via their interaction with Chromatin, other RNA molecules, and various proteins such as transcription factors[51,52]. Known lncRNA examples include: Xist, which plays an essential role in chromosome X-inactivation (XCI) of female cells by wide repression of gene expression[53], COOLAIR, which participate in floral regulation in plants through antisense silencing of the Flowering Locus C (FLC)[54], MALAT1[55], HOTAIR[56] and Gas5[57].

1.3.1. Transcriptional regulation by lncRNA

Through the years, it has been demonstrated that lncRNA molecules are often associated with chromatin, influencing its structure and modifications[58,59]. Intriguing findings also showed that purified chromatin contained twice as much RNA as DNA, indicating the close connection between the two[60] and support the idea of gene regulation by these RNA species.

Several hypothesis regarding the way lncRNA interact with the genome have been raised, including (i) RNA-DNA-binding proteins mediation, (ii) RNA:dsDNA triplex formation and (iii) RNA:RNA hybrid of lncRNA with a nascent transcript. In parallel, the ability of lncRNA to regulate gene expression is directly linked to their capability to interact with protein partners. There are three established mechanisms in which lncRNA may act: decoys, scaffolds and guides[61]. Decoys are lncRNA that can bind regulatory proteins and preclude their access to the DNA, while guides lncRNA are involved in the localization of specific proteins to the exact genomic locus. Yet, the most documented and interesting theme is scaffold lncRNA which serves as adaptor to bring together two or more proteins into a discrete complex[62,63]. Perhaps the most well-known example of scaffold lncRNA is HOTAIR, a marker of breast, colon and liver cancers, indicating its general oncogenic trait. HOTAIR provides secondary docking structure for both PRC2 (Polycomb repressive complex 2) and LSD1-CoREST complex, leading to H3K27 methylation and H3K4me2 demethylation which induce gene silencing at specific genomic loci[64].

1.3.2. Synthetic long non-coding RNA as regulatory elements

Until recently, gene regulation was mainly achieved directly by protein effectors such as transcription factors and repressors, due to the large, well-studied repertoire of natural regulators. However, as we deepen our understanding of

the principles governing RNA folding and functionality, more and more RNA-based applications are being developed.

Inspired by natural regulatory lncRNA, many researchers have tried to engineer genetic regulatory systems using synthetic or semi-synthetic lncRNA in bacteria, yeast and mammalian cells[65–68]. Already in 2011, Delebecque and colleagues[65] have shown an elegant design of RNA scaffold for spatial organization of metabolic pathway, demonstrating increased hydrogen biogenesis in bacteria. One year later, the bacterial type II CRISPR (clustered regularly interspaced short palindromic repeats) system[69,70] emerged in the biotechnology field and dramatically changed our ability to target the genome. Aside for genome editing using Cas9, many groups have tried to use the deficient version of Cas9 (*i.e.* dead-Cas9, dCas9) to localize longer and functional RNA cargos to specific DNA loci with the purpose of gene regulation. Zalatan *et al.*[67] established a CRISPR-based transcriptional program using modular synthetic RNA scaffolds to manipulate metabolic pathway in yeast cells, while Shechner *et al.*[66] expanded the CRISPR tool-box by showing the ability of dCas9 to target genomic DNA with natural lncRNA.

RNA is fundamentally modular and programmable, therefore it is highly suitable for regulating complex biological systems which require precise gene expression control. This will open the door to advanced biotechnology applications such as biological compound production (also in unique organisms which lack alternative regulation), novel sensors and cell-differentiation control.

1.3.3. Challenges in RNA engineering

The work of an engineer is highly dependent on her ability to understand the characteristics and behavior of materials, leading to the biggest challenge in molecular biology in general and in RNA engineering in particular. One of the most intriguing issue related to RNA biology is the complex relationship between sequence, structure and function. RNA is known for its unique ability to adopt specific secondary and tertiary structures, while some folding states are protein-dependent (RNA chaperones and other RNA-binding factors[71]). The secondary structure of RNA in mostly determined by Watson-Crick base-pairing to form single-stranded regions, loops and bulges. Long-term interactions may result in more complex structures of non-canonical base-pairing, pseudoknots and unique

tertiary motifs. Few studies established that while lncRNA sequences are poorly conserved across evolution, their secondary and tertiary structures show high conservation[48–50], indicating an evolutionarily-conserved relationship between RNA structure to its functionality.

To tackle the challenge of lncRNA design, it is necessary to take a more systematic approach to the design such as deep characterization or libraries of lncRNA sequences. Recently, there is a big emphasis on genome-scale RNA characterization in cellular environment ('structurome'), using chemical probing of flexible RNA nucleotide and next-generation sequencing (SHAPE-Seq[44,72] and SHAPE-MaP[73]). These recently developed techniques generate large database of information, expanding our knowledge on RNA folding and structure *in vivo*. Alternatively, one can design a library of putative function slncRNAs *de novo*, and subsequently screen for functional variants. In this manner, we can cover many putative designs of the RNA under investigations and select only the functional variants. In the past few years, DNA synthesis technologies have greatly improved, allowing large-scale production of thousands of single-stranded DNA oligonucleotides (oligos) efficiently and affordably. Therefore, many researchers are turning to high-throughput methodologies to explore different RNA features in cells. For example, Shukla and others[74] presented a massively parallel RNA assay to identify RNA-based nuclear localization domains harbored in lncRNA by screening a pool of ~12,000 oligos representing different human lncRNA. In another recent work[35], the authors designed 55,000 oligos to screen for novel cap-independent translation sequences, and to decipher regulatory elements driving IRES activity. Although promising, the major hurdle for implementation of library-based approach is the need for compatible high-throughput screening assay, to enable identification of functional variants.

In this thesis, I present the design of 40,000 synthetic lncRNA variants, which encompass variable RNA-binding sites of the MS2 phage-coat protein (MCP), combined with a screening assay using synthetic DNA-RNA-binding protein (sDRBP) and a fluorescence reporter gene. In the presence of functional slncRNA scaffold, MCP fused to transcription activator can be assembled nearby the minimal promoter and activate the expression of the reporter gene. By sorting and analyzing the activated cells we can gain an insight into the central factors of functional slncRNA.

2. Research objectives

As described above, alongside the growing interest in non-coding RNA, recent advances in DNA technologies of synthesis and sequencing are allowing us to explore and engineer RNA in a high-throughput manner. This thesis is divided into two main parts, employing these advanced abilities:

2.1. Understanding regulation of translation through RNA structure

Here I aimed to characterize the translational regulatory effect controlled by an RNA-binding protein (RBP) bound to a hairpin within a bacterial mRNA.

2.2. Engineering regulatory synthetic long non-coding RNA

Here I aimed to develop both experimental and design tools for the design of functional synthetic lncRNA (slncRNA) scaffolds. In order to accomplish the above, the research focuses on the following aims:

1. Developing and optimizing a reporter system in cells for screening of slncRNA transcription regulators using synthetic DNA-RNA-binding proteins (sDRBP).

2. Designing slncRNA library and integrate it into an artificial genome of mammalian cell-line.

3. Screening for functional slncRNA variants using the reporter system and flow-cytometry sorting.

3. Materials and Methods

3.1. SHAPE-Seq

3.1.1. Strains and constructs

SHAPE-Seq experiments were performed on 3 *E. coli* strains named PP7-wt δ=6, PP7-wt δ=-29 and PP7-USs δ=-29. All strains harbor a set of 2 plasmids: fusion-RBP plasmid and binding-site plasmid. The fusion-RBP plasmid (Ampicillin resistance) consists of PP7 phage coat protein (PCP) fused to an mCerulean gene under the so-called RhlR promoter[75], induced by N-butyryl-L-homoserine lactone (C_4-HSL). The Binding-site plasmids (Kanamycin resistance) contain one wild-type or mutated RBP binding sites, at varying distances, either upstream (δ<0) or downstream (δ>0) to the RBS of an mCherry gene. Strains were obtained from previous work done in our lab by Noa Katz and others[76,77].

3.1.2. Experimental setup

LB medium supplemented with appropriate concentrations of Amp and Kan was inoculated with glycerol stocks of bacterial strains harboring both the RBP-fusion plasmid and the binding-site plasmid and grown at 37°C for 16 hr while shaking at 250 rpm (see **Figure 1** for SHAPE-Seq methodology). Overnight cultures were diluted 1:100 into semi-poor medium (95% BA and 5% LB). Each bacterial sample was divided into a non-induced sample and an induced sample in which RBP protein expression was induced with 250 nM N-butanoyl-L-homoserine lactone (C_4-HSL), as described above.

Bacterial cells were grown until OD_{600}=0.3, 2 mL of cells were centrifuged and gently resuspended in 0.5 mL semi-poor medium. For *in vivo* SHAPE modification, cells were supplemented with a final concentration of 30 mM 2-methylnicotinic acid imidazole (NAI) suspended in anhydrous dimethyl sulfoxide (DMSO, Sigma-Aldrich)[78], or 5% (v/v) DMSO. Cells were incubated for 5 min at 37°C while shaking and subsequently centrifuged at 6000 g for 5 min. RNA isolation of 5S rRNA was performed using TRIzol-based standard protocols. Briefly, cells were lysed using Max Bacterial Enhancement Reagent followed by TRIzol treatment (Life Technologies). Phase separation was performed using chloroform. RNA was precipitated from the aqueous phase using isopropanol and ethanol washes, and then resuspended in RNase-free water.

For the strains harboring PP7-wt δ=-29 and PP7-USs δ=-29, column-based RNA isolation (RNeasy mini kit, QIAGEN) was performed. Samples were divided into the following sub-samples (except for 5S rRNA, where no induction was used):

1. induced/modified (+C_4-HSL/+NAI)
2. non-induced/modified (-C_4-HSL/+NAI)
3. induced/non-modified (+C_4-HSL/+DMSO)
4. non-induced/non-modified (-C_4-HSL/+DMSO).

In vitro modification was carried out on DMSO-treated samples (3 and 4) and has been described elsewhere[44]. Briefly, 1500 ng of RNA isolated from cells treated with DMSO were denatured at 95°C for 5 min, transferred to ice for 1 min and incubated in SHAPE-Seq reaction buffer (100 mM HEPES [pH 7.5], 20 mM $MgCl_2$, 6.6 mM NaCl) supplemented with 40 U of RiboLock RNAse inhibitor (Thermo Fisher Scientific) for 5 min at 37°C. Subsequently, final concentrations of 100 mM NAI or 5% (v/v) DMSO were added to the RNA-SHAPE buffer reaction mix and incubated for an additional 5 min at 37°C while shaking. Samples were then transferred to ice to stop the SHAPE-reaction and precipitated by addition of 3 volumes of ice-cold 100% ethanol, followed by incubation at -80°C for 15 min and centrifugation at 4°C, 17000 g for 15 min. Samples were air-dried for 5 min at room temperature and resuspended in 10 µL of RNAse-free water.

Subsequent steps of the SHAPE-Seq protocol, that were applied to all samples, have been described elsewhere[72], including reverse transcription, adapter ligation and purification as well as dsDNA sequencing library preparation. In brief, 1000 ng of RNA were converted to cDNA using the reverse transcription primers for mCherry or 5S rRNA that are specific for either the mCherry transcripts (PP7-wt δ=6, PP7-USs δ=-29 or PP7-wt δ=-29). The RNA was mixed with 0.5 µM primers and incubated at 95°C for 2 min followed by an incubation at 65°C for 5 min. The Superscript III reaction mix (Thermo Fisher Scientific; 1x SSIII First Strand Buffer, 5 mM DTT, 0.5 mM dNTPs, 200 U Superscript III reverse transcriptase) was added to the cDNA/primer mix, cooled down to 45°C and subsequently incubated at 52°C for 25 min. Following inactivation of the reverse transcriptase for 5 min at 65°C, the RNA was hydrolyzed (0.5 M NaOH, 95°C, 5 min) and neutralized (0.2 M HCl). cDNA was precipitated with 3 volumes of ice-cold 100% ethanol, incubated at -80°C for 15 min, centrifuged at 4°C for 15 min at 17000 g and resuspended in 22.5 µl ultra-pure water. Next, 1.7 µM of 5'

phosphorylated ssDNA adapter was ligated to the cDNA using a CircLigase (Epicentre) reaction mix (1xCircLigase reaction buffer, 2.5 mM $MnCl_2$, 50 µM ATP, 100 U CircLigase). Samples were incubated at 60°C for 120 min, followed by an inactivation step at 80°C for 10 min. cDNA was ethanol precipitated (3 volumes ice-cold 100% ethanol, 75 mM sodium acetate [pH 5.5], 0.05 mg/mL glycogen [Invitrogen]). After an overnight incubation at -80°C, the cDNA was centrifuged (4°C, 30 min at 17000 g) and resuspended in 20 µl ultra-pure water. To remove non-ligated adapter, resuspended cDNA was further purified using the Agencourt AMPure XP beads (Beackman Coulter) by mixing 1.8x of AMPure bead slurry with the cDNA and incubation at room temperature for 5 min. The subsequent steps were carried out with a DynaMag-96 Side Magnet (Thermo Fisher Scientific) according to the manufacturer's protocol. Following the washing steps with 70% ethanol, cDNA was resuspended in 20 µL ultra-pure water and were subjected to PCR amplification to construct dsDNA library as detailed below.

3.1.3. *In vitro* SHAPE-Seq with recombinant protein

In vitro modification with recombinant protein was carried on non-induced, DMSO-treated samples, similarly to the detailed above with the following changes: after RNA refolding, 15.6 pmol (based on 1:2 molar ratio between RNA:PP7 protein) of highly-purified recombinant PP7 coat-protein (Genscript) were added to the RNA samples and incubated at 37°C for 30 min. Subsequently, final concentrations of 100 mM NAI or 5% (v/v) DMSO were added to the RNA-PP7 protein reaction mix and incubated for an additional 10 min at 37°C. Downstream steps kept unchanged.

3.1.4. Library preparation and sequencing

To produce the dsDNA for sequencing 10 µL of purified cDNA from the SHAPE procedure (see above) were PCR amplified using 3 primers: 4nM mCherry selection or 5S rRNA selection primer, 0.5µM TruSeq Universal Adapter and 0.5µM TrueSeq Illumina indexes with PCR reaction mix (1x Q5 HotStart reaction buffer, 0.1 mM dNTPs, 1 U Q5 HotStart Polymerase [NEB]) (see **Figure 1** for SHAPE-Seq methodology). A 15-cycle PCR program was used: initial denaturation at 98°C for 30 sec followed by a denaturation step at 98°C for 15 sec, primer annealing at 65°C for 30 sec and extension at 72°C for 30 sec, followed by a final

extension 72°C for 5 min. Samples were chilled at 4°C for 5 min. After cool-down, 5 U of Exonuclease I (ExoI, NEB) were added, incubated at 37°C for 30 min followed by mixing 1.8x volume of Agencourt AMPure XP beads to the PCR/ExoI mix and purified according to manufacturer's protocol. Samples were eluted in 20 µL ultra-pure water. After library preparation, samples were analyzed using the TapeStation 2200 DNA ScreenTape assay (Agilent) and the molarity of each library was determined by the average size of the peak maxima and the concentrations obtained from the Qubit fluorimeter (Thermo Fisher Scientific). Libraries were multiplexed by mixing the same molar concentration (2-5 nM) of each sample library, and library and sequenced using the Illumina HiSeq 2500 sequencing system using either 2X51 paired end reads for the 5S-rRNA control and *in vitro* experiments or 2x101 bp paired-end reads for all other samples.

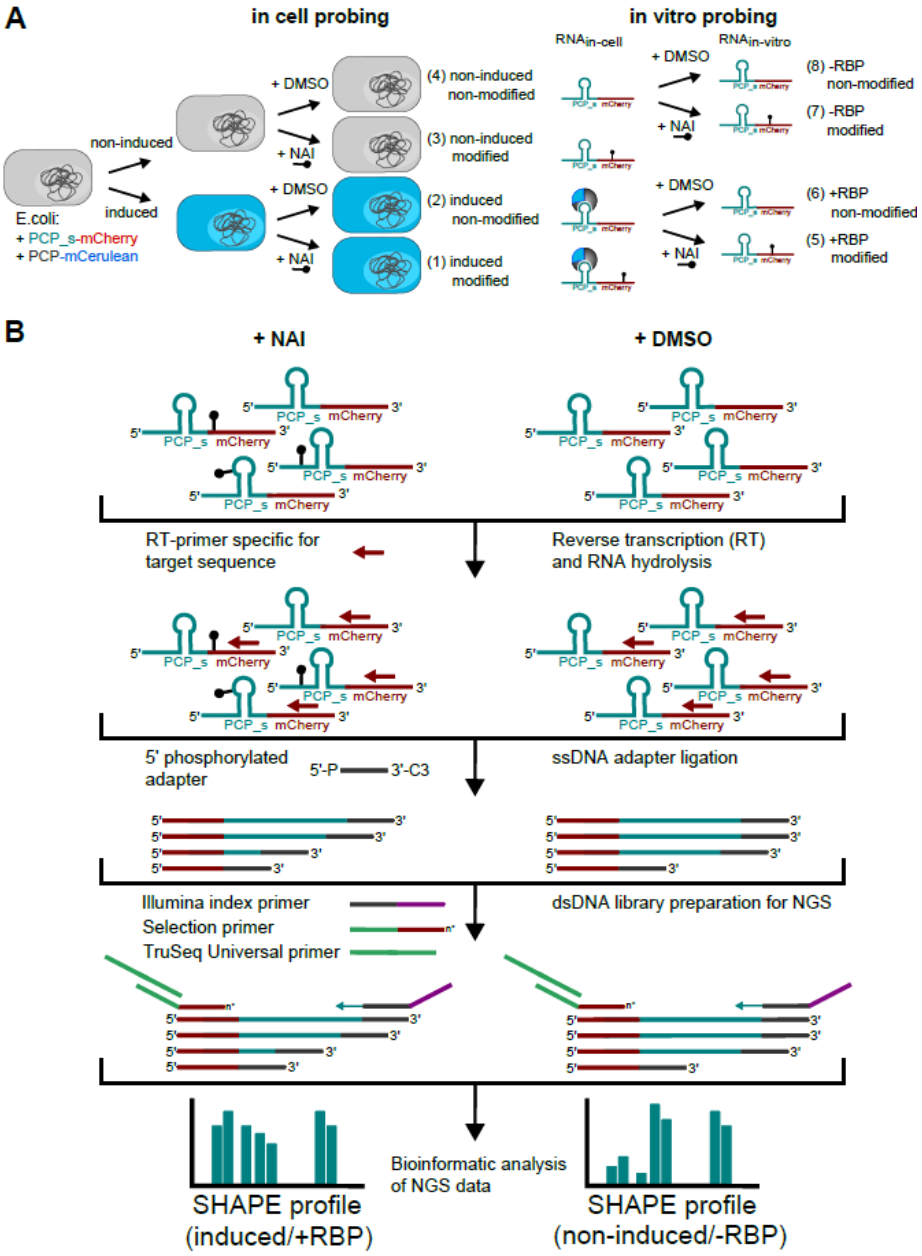

Figure 1: *Schematic overview of SHAPE-seq experiment.*

(A) *Overnight-grown bacterial strains harboring both the RBP-binding site plasmid (containing the mCherry reporter) and the RBP-fusion plasmid (PCP-mCerulean) are split into two samples and PCP-mCerulean expression is induced (using C4-HSL) for one of them. Following protein expression, each bacterial sample is further split and treated with either DMSO (as the non-modified control) or NAI. Subsequently, RNA is isolated and either further chemically probed (samples 2+4) or directly used for subsequent steps of SHAPE-seq (samples 1+3).* ***(B)*** *Following 2' hydroxyl acylation and subsequent RNA isolation, RNA samples are reverse-transcribed using a gene-specific primer that binds in the target transcript. During reverse transcription, reverse transcriptase is stalled one nucleotide before the modification. Subsequently, a single-stranded 5' phosphorylated (5'P) and 3-carbon spacer (3'C) adapter sequence is ligated to the obtained cDNAs, which serves in the next step as a primer-binding site for the Illumina index primers to prepare double-stranded DNA for Illumina next generation sequencing.*

3.1.5. SHAPE-Seq analysis

3.1.5.1. Initial reactivity analysis

Illumina reads were first adapter-trimmed and were aligned against a composite reference for mCherry or *E. coli* 5S rRNA sequences.

Reverse transcriptase (RT) drop-out positions were indicated by the end position of Illumina Read 2 (the second read on the same fragment). Reads that were aligned only to the first 19 bp were eliminated from downstream analysis, as these correspond to the RT primer sequence. For each position upstream of the RT-primer, the number of drop-outs detected was summed (see **Figure 2** for SHAPE-Seq analysis). To facilitate proper signal comparison, all libraries were normalized to have the same total number of reads. For each library j and position x=1...L, we normalized the number of drop-outs $D_j(x)$ according to:

$$\widehat{D}_j^0(x) = \frac{D_j^0(x)}{\sum_{i=1}^{L} D_j^0(x)} \qquad (1)$$

where L is the length of the sequence under investigation after RT primer removal.

3.1.5.2. Bootstrap analysis

To compute the mean read-ratio, reactivity, and associated error bars, we employed bootstrap statistics in a classic sense. Given *M* reads per library, we first constructed a vector of length M, containing the index of the read # (1...M) and an associated nucleotide position x per index. Next, we used a random number generator (MATLAB) and pick a number between 1 and *M*, *M* times to completely resample our read space. Each randomly selected index number was matched with a position *x.* The length *x* was obtained from the matching index in the original non-resampled library $\widehat{D}_j^0(x)$. We repeated this procedure 100 times to generate 100 virtual libraries from the original $\widehat{D}_j^0(x)$ to generate $\widehat{D}_j^k(x)$, where k = {1...100}.

3.1.5.3. Signal-to-noise (read-ratio) computation

SHAPE-Seq read-ratio was computed as the ratio between each pair of NAI-modifed and umodified (DMSO) samples, defined for each individual nucleotide. Furthermore, mean read-ratio vector and associated standard errors were also computed.

For each pair of NAI-modifed and umodified (DMSO) resampled libraries for a particular sample s $\left[\widehat{D}^k_{s,mod}(x), \widehat{D}^k_{s,non-mod}(x)\right]$ we computed the SHAPE-seq read-ratio for each position i to generate a read-ratio matrix as follows:

$$R^k_s(x) = \frac{\widehat{D}^k_{s,mod}(x)}{\widehat{D}^k_{s,non-mod}(x)} \qquad (2)$$

where the read-ratio is a signal-to-noise observable defined for each individual nucleotide. To obtain the mean read-ratio vector and associated standard errors, we computed the mean and standard deviation of the read-ratio per position as follows:

$$\langle R_s(x)\rangle = \frac{1}{100}\sum_{k=0}^{100}\frac{\widehat{D}^k_{s,mod}(x)}{\widehat{D}^k_{s,non-mod}(x)} \qquad (3)$$

$$\sigma_s(x) = \langle R_s(x)\rangle - \langle R_s(x)\rangle^2 \qquad (4)$$

3.1.5.4. Reactivity computation

The literature has several redundant definitions for reactivity, and no consensus on a precise formulation[41,79,80]. The simplest definition of reactivity is the modification signal that is obtained above the background noise. As a result, we define the reactivity as follows:

$$\rho^k_s(x) = \left(R^k_s(x) - 1\right)\Theta\left(R^k_s(x) - 1\right) \qquad (5)$$

Where,

$$\Theta(x) = \begin{cases} 0 \; if \; x < 0 \\ 1 \; if \; x \geq 0 \end{cases} \qquad (6)$$

For the average reactivity score obtained for each position for a given sample s:

$$\rho_s(x) = (\langle R_s(x)\rangle - 1)\Theta(\langle R_s(x)\rangle - 1) \qquad (7)$$

For the running-average reactivity plots, we used the following procedure: first, we computed an average reactivity per position based on two bootstrapped

mean reactivity scores that were obtained from the two biological replicates. We then computed a running average of 10 nt window for every position X.

Error bars were computed from the bootstrapped sigma of a certain technical repeat and from the standard deviation of the read-ratio values for each *N* repeat. Finally, the error bar of each position was computes in accordance with the running average.

3.1.5.5. Determining protected regions and differences between signals

To determine regions of the RNA molecules that are protected by the RBP, we employ a Z-factor analysis on the difference between the read-ratio scores. Z-factor analysis is a statistical test that allows comparison of the differences between means taking into account their associated errors. If Z > 0 then the two means are considered to be "different" in a statistically significant fashion (*i.e.* > 3σ). The regions that were determined to generate a statistically different mean reactivity values, and also resulted in a positive difference between the -RBP and +RBP cases were considered to be protected and marked accordingly.

3.1.5.6. Structural visualization

For the structural visualization (as in **Figure 8**), the mRNA SHAPE-Seq fragment of PP7-wt and PP7-USs δ=-29 constructs was first folded *in silico* using RNAfold in default parameters. For visualization purposes, the SHAPE-Seq reactivity scores were used as colormap to overlay the reactivity on the RNAfold predicted structure and to generate the structure image.

3.1.5.7. Using the empirical SHAPE-Seq data as constraints for structural prediction

In order to predict more accurate structural schemes for PP7-wt and PP7-USs δ=-29 constructs (as in **Figure 9**) we used the SHAPE-Seq data as constraints to the computational structure prediction. We computed the inferred structures by using the calculated reactivities of each sample as perturbations that minimizes the discrepancies between the predicted and empirically inferred pairing probabilities. Based on the structural ensemble, the resulted probability of pairing for each nucleotide was calculated.

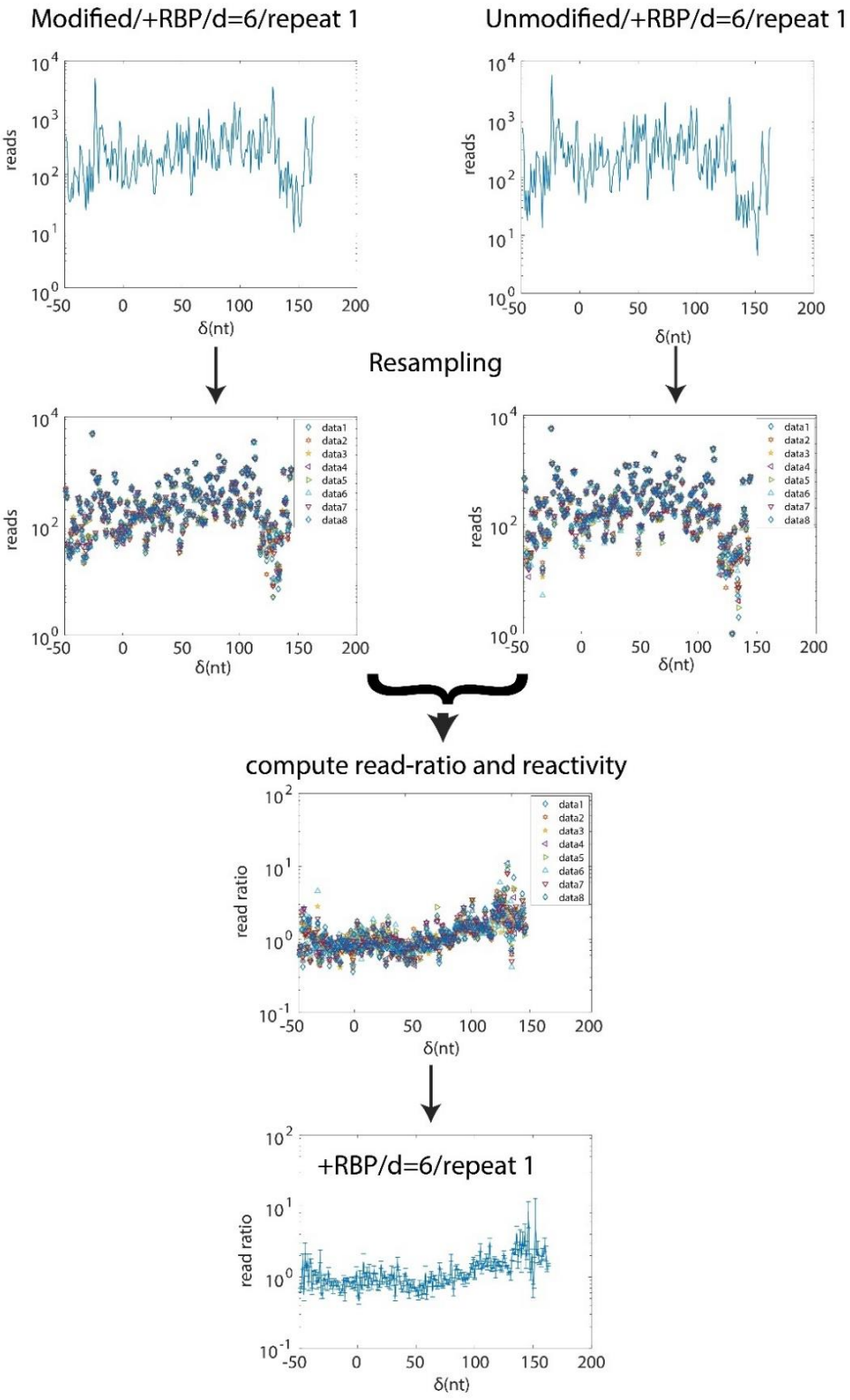

Figure 2: *Schematic overview for SHAPE-Seq analysis for a given data-set.*

Initially, the number of drop-outs (read end-point) at each position is summed. Next, the original data is resampled (bootstrapping) to enable computation of mean read-ratio, reactivity and the associated error bars. The read-ratio is computed between each pair of modified and unmodified samples at the single nucleotide level.

3.2. Cloning of bacterial and mammalian plasmids

Cloning procedures were done using standard molecular biology techniques such as Polymerase chain reaction (PCR), oligonucleotide annealing, restriction enzymes, ligation and Gibson assembly[81] (all enzymes purchased from New England Biolabs, NEB). Recombinant DNA was transformed into *E.coli* Top10 (Invitrogen) using a standard heat-shock transformation, after which the bacteria were grown overnight at 37°C on an LB agar plate containing the appropriate selection antibiotics.

Bacterial colony PCR analysis was performed to isolate desired clones (Taq Ready Mix (2X), hy-labs). Extraction and purification of DNA from cells (miniprep) was done using NucleoSpin Plasmid Easy Pure Kit (Macherey-Nagel), DNA purification from gels and *in-vitro* enzymatic reactions was carried with Wizard SV Gel and PCR Clean-Up system (Promega).

3.3. Activation domains screening in mammalian cells

3.3.1. Design and construction of pTRE-mCherry reporter plasmid

The pTRE-mCherry plasmid comprised of 7 repeats of a tetracycline operator (tetO) sequence upstream to minimal CMV promoter which regulate the transcription of mCherry protein (see **Figure 3A**). It was derived from the pTRETightBI-RY-0 vector (ordered from Addgene #31463) by digesting it with restriction enzymes (XbaI and XhoI) followed by gel extraction of the backbone without the eYFP. Subsequently, the vector was ligated with a 'filling' fragment of 33 random nucleotides (annealed oligos ordered from Sigma-Aldrich).

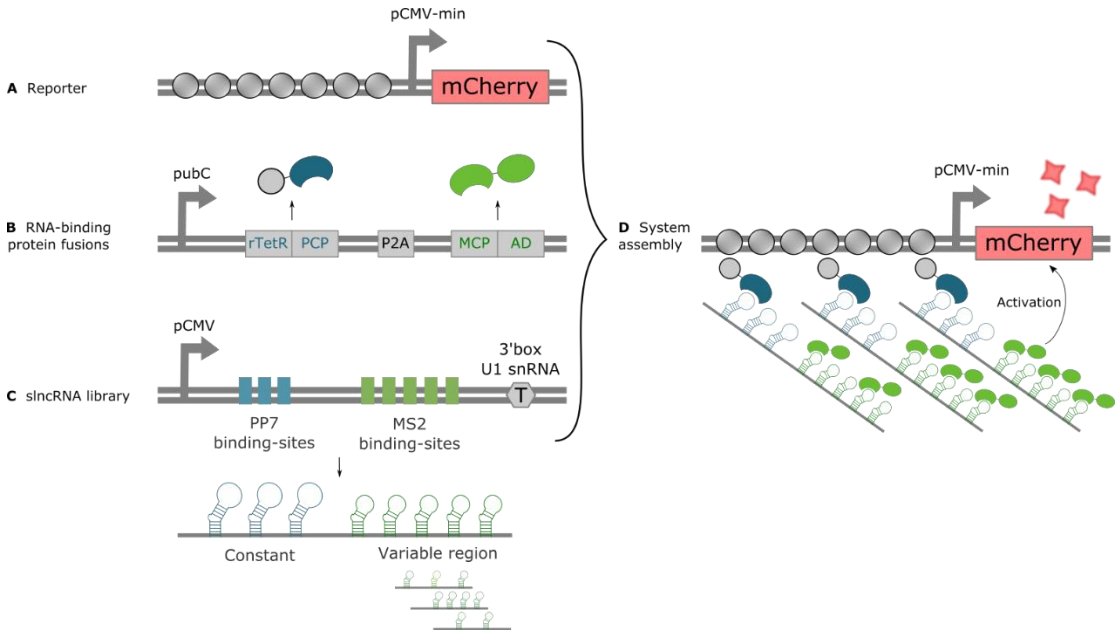

Figure 3: *Schematic of the basic parts in the sIncRNA screening system.*

(A) *The inducible reporter construct consists of the TRE promoter (tetO binding-sites and minimal CMV promoter) and mCherry fluorescent protein.* ***(B)*** *The vector pUC57-sRBP encodes the synthetic RNA-binding protein fusions rTetR-PCP (sDRBP) and MCP-P65-HSF1 (RBP-AD). The proteins are translated separately due to the P2A self-cleavage peptide.* ***(C)*** *sIncRNA library was ordered as an oligo-pool consists of a constant region of 3 PP7 binding-sites (BS) and a variable region of 0 to 5 MS2 BS. (**D**) Assembly of the full system: the rTetR-PCP links the DNA and sIncRNA while the MCP-P65-HSF1 assembled on the sIncRNA and activates mCherry transcription.*

3.3.2. Design and construction of rTetR-activation domain fusions

The construction of pubC-rTetR-AD-YFP variants (see illustration in **Figure 4**) was carried in 2 subsequent steps. First, Gibson assembly of 3 parts: pubC-YFP backbone (PCR amplified), ~700bp sequence encodes for rTetR protein (PCR amplified) and dsDNA fragment consists of nuclear localization signal (NLS), KpnI and EcoRI restriction sites and P2A self-cleavage peptide[82] (annealed oligos, ordered from Sigma-Aldrich). Second, the verified pubC-rTetR-YFP backbone was restricted by KpnI and EcoRI and ligated with one of the inserts encoding for an activation domain (PCR amplified). Four activation domains (AD) were tested: VP64[83], P300[84], VPR[85] and HSF1-p65[86] (see **Table 1** for sequence details).

Table 1: *DNA fragments used for pubC-rTetR-AD-YFP cloning. The DNA source for each PCR amplified fragment*

Amplified fragment	DNA source
rTetR	pBSKΔB-CAG-rtTA2sM2-IRES-tTSkid-IRES-Neo (Addgene #62346)
P300	pcDNA-dCas9-p300 Core (Addgene #61362)
VPR	pAAV-CMV-Cas9C-VPR (Addgene #80933)
HSF1-p65	lenti MS2-P65-HSF1_Hygro (Addgene #61426)
NLS-RE-P2A	Custom design and annealing of oligos

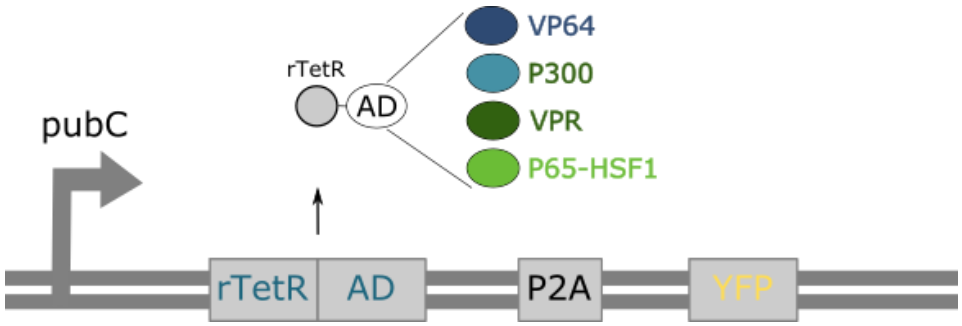

Figure 4: *rTetR-AD construct for activation domains screening.*

Fusion of the DNA-binding protein rTetR with an activation domain of either VP64, P300, VPR or P65-HSF1. YFP protein is expressed separately due to P2A self-cleavage peptide, and act as a marker for expression during flow cytometry analysis.

3.3.3. HEK293 cell culture growth and transfection

The human embryonic kidney cell line (HEK293, kindly provided by Arie Admon's lab, Technion) was incubated and maintained in 10 cm cell culture dishes (Nunclon cell culture treated, Thermo Scientific) under standard cell culture conditions at 37 °C in humidified atmosphere containing 5% CO_2.

Cells were passaged at 85% confluence by treatment of 1x PBS wash and trypsin followed by incubation at 37 °C for 2 min. Growing media DMEM (Dulbecco Eagle's Minimum Essential Medium) was complemented with 10% FBS (fetal bovine serum) and 5% penicillin-streptomycin solution (all purchased from Biological Industries, BI).

Transient transfection of HEK293 carried in 96-well tissue culture plate (Nunclon cell culture treated, Thermo Scientific) by seeding 40,000 cells at the day of transfection. DNA complexes were prepared in total volume of 10 µL OptiMEM (Gibco/Life Technologies) by mixing 100 ng of DNA with PEI (PolyEthyleneImine) in a ratio of 1:5. After 15 min incubation at room temperature cells were added to the DNA mixture and the plate was incubated at 37 °C overnight. Culture medium was replaced after 12-18 hr.

3.3.4. Flow cytometry experiment

48 hr post-transfection, HEK293 cells were washed once with 1x PBS and incubated for 4 min at 37 °C with trypsin. Subsequently, cells were suspended in FACS running buffer (1x PBS complemented with 1% FBS and 3 mM EDTA). Data acquisition was performed on the MACSQuant (Miltenyi Biotec) analyzer using the proprietary MACSQuantify software. Histograms were adjusted according to the auto-fluorescence of non-transfected cells. Data collected from the experiments were analyzed using FlowJo analysis software (FlowJo LLC). The percentages of cells expressing mCherry as well as the median fluorescence intensities were exported and used to calculate activation of the reporter gene in each sample.

3.4. Reporter gene cell-line construction

3.4.1. Design and construction of vectors

Since genomic integration of recombinant DNA requires selection marker for mammalian cell culture, pTRE-mCherry vector (described in Method section 3.1) was cloned with Blasticidin resistance gene. Blasticidin sequence was PCR amplified from pMSCV-Blasticidin (Addgene #75085) with primers adding restriction sites of BglII and XbaI. Backbone vector and PCR product were digested with BglII and XbaI enzymes followed by standard ligation protocol.

3.4.2. CHO cell culture growth, transfection and random genomic integration

CHO-K1-MI-HAC (kindly provided by Y. Kazuki and M. Oshimura and hereby referred to as simply CHO cells) were grown in F-12 Nutrient Mixture (HAM's) medium (BI), supplemented with 10% FBS and 1% Penicillin-Streptomycin solution (BI), and cultured at 37 °C and 5% CO_2 in humidified atmosphere. CHO cells were subcultures 2 times a week in 1:10 ratio.

For transient transfection, 10,000 CHO cells were seeded in 96-well tissue culture plate 24 hr prior transfection. At time of transfection 100 ng DNA were mixed with 0.3 µL PolyJet (SignaGen) in final volume of 10 µL serum-free medium. DNA-PolyJet mix was incubated for 10 min at room temperature and subsequently added drop-wise to the cells. 16 hr post transfection medium was replaced and 24 hr later the cells were analyzed by FACS (for details see Flow cytometry section 3.4 above).

When transfection was carried with more than one plasmid, I used the pUC19 plasmid as an empty plasmid for control samples (to keep the amount of DNA constant).

Random integration of pTRE-mCherry-Blasticidin construct into the genome of CHO cells was carried in 6-well plate by transfection of 150,000 cells with 2.5 µg DNA and 12.5 µL PEI (1 mg/mL) in total volume of 150 µL OptiMEM (Gibco/Life Technologies). Cells were incubated overnight at 37 °C and subsequently passed into 10 cm dish with selective medium of 8 µg/mL Blasticidin (InvivoGen). The generated cell-line will be termed hereby 'CHO-mCherry'.

3.4.3. Cell sorting and single variant selection

CHO-mCherry cells were sorted to single cells using the FACSAria cell sorter (Becton-Dickinson) and were collected in 96-well plate contains complete F-12 media (10% FBS, 1% PS) enriched with 5% FBS to facilitate cell recovery. Cells with low mCherry levels were sorted into 96-well plate (FACS parameters were calibrated according to native CHO cells).

Cells were cultured and expanded from 96-well to 24-well format for 1 month and were then subjected to mCherry activation experiment by transient transfection of vector encoding for rTetR-p65-HSF1 fusion. Levels of mCherry were measures by flow-cytometry, the selected variant presented a profile of low basal mCherry with strong mCherry expression upon induction of doxycycline.

3.5. Design of RNA-binding proteins fusions cassette

The vector pUC57-sRBP encodes the synthetic RNA-binding protein fusions (sRBP) was ordered from GenScript as a custom 4350bp sequence cloned between AatII/EcoRV restriction sites. The vector (see illustration in **Figure 3B**)

encoding the sRBP fusions, linked by P2A self-cleavage peptide[82], so that they are transcribed in the same mRNA but translated independently. The first fusion is the synthetic DNA-RNA-binding protein (sDRBP) consist of rTetR and tandem PCP, while the second is an MCP (N55K mutant) fused to the activation domains p65 and HSF1. Both fusion proteins carry nuclear localization sequence (NLS) and fluorescent protein as a marker for expression (eCFP and eYFP, respectably).

3.6. slncRNA library

3.6.1. Backbone vector for HAC integration

The pNeo-attB(ΦC31)-CMV-3'box construct was cloned from a backbone containing the ΦC31 attB site, Neomycin resistance gene and a CMV promoter. This construct was designed such that after integration, the neomycin gene would be expressed from a PGK promoter situated upstream of the ΦC31 site in the human artificial chromosome (HAC) of CHO cells. The ΦC31 constructs (including the ΦC31 integrase) were a gift from the Oshimura Lab[87].

Backbone was digested with AgeI/NotI restriction enzymes and cleaned from gel. Double-stranded linear DNA fragment was ordered from IDT as a gBlock encoding for AgeI restriction site, x3-PP7 binding-sites, EcoRI and AvrII restriction sites for library insertion, the sequence of 3'box as a non-polyadenylated terminator and NotI restriction site. Sequences of PP7 binding-sites and 3'box were derived from pCMV/3«Box_(GLuc)_INT[66] (Addgene #68436). 1 µg gBlock were digested by AgeI and NotI, cleaned and ligated with the pNeo-attB(ΦC31) backbone described above.

To prepare the backbone (pNeo-attB(ΦC31)-CMV-3'box) for ligation with the oligo library it was digested twice with EcoRI and AvrII, followed by dephosphorylated (CIP) in order to ensure as little self-ligation as possible (background noise).

3.6.2. Oligo-library design

Single-stranded oligo-pool was ordered from TWIST Bioscience as a library of 40,000 variants of length 171bp. Each variant consists of a constant sequence at the 3' and 5' ends for primer binding sites and restriction sites (EcoRI/AvrII). The 101bp variable region encodes for 0 to 5 MS2 binding-sites variants with random spacer sequences. **Figure 3C** illustrate the general design of slncRNA library.

The oligo-pool sequences were generated in collaboration with Leon Anavy from the Department of Computer Science at the Technion, by using a customized Python script. General description of the library design is presented in **Table 2**.

The MS2 binding-sites mutated variants were obtained from concurrent research carried out in the lab by Noa Katz (unpublished). As for the spacer sequences, since there are no defined rules for slncRNA design, the spacers were designed from random sequences, either linear or hairpin-structured, under the assumption that hairpins may assist in stabilizing the structure.

Table 2: *slncRNA library general design. Number of MS2 binding-sites, spacers and total amount of variants at each group*

Binding-sites of MS2	**Spacers**		**Total variants**
Number	**Number**	**Length (bp)**	**Number**
5	6	6	3000
4	5	25	6000
3	4	44	18000
2	3	63	9000
1	2	82	3000
0	1	101	500
			39500

3.6.3. Oligo-library cloning

Oligo-pool was reconstituted in Ultra-pure water (BI) and was then PCR amplified in 96-well plate for 14 cycles. Each well contained 10 ng ssDNA, 5 µL from specific forward and reverse primers (10 µM), 10 µL 5X-Q5 Reaction Buffer (NEB), 1 µL dNTPs (10 mM, Sigma-Aldrich), 0.5 µL Q5 High-Fidelity DNA Polymerase (NEB), completed to final volume of 50 µL with Ultra-pure water. Next, residual ssDNA were digested with 5 µL exonuclease ExoI (20,000 units/mL, NEB) at 37˚C for 30 min. Subsequently, all wells were collected and cleaned (Wizard SV Gel and PCR Clean-Up system, Promega). To verify a product size of 163bp, the dsDNA was analyzed using the ScreenTape assay (2200 Tapestation, Agilent). Next, dsDNA was digested with EcoRI and AvrII and cleaned. 3.3 ng of library were ligated

with 100 ng of backbone (see section 6.1) in ratio of 1:1 to generate pNeo-attB(ΦC31)-CMV-library-3'box.

Ligation mix (2 µL) was transformed into E.cloni 10G (Lucigen) and cells were plated on 15cm petri-dish and incubated overnight at 37˚C. The day after, colonies were collected in LB using cell scraper (Biologix group) and centrifuged at 5000 rpm for 15 min. DNA was purified from cell pellet using NucleoBond Xtra Midi kit (MN).

3.6.4. Integration into HAC of CHO cells

Integration of the recombinant DNA (GFP or slncRNA library) into the HAC of CHO cells was performed by co-transfecting 3 µg recombinant DNA plasmid and 1 µg ΦC31 integrase plasmid, using PolyJet (SignaGen). The transfection was performed on 1M cells in 6-well plates. 48 hr post transfection cells medium was changed to selective medium with 600 ng/µL Neomycin (G418, Sigma-Aldrich). Cells were selected for 14 days, expanded and frozen in 5% DMSO in liquid nitrogen.

Integration specificity into the HAC was supported by control samples of cells transfected with only 1 out of 2 required constructs, either the integrase plasmid (pCMV-ΦC31) or integration backbone (GFP or pNeo-attB(ΦC31)-CMV-library-3'box). In both controls the cell didn't survive the antibiotic selection.

3.6.5. Genomic PCR of HAC integration

Genomic PCR was done to accomplish 2 goals: First, to verify the integration of the desired construct into the HAC and second, to amplify the integrated variants for NG-sequencing.

Genomic DNA (gDNA) was purified from 5M pellet resuspended in 200 µL PBS using ExgeneTM Cell SV, mini (GeneAll) according to the kit manual.

PCR was performed on 10 ng gDNA with primers for either the pNeo-attB(ΦC31) backbone for general gel analysis on a broader region (1265 bp) or primers adjacent to the variable region of the library for sequencing purposes (230 bp).

4. Results

4.1. Understanding regulation of translation through RNA structure

A previous work done by us (Katz *et al.*)[76,77] showed translational regulatory effect controlled by an RBP bound to its cognate hairpin binding-site, located within bacterial mRNA encoding for a reporter protein (mCherry). The hairpin was positioned in either the ribosomal initiation region, namely downstream to the AUG (annotated as δ>0) or upstream, at the 5' UTR (δ<0). To obtain a dose-response function, a fusion of the RBP and mCerulean fluorescent protein was expressed from the inducible promoter RhlR (see system setup in **Figure 5**).

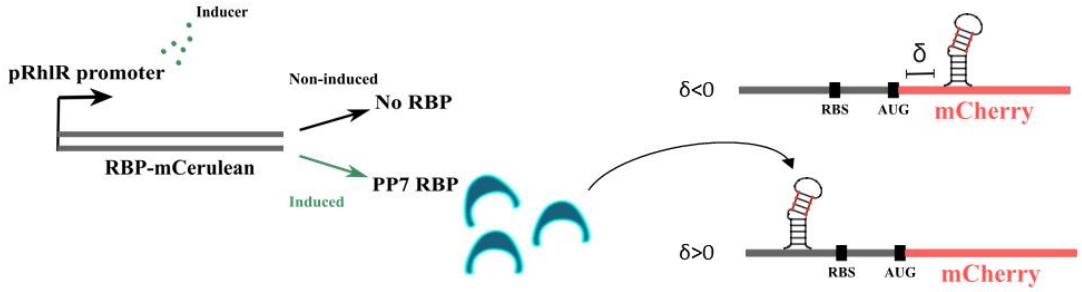

***Figure 5**: Translational regulation circuit by a RBP-hairpin complex.*

PP7 RBP-mCerulean expression is under the control of pRhlR, activated by the C_4-HSL inducer. Upon expression, the RBP can bind to its cognate RNA site. The mCherry reporter mRNA (expressed constitutively) encoding a folded RBP binding-site in either the ribosomal initiation region (δ>0) or in the 5' UTR (δ<0).

In this thesis I will discuss the results of several designs involve the PP7 phage coat protein (PCP) and its cognate hairpin in two different conformations: PP7-wildtype (PP7-wt) and the mutated PP7-Upper Stem short (PP7-USs), which featured with a deletion of two nucleotides in the upper stem of the hairpin (see **Figure 6B**). The production rate of mCherry was measured as a function of increasing concentrations of the RBP. The results for the δ>0 and δ<0 constructs are presented in **Figure 6**, separated to panel A and B, respectively.

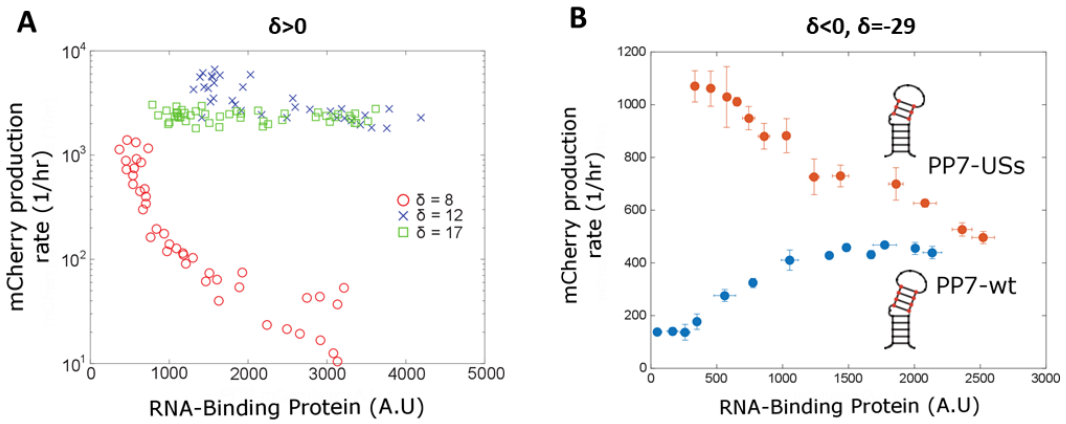

Figure 6: *Dose-response functions for the RNA-binding protein PP7 with a reporter mRNA encoding PP7-wt.*

(A) *PP7-wt binding site was located downstream to the AUG (δ>0) at three positions: δ=8 (red), δ=12 (blue) and δ=17 (green) nt.* ***(B)*** *Dose-response functions for two strains containing the PP7-wt (blue) and PP7-USs (red) binding sites at δ=-29 nt from the AUG. Each data point is an average over multiple mCerulean and mCherry measurements taken at a given inducer concentration.*

In order to understand the observed regulatory phenomena from a structural perspective, we chose these representative constructs for further investigation using SHAPE-Seq, a method to study RNA secondary structures.

4.1.1. SHAPE-Seq on 5S rRNA (control)

We first applied SHAPE-Seq to ribosomal 5S rRNA both *in vivo* and *in vitro* as a control that the protocol was producing reliable results. We analyzed the SHAPE-Seq read count by computing the "reactivity" of each base corresponding to the propensity of that base to be modified by NAI (SHAPE reagent). Bases that are highly modified or "reactive" are more likely to be free from interactions (*e.g.* secondary, tertiary, RBP-based, etc.) and thus remain single stranded. We plot in **Figure 7A** the reactivity analysis for 5S rRNA both *in vitro* and *in vivo*. The data shows that for the *in vitro* sample (red signal) distinct peaks of high reactivity can be detected at positions which align with single stranded segments of the known 5S rRNA[72,88].

By contrast, the *in vivo* reactivity data (blue line) is less modified on average and especially in the central part of the molecule, which is consistent with these regions being protected by the larger ribosome structure in which the 5S rRNA is embedded. The reactivity scores obtained here for both the *in vitro* and *in vivo* samples (**Figure 7B**) are comparable to previously published 5S-rRNA reactivity analysis[72,88].

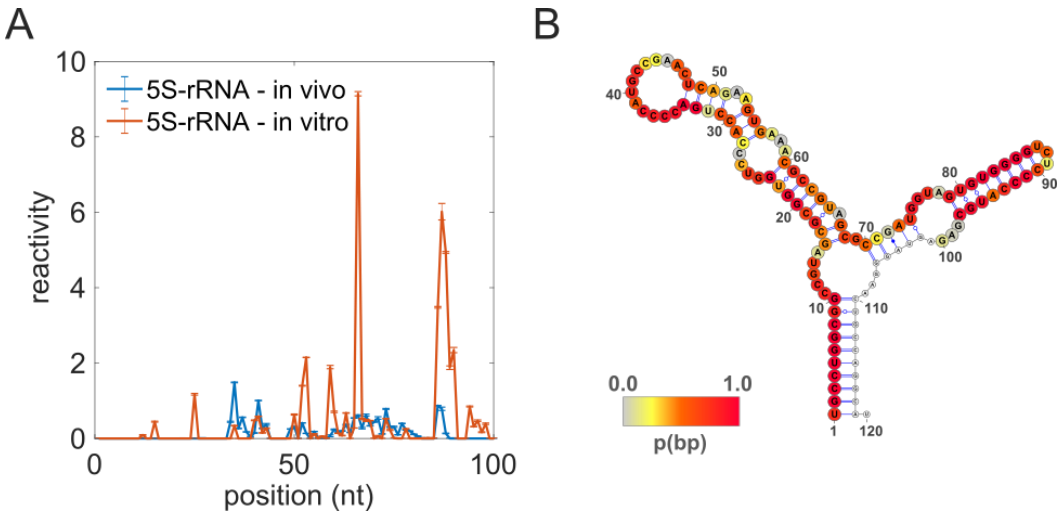

Figure 7: *5S-rRNA control.*

(A) *Reactivity scores for in vivo (blue) and in vitro (red) SHAPE-Seq measurements of 5S-rRNA.* ***(B)*** *5S-rRNA base-paring probabilities were calculated using RNAfold and RNApvmin (by using the in vitro SHAPE-Seq data ad constraints) and overlaid as heat-map for each nucleotide on the known 5S-rRNA structure.*

4.1.2. Binding-site positioned in the ribosomal initiation region (δ>0)

First, I studied the PP7-wt δ=6 construct, where the binding site is positioned 6 nucleotides downstream to the AUG. This construct exhibited a strong repression effect in response to the PCP binding (see plot A in

Figure 6). SHAPE-Seq experiment was carried on this mRNA both *in vitro* and *in vivo* to generate a comprehensive observation into the molecular structure and interaction of the mRNA in the presence and absence of the protein.

4.1.2.1. *In vitro* SHAPE-Seq with recombinant protein

In attempt to uncover the underlying structure of the PP7-wt δ=6 construct and its resulted regulation effect controlled by the PCP, we sought to investigate the *in vitro* RNA structures in the presence and absence of the PCP, similarly to *in*

vivo induction. To do this, I developed an extension to the *in vitro* SHAPE-Seq protocol by adding a purified recombinant protein to the SHAPE reaction buffer after refolding of the RNA. Hence, the RNA is being modified while bound to the PCP, and the binding signature can be captured to enable us better observation on the RNA regions protected by the protein.

Presented in **Figure 8A** are the results for the reactivity analysis carried out on the *in vitro* SHAPE-Seq data for the PP7-wt δ=6 construct with (red line, +RBP) and without (blue line, -RBP) the presence of a recombinant PCP (PP7 phage-coat protein) in the reaction solution. Reactivities are presented as a running average over a 10 nt window to eliminate high frequency noise.

The plot shows that for the –RBP case (blue line) the reactivity pattern is a varying function of nucleotide position, reflecting a footprint of some underlying structure. Namely, the segments that are reactive (*e.g.* -20 to 40 nt range), and those which are not (*e.g.* 110-140 nt range), indicate non-interacting and highly sequestered nucleotides, respectively.

With the addition of the RBP (red line), the reactivity level in the -50 to 80 nt range is predominantly 0 over that range.

Indicated in gray shades are statistically significant differences between the reactivity signals of samples, as determined by Z-factor analysis. We can observe such segments span a range of ~±50 nt from the position of the binding site.

4.1.2.2. *In-vivo* SHAPE-Seq

Next, to provide an insight into the regulatory phenomenon, we studied the PP7-wt δ=6 construct *in vivo*. The SHAPE-Seq experiments were carried on at two induction states (**Figure 8B**): 0 nM of C_4-HSL (blue line - *i.e.*, no PCP-mCerulean present), and 250 nM of C_4-HSL (red line – PCP-mCerulean fully induced). The experiments for both conditions were carried in duplicates on different days. To ensure that a proper comparison between the two induction states was carried out, we first checked that the RNA levels at both states were the same using quantitative PCR (**Figure 8B**-inset).

We plot in **Figure 8B** the reactivity results for both the induced (red) and non-induced (blue) cases. For the non-induced case, we observe a strong reactivity signal (>0.5) over the range spanning -45-110 nt, which diminishes to no

reactivity for positions > 110. This picture is flipped for the induced case, displaying lower- or no-reactivity for the -40 to 110 nt range and a sharp increase in reactivity for positions > 130 nt. Next, we computed the Z-factor for the regions where the differences between the two reactivity signals was statistically significant (Z>0). In the plot, we marked in gray shades the region where the non-induced reactivity was significantly larger than the induced-reactivity. This shaded region flanks the binding site by ~50 nt both upstream and downstream.

A closer examination of the *in vivo* SHAPE-Seq data reveals two major differences from the *in vitro* SHAPE-Seq. First, the non-induced case generates significantly higher values of reactivity in the -50-110 nt range as compared with the –RBP *in vitro* case. Second, while in the *in vitro* experiments no significant difference was found between the – and +RBP cases over the 80-180 range, in the *in vivo* case a significant difference was observed. In particular, the non-induced signal becomes sharply non-reactive over this range. To gain a structural perspective for the extent of these differences, we plot in **Figure 8C** two structures. The structures were computed using RNAfold[89] for the sequence of this molecule and overlaid by its *in vivo* non-induced (left structure) or induced (right structure) reactivity scores (depicted by a heat-map). We demark the RBS (orange oval), PP7-wt binding site (purple oval), and the putative RBP-protected region computed via Z-factor analysis (gray circle on right structure).

Consequently, the SHAPE-Seq analysis *in vivo* reveals significant structural differences between the induced and non-induced cases that are consistent with their RBP-bound states, resultant translational level, and the observed post-transcriptional repression. Furthermore, a comparison between the *in vitro* and *in vivo* SHAPE signal in the presence of the PCP (red: +RBP/induced) show little to no difference between these 2 cases. On the other hand, when the PCP is absent (blue: -RBP/non-induced) we see significantly higher signal *in vivo*, implying on the destabilizing effect of a translationally active ribosome on the mRNA secondary structure.

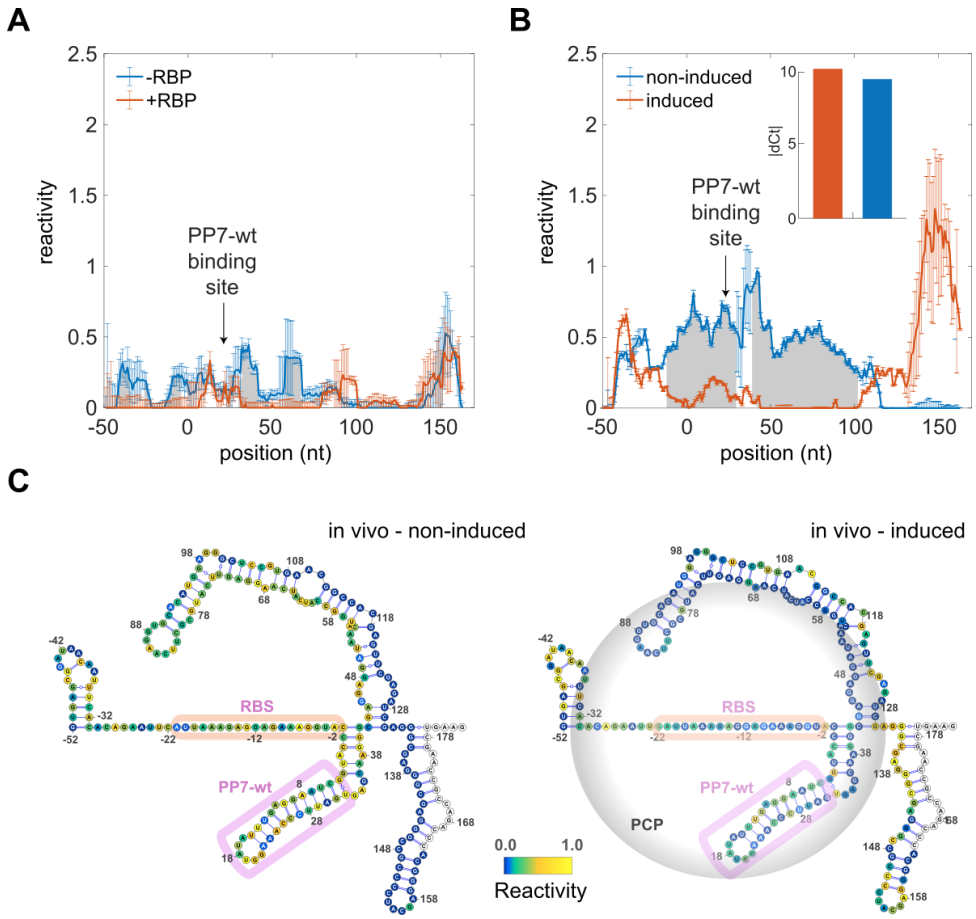

Figure 8: *SHAPE-Seq analysis of the PP7-wt binding site in the absence and in the presence of RBP.*

(A) *In vitro reactivity. Scores for the SHAPE-Seq reactions carried out on refolded mCherry reporter mRNA molecules containing a PP7-wt binding site at δ = 6 with (red) and without (blue) a recombinant PCP present in the reaction buffer.* ***(B)*** *In vivo reactivity. Scores for the SHAPE-Seq reactions carried out in vivo on the PP7-wt δ = 6 construct with the PCP-mCerulean protein non-induced (blue) or induced (red). For both A and B panels, gray shades signify segments of RNA where a statistically significant difference in reactivity scores (as computed by a Z-factor analysis) was detected between the +RBP and −RBP (A), and induced and non-induced (B) cases, respectively. Error bars were computed using boot-strap resampling and subsequent averaging over two biological replicates.* ***(C)*** *Structural schematics of the segment of the PP7-wt δ = 6 construct that was subjected to SHAPE-Seq in vitro. The structures are overlaid by the reactivity scores (represented as heatmaps from blue, low reactivity, to yellow, high reactivity) for the non-induced (left) and induced (right) cases, respectively. Binding site and RBS are highlighted magenta and orange ovals, respectively. Gray circle in right structure corresponds to the range of protection by a bound RBP. Non-colored bases correspond to position of the reverse transcriptase primer.*

4.1.3. Binding-site positioned in the 5' UTR (δ<0)

Next, I proceeded to study the different regulatory effects observed in two similar mRNA constructs, PP7-wt and PP7-USs, differ in only one base-pair in the upper stem of the hairpin. As presented in

Figure 6, the mCherry expression of the PP7-USs mRNA strain is being down-regulated while the PP7-wt strain is surprisingly up-regulated. Moreover, the expression level of the two strains is converging upon binding of the PCP. Aiming to find the underlying mechanism behind these two different regulatory effects, I carried SHAPE-Seq experiments both *in vitro* and *in vivo* to look on the structural features of each strain.

4.1.3.1. *In vitro* SHAPE-Seq

In order to unravel the connection between the structure of the 5' UTR and resultant dose-response functions, we subjugated the PP7-wt and PP7-USs constructs at δ=-29 to SHAPE-Seq *in vitro*. We chose to modify a segment that includes the entire 5' UTR, and in addition another ~140 nt of the mCherry reporter gene. We hypothesized that SHAPE-Seq data can provide a foot-print or echo for the mRNA structure in the 5' UTR as it did for the ribosomal initiation region with and without a bound RBP.

In **Figure 9A** we plot the reactivity signals as a function of nucleotide obtained for both the PP7-wt (blue line) and PP7-USs (red line) constructs at δ=-29 using *in vitro* SHAPE-Seq. The reactivity of each base corresponds to the propensity of that base to be modified by NAI. For each data-point in the plots, error-bars are computed from two biological replicates for each variant, and additional boot-strapping analysis.

Since the two constructs differ by a deletion of two nucleotides at positions -45 and -38, we reasoned that in order to facilitate a proper alignment between the PP7-USs and PP7-wt reactivity scores downstream to the binding sites, the reactivities at those positions should be omitted from the plot (**Figure 9A**). When doing so both *in vitro* reactivity signals look nearly identical for the entire modified segment of the RNA. This is further confirmed by Z-factor analysis (lower panel), which only yields significant distinguishability for a narrow segment within the coding region (~+30 nt).

Next, we used the *in vitro* reactivity data to guide the computational prediction of the RNA structure. It was shown previously[90–93] that using experimental constraints for RNA, 2D structure computation can increase the similarity of the predicted to the solved structure. Therefore, the free-energy minimizing structure that results is different from the one that would be obtained from computations that are based on the sequence alone.

In **Figure 9B** we plot the structures for both variants, as computed using constraints from the *in vitro* SHAPE-Seq data. Examination of the computed structures reveals two 5' UTR features that consistently appear. The first corresponds to the binding site (-56 to -30) as expected, while the second corresponds to a downstream satellite structure (-23 to -10). The secondary hairpin encodes a putative short anti-Shine-Dalgarno (aSD) motif (CUCUU) [94], which may partially sequester the RBS. While RBS-sequestration by an aSD motif can explain the up-regulation effect observed for PP7-wt, it cannot at the same time explain the down-regulatory phenomenon observed for PP7-USs, nor its high basal production rate levels.

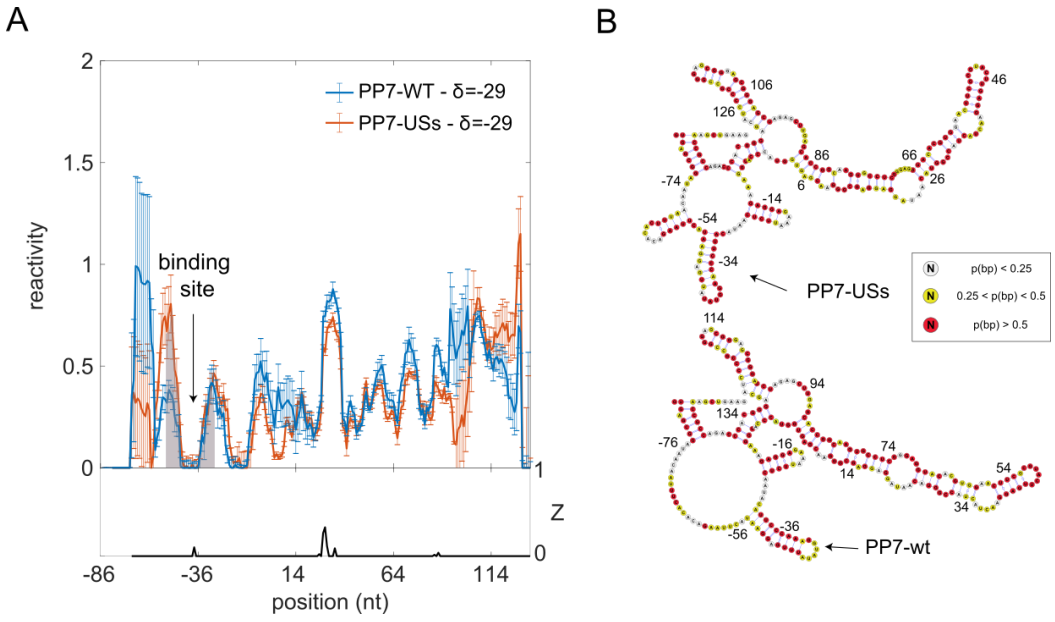

Figure 9: *in vitro SHAPE-Seq analysis for PP7-wt and PP7-USs strains.*

(A) *In vitro reactivity analysis for SHAPE-Seq data obtained for two constructs PP7-wt (blue) and PP7-USs (red) at δ=-29. Error-bars are computed by using boot-strapping re-sampling of the original modified and non-modified libraries for each strain and also averaged from two biological replicates.* ***(B)*** *Inferred in vitro structures for both constructs and constrained by the reactivity scores from (A). Each base is colored by its base pairing probability (red-high, yellow-intermediate, and white-low) calculated based on the structural ensemble.*

4.1.3.2. *In-vivo* SHAPE-Seq

We proceeded to carry out the SHAPE-Seq protocol *in vivo* on induced and non-induced samples for both the PP7-wt and PP7-USs δ=-29 variants. We used biological duplicates for every variant/induction level pair. In **Figure 10A**, we plot the non-induced (RBP-) reactivity obtained for PP7-wt (blue) and PP7-USs (red). The data shows that PP7-USs is more reactive across nearly the entire segment, including all of the 5' UTR and >50 nt into the coding region. Z-factor analysis reveals that this difference is statistically significant for a large portion of the 5' UTR and the coding region, suggesting that the PP7-USs is overall more reactive and thus less structured than the PP7-wt fragment. Alternatively, in **Figure 10B** we show that in the induced state (RBP+) both constructs exhibit a weaker reactivity signal that is statistically indistinguishable in the 5' UTR (*i.e.* Z-factor ~0). Moreover, the region associated with the binding site is unreactive (marked in grey), while both the adjacent upstream and downstream regions exhibit a moderate reactivity signal.

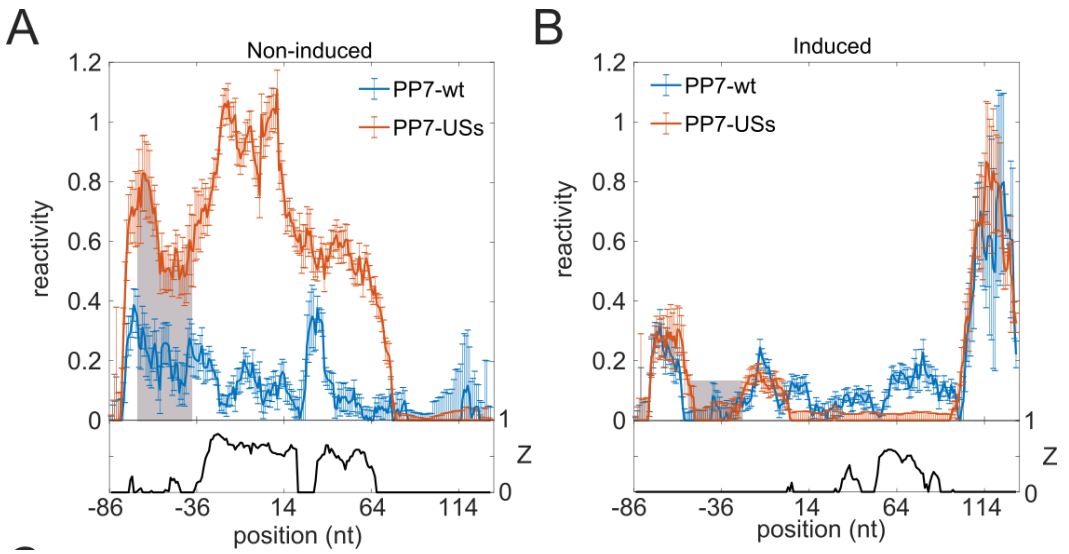

Figure 10: *in vivo SHAPE-Seq analysis for PP7-wt and PP7-USs strains.*

(A-B) *Comparison of reactivity analysis computed using in vivo SHAPE-Seq data for the non-induced (A) and induced (B) states of PP7-wt (blue) and PP7-USs (red) at δ=-29. Error-bars are computed by using boot-strapping re-sampling of the original modified and non-modified libraries for each strain, and also averaged from two biological replicates.*

To further explore the reactivity signal of the 5' UTR in the induced cases, we plot the induced versus non-induced reactivities for each construct (**Figure 11**). The plots reveal that for the PP7-wt construct (**Figure 11A**), the binding site location coincides with a statistically distinguishable protected region that becomes non-reactive upon induction. For PP7-USs (**Figure 11B**), no such identification can be made due to the radically different reactivity signals observed for the two states. Taken together, PCP-mCerulean induction seems to trigger structural changes in the mRNA molecules. For PP7-USs, RBP binding likely leads to a moderate re-structuring of the 5' UTR, which in turn triggers reduced translation. Whereas, for the PP7-wt construct a signature for RBP binding can be discerned and taking into account the nearly identical reactivity signal to that of PP7-USs in the induced case a likely structural shift ensues as well.

To provide further evidence for the correlation between translational activity and resultant reactivity signature, we examined the reactivity and gene-expression data for a PP7-wt construct (δ=5) that was positioned in the ribosomal initiation region (**Figure 11C**). When positioned in the ribosomal initiation region locations, there is a moderate level of expression in the absence of PCP induction, and complete repression in the induced state (see **Figure 11C** inset). In this case (PP7-wt-δ=5), the reactivity signature in the non-induced state is similar to what was observed for PP7-USs and radically different from the signature observed for the PP7-wt construct at δ=-29. However, in the induced state a structured reactivity signature is observed, which is similar for all three constructs. Thus, the up-regulating PP7-wt δ=-29 construct can be differentiated by its reactivity signature from the rest of the down-regulating variants consistent with it being non-translated in the non-induced state, as compared with the two translationally active variants.

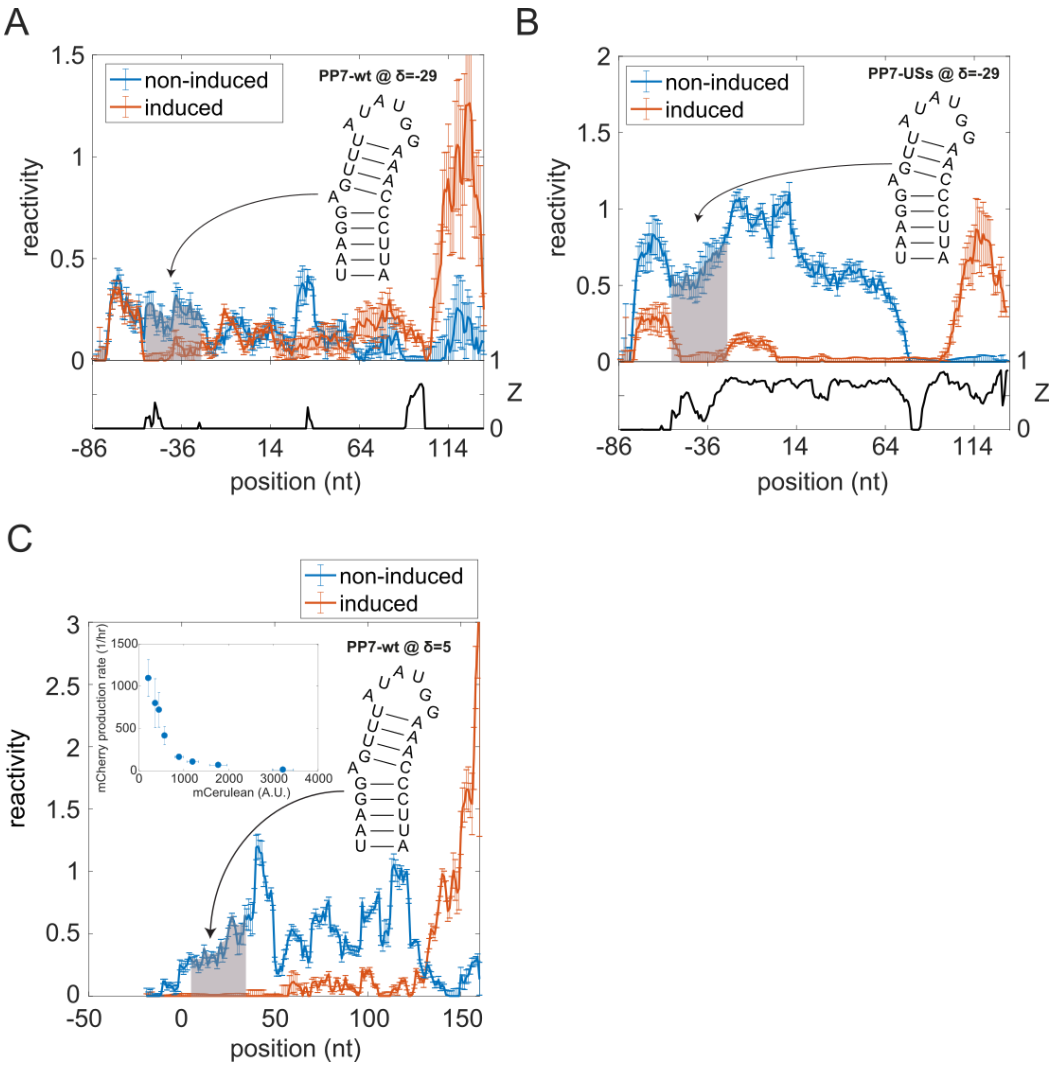

Figure 11: *Induced vs Non-induced plots for PP7-wt and PP7-USs in vivo.*

(A-B) *Plots displaying the reactivities and Z-factor analysis (black) between the non-induced (blue) and induced (orange) strains for PP7-wt (A) and PP7-USs (B). Note the massive difference between the non-induced and induced states of PP7-USs in comparison to PP7-wt where only a small difference is observed in the vicinity of the binding site.* ***(C)*** *Plot comparing the non-induced (blue) to induced (orange) reactivity signals for PP7-wt when positioned at the ribosomal initiation region (δ=5) (Insets) Dose response plotted as mCherry production rate vs mCerulean fluorescence for PP7-wt (δ=5).*

To generate a structural insight, we implemented the constrained structure computation that was used for the *in vitro* samples on the PP7-wt (δ=-29) and PP7-USs (δ=-29) variants. This was done in order to derive structures of the RNA molecules that are consistent with the reactivity data obtained for the different induction states. The structures with nucleotides overlaid by base-pairing probabilities are plotted in **Figure 12**. In the top schema, we plot the derived

PP7-USs non-induced variant, which is non-structured in the 5’ UTR exhibiting a predominantly yellow and white coloring of the individual nucleotide base-pairing probabilities. By contrast, in the PP7-wt non-induced structure (bottom) there are three predicted closely spaced smaller hairpins that span from -60 to -10 that are predominantly colored by yellow and red except in the predicted loop regions. Both top and bottom structures are markedly different from the *in vitro* structures (**Figure 9B**). Neither displays the PP7-wt or PP7-USs binding site, and the secondary aSD hairpin only appears in the PP7-wt non-induced strain. In the induced state, a structure reminiscent of the *in vitro* structure is recovered for both variants with three distinct structural features visible in the 5’ UTR: the upstream flanking hairpin, the binding site, and downstream CUCUU anti-Shine Dalgarno hairpin. These variety of predicted structures for each state *in vivo* suggests that the level of translation may be mostly dependent on a particular arrangement of sub-structures in the 5’ UTR, and to a lesser extent on the presence of the aSD motif.

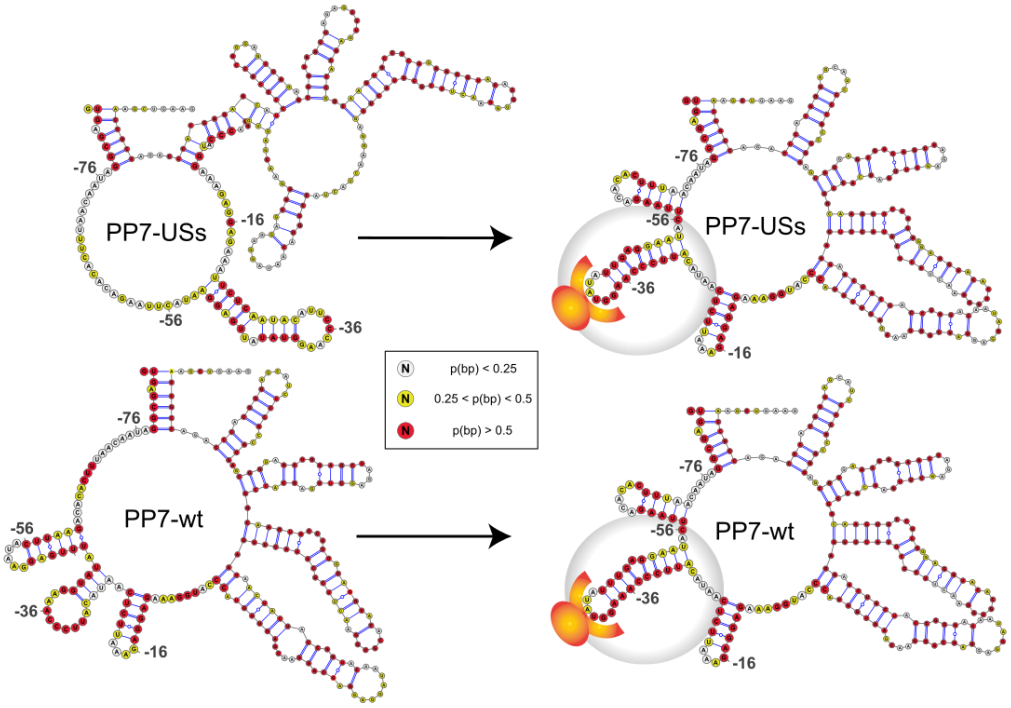

Figure 12: *predicted structures of PP7-wt and PP7-USs strains in vivo combined with SHAPE-Seq reactivity scores.*

Inferred in vivo structures for all 4 constructs and constrained by the reactivity scores (shown in ***Figure 10****). Each base is colored by its base pairing probability (red-high, yellow-intermediate, and white-low) calculated based on the structural ensemble. For both the PP7-wt and PP7-USs the inferred structures show a distinct structural change in the 5’ UTR as a result of induction of the RBP.*

4.2. Engineering regulatory synthetic long non-coding RNA

In this part of the thesis I will describe the experimental results and characterization of the synthetic lncRNA library and the corresponding screening system. To optimize the functionality of the system I first tested different activation domain to choose the strongest one to be integrated into the final design. Next, I also calibrated the transfection conditions for CHO cells to facilitate DNA transfection during the random genomic integration of the DNA reporter construct and the subsequent activation experiment with the DNA-binding activator (rTetR-P65-HSF1). Additionally, characterization of the RBP fusions expression and functionality was performed. Lastly, I tested the slncRNA library sequences and the efficiency of the HAC-based integration using a reporter plasmid with GFP, as was described elsewhere.

4.2.1. Screening transcription activation domains

For development of robust reporting system, a strong transcription activation response is required, thus I screened different activation domains (ADs), while the most potent one will be integrated into the final design. Screening of ADs was carried out by fusing the ADs to rTetR DNA-binding protein and testing the transcription activation effect of each on a reporter plasmid consisting of an mCherry gene under a TRE promoter (tetO operator and minimal CMV promoter). Upon induction of doxycycline (tetracycline analog), rTetR-AD binds the DNA and activates transcription of mCherry. Four ADs were tested: VP64 (tetramer of VP16), p300, P65-HSF1 and VPR (VP64-P65-Rta).

HEK293 cells were co-transfected with pTRE-mCherry and rTetR-AD plasmids and after 24 hr induction with doxycycline mCherry fluorescence was measured by flow-cytometry. Fold-change of mCherry activation (**Figure 13**) was calculated with respect to a control sample transfected with rTetR lacking AD.

Across all ADs tested in this experiment, no significant activation was observed at both 0 and 10 ng/mL Doxycycline levels, with fold-change range of 2 to 4. However, induction of 1000 ng/mL Doxycycline led to strong activation of mCherry expression, which varied from 3-fold with VP64, 5-fold with P300, 7-fold with VPR and 13-fold with P65-HSF1.

A closer look at the maximum induction level (1000 ng/mL) shows that the highest activation was observed with the synthetic transactivators VPR and P65-

HSF1 with up to 7 and 13-fold increase in mCherry fluorescence, respectively. Finally, the chosen AD for further research was P65-HSF1 with the best performance in this experiment.

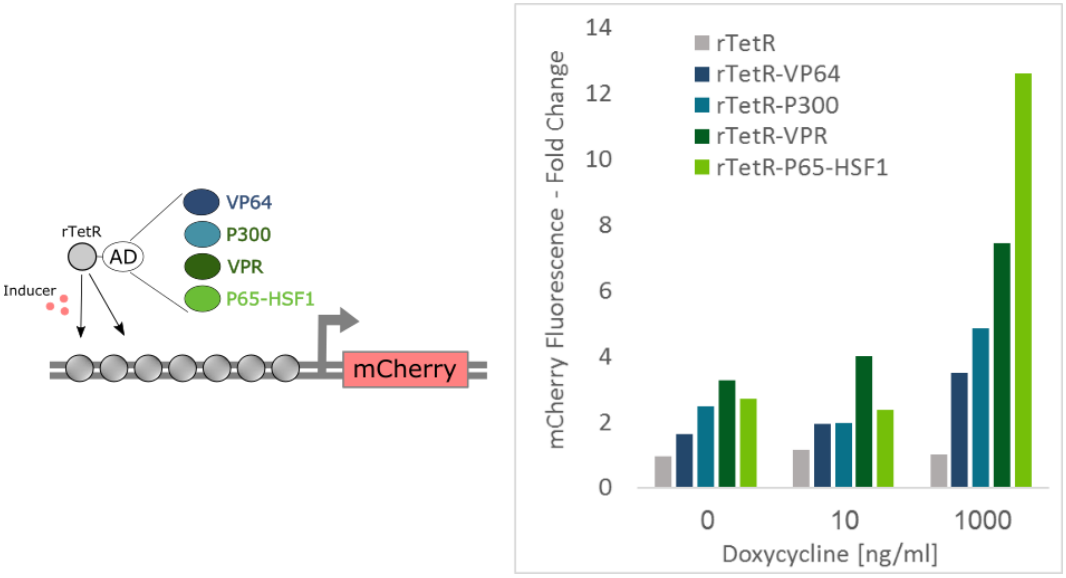

Figure 13: *Fold-change in mCherry expression by rTetR-AD in 3 induction states.*

HEK293 cells were transfected with pTRE-mCherry plasmid and one of the rTetR-AD variants (AD: VP64, P300, VPR or P65-HSF1). Subsequently, transfected cells were induced with Doxycycline (tetracycline analog) and mCherry fluorescence was measured using flow cytometry. Fold-change of mCherry expression was calculated relative to a control sample of cells transfected with the rTetR protein only (grey bar). Since rTetR can bind its cognate DNA site in the presence of Doxycycline, we observe activation of mCherry expression in the higher induction level (1000 ng/mL) with all 4 fusions, while the rTetR-P65-HSF1 fusion provides the strongest activation effect with up to 13-fold induction (dark blue bar).

4.2.2. Transfection calibration of CHO cells

Since the final goal of my study is to integrate an oligo library into the human artificial chromosome (HAC) of CHO cells, I had to optimize the transfection conditions for this cell-line to obtain better transfection efficiencies. The transfection strategies tested in this experiment were based either on the commercial transfection protocols of PolyJet (SignaGen) and Lipofectamine 2000 (Thermo Fisher Scientific), or on literature protocols using the PEI reagent (PolyEthyleneImine)[95,96]. Additionally, I tested the optimal time period from transfection to FACS analysis (24, 48 or 96 hr).

CHO cells were transfected with pCMV-mKate plasmid, known for its strong fluorescence, and were then analyzed by flow-cytometry (FACS). To evaluate the transfection efficiency I calculated the weighted fluorescence intensity by multiplying the percent of fluorescent cells with the median mKate intensity measured in these cells (**Figure 14**). This calculation has been previously described to represent a valid method to quantify fluorescent levels[97].

For the conditions tested in this experiment, I observed a distinct advantage for using the commercial reagents, PolyJet and Lipofectamine, with significant higher transfection efficiencies with PolyJet. In contrast, PEI transfections presented very pour mKate intensities.

Using PolyJet transfection, the best mKate measurement was after 48 hr from transfection, while for Lipofectamin 2000 it is recommended to read the samples 24 hr after transfection.

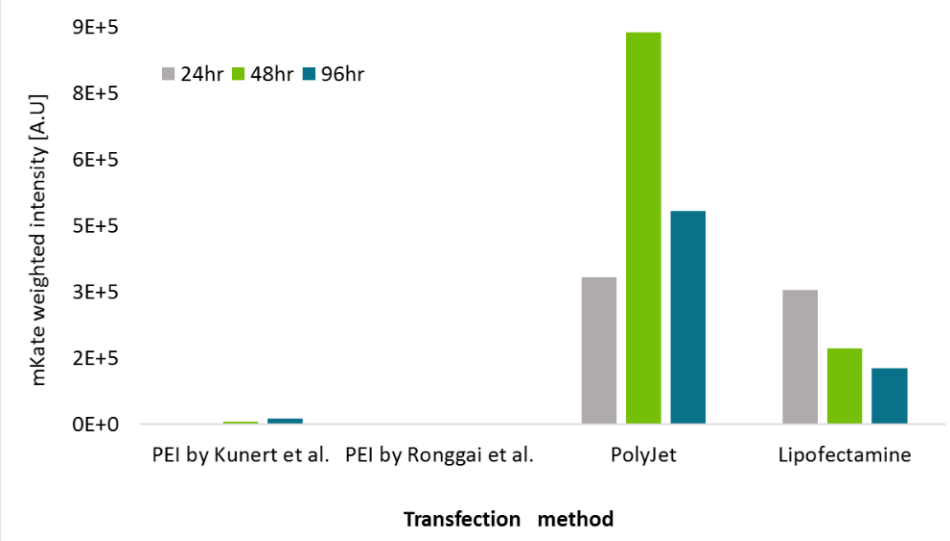

Figure 14: *Calibration of transfection conditions for CHO cells.*

CHO cells were transfected with pCMV-mKate plasmid to evaluate 4 different transfection protocols. Cells were transfected with either the transfection reagent PEI (PolyEthyleneImine) according to literature methods[95,96] or using the commercial transfection reagents PolyJet and Lipofectamine. Additionally, cells were analyzed 24 hr, 48 hr and 96 hr post transfection to assess the optimal time for protein expression. mKate weighted intensity was calculated by multiplying the percentages of positive mKate cells by the median value of mKate fluorescence. Transfection with PEI show poor mKate expression while using PolyJet and Lipofectamine resulted in significantly higher expression. Among transfection method tested, PolyJet produced the best mKate expression after 48 hr.

4.2.3. mCherry activation in stable CHO-mCherry cells

Next, I proceeded to construct the cell lines encoding the slncRNA functionality. First, pTRE-mCherry construct was randomly integrated into the genome of CHO cells using Blasticidin pressure to select the resistant clones. Since the pTRE-mCherry construct consists of minimal CMV promoter, I expected to see low mCherry expression levels when the cells are not induced (basal level). In practice, the genomic integration resulted in rather diverse cell population with 3 observed fluorescence peaks of low and high mCherry (at $\sim 10^2$ and 10^4 respectively), and a sub-population of intermediate intensity ($\sim 10^3$) (**Figure 15**). The low mCherry population overlaps the auto-fluorescence observed for native CHO cells, but may also include cells with the pTRE-mCherry construct expressed in very low basal levels.

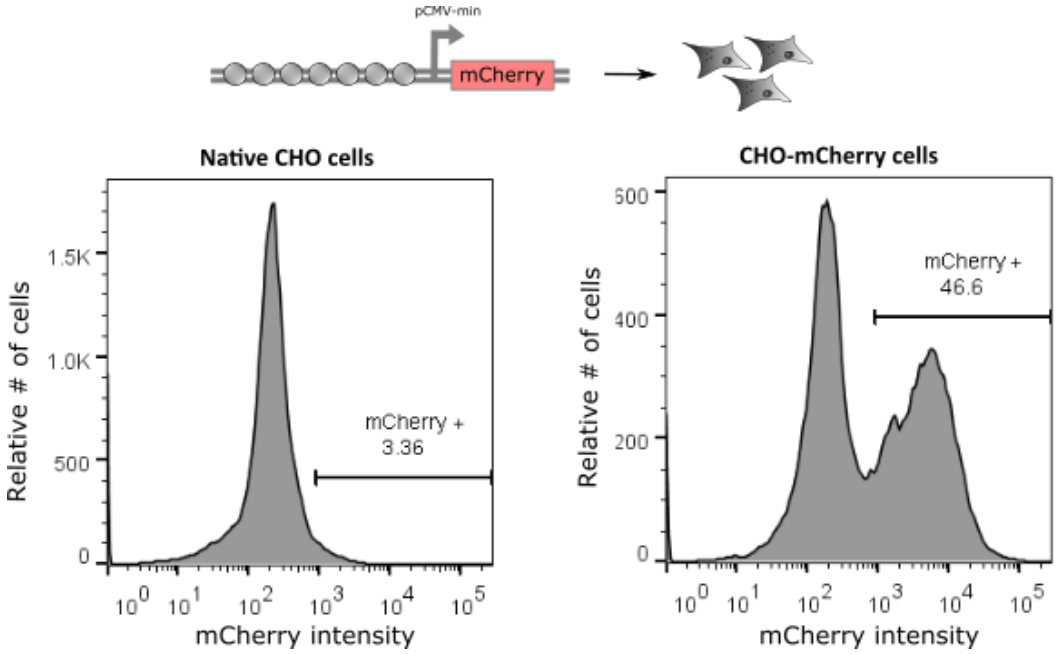

Figure 15: *mCherry intensity distribution in CHO-mCherry cell-line after random integration of the reporter construct.*

pTRE-mCherry was randomly integrated into the genome of CHO cells. After 3 weeks of Blasticidin selection, cells were pooled and analyzed by flow cytometry for mCherry measurement. CHO-mCherry population present 3 peaks of low, medium and high mCherry levels. As a control, native CHO cells before integration was analyzed as well (left panel). CHO-mCherry cells were subsequently subjected to sorting according to mCherry level.

To further improve the reporter cell-line, I sought to establish a unified cell-line characterized by low basal mCherry levels on one hand, and a strong response to induced activation on the other hand. Therefore, I started from selection of cells demonstrating low mCherry levels, which will later be tested for their mCherry activation response. I used a cell sorter to select and isolate single cells with low mCherry basal levels into 96-well plate (one cell per well), and then cultured and expanded them in their wells for approximately 1 month to obtain enough cells for further experiments. Out of 96 single clones that were initially collected, only 18 clones endured this procedure.

The 18 clones were then subjected to an activation experiment using rTetR-P65-HSF1 in order to select a single clone demonstrating strong response to induced activation. Transfection of transactivator was supported by YFP expression (**Figure 16A**), encoded downstream to the rTetR-P65-HSF1 fusion. Most clones didn't respond to activation (not shown), with constant negative mCherry rates across the different induction levels. Only 2 clones, B4 and C1, showed mCherry activation upon induction as shown in **Figure 16**. For both clones the basal level of mCherry expression is ~5000 A.U and there is a correlation between induction and mCherry activation. However, a closer look reveals almost 2 orders of magnitude higher mCherry intensities for clone B4 (**Figure 16B**) at induction levels of 100-1000 ng/mL Doxycycline, in comparison to the corresponding results for clone C1 (**Figure 16C**). Additionally, clone B4 show 100-fold change between induction levels of 0 to 1000 ng/mL, while the fold induction measured for clone C1 is only ~1.5. Consequently, clone B4 was selected as the CHO-mCherry cell-line.

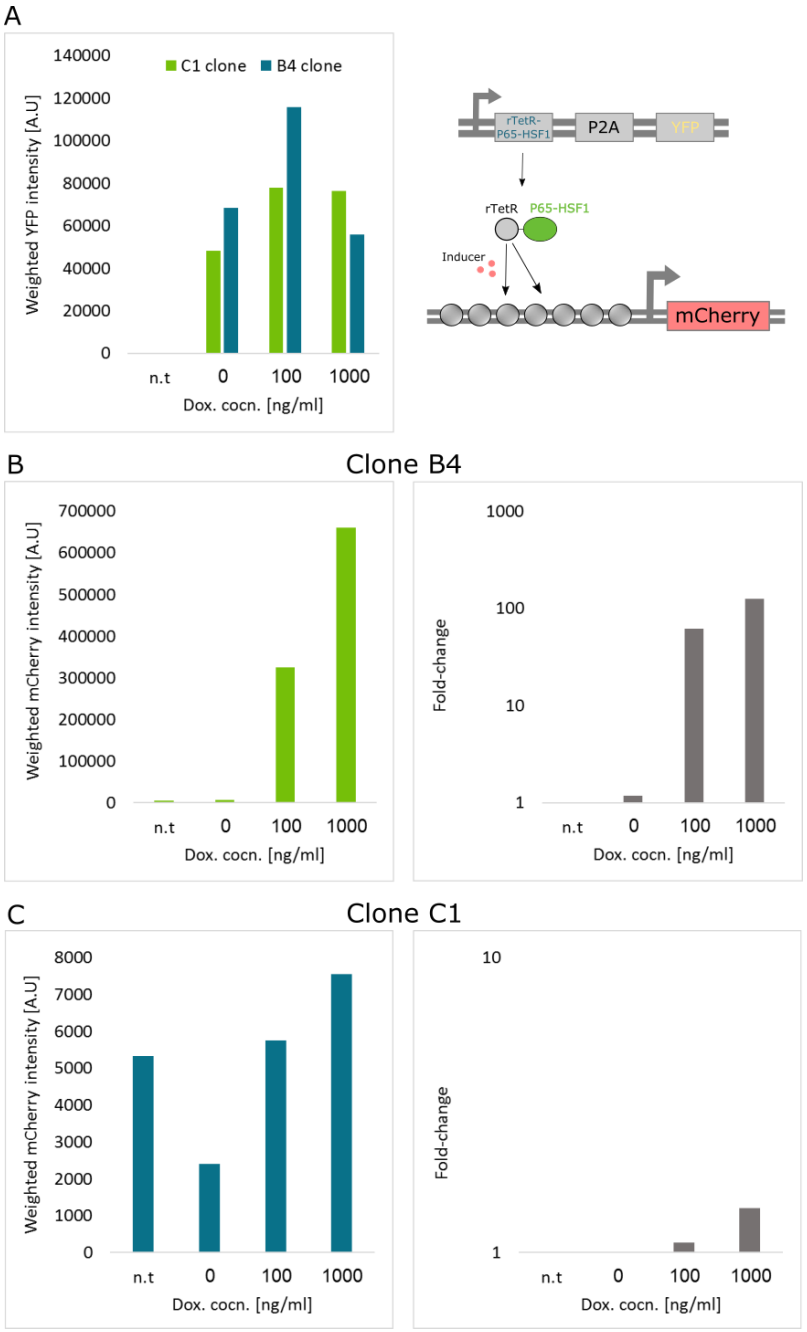

Figure 16: *Flow cytometry analysis of selected CHO-mCherry clones, B4 and C1.*

Sorted CHO-mCherry single clones were transfected with a plasmid encoding for rTetR-P65-HSF1 fusion and YFP, both are constantly expressed from the ubC promoter (see ***Figure 4:*** *rTetR-AD construct for activation domains screening.. Cells were induced with Doxycycline for transcription activation of mCherry. Non-transfected (n.t) control was carried to evaluate the mCherry basal levels. Among 18 clones tested (not shown), only two clones, named B4 and C1, respond to activation.* ***(A)*** *YFP was used as transfection marker, its fluorescence is constant and independent of Doxycycline induction.* ***(B-C)*** *mCherry intensities of clones B4 and C1 in 3 induction levels. mCherry basal level in the non-transfected (n.t) control is similar for both B4 and C1 clones and equals to ~5000 A.U Fluorescence intensities of clone B4 are higher in almost 2 magnitudes of order in comparison to clone C1.*

4.2.3.1. Choosing induction levels for selected CHO-mCherry cells

Following the selection of clone B4 as the final CHO-mCherry cell-line, I ran additional experiments to determine the suitable induction levels.

The CHO-mCherry cells were again transfected with rTetR-P65-HSF1 and induced in 4 induction levels. First, I tried induction levels of 0, 10, 100 and 1000 ng/mL doxycycline (**Figure 17A**), which resulted in only 3 distinct levels (0, 10 and 100 ng/mL), while induction level of 1000 ng/mL responded similarly to 100 ng/mL. In the second experiment, I tested induction levels of 0, 1, 10 and 100 ng/mL (**Figure 17B**), showing that induction levels of 0 and 1 ng/mL were indistinguishable, and 10-100 ng/mL acting as expected.

As a consequence, three induction levels of 0, 10, 100 ng/mL are sufficient to see mCherry activation in this particular experimental setup.

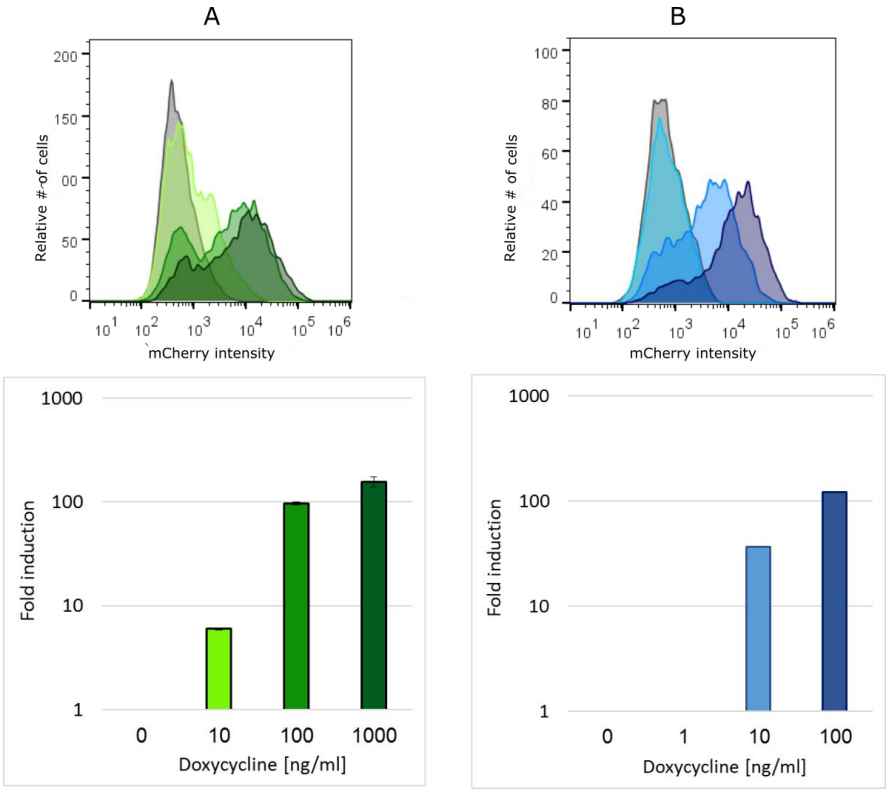

Figure 17: *Comparison of mCherry fold-change in different induction levels.*

CHO-mCherry cells were transfected with rTetR-P65-HSF1, induced by Doxycycline (Dox) and analyzed in flow cytometry for mCherry intensity measurement. When induction levels of 0, 10, 100 and 1000 ng/mL Dox were used ***(A)*** *no significant improvement in induction observed in the 1000 ng/mL level relative to 100 ng/mL. Moreover, induction levels of 0, 1, 10 and 100 ng/mL Dox* ***(B)*** *resulted in mCherry activation only in the 10 and 100 levels. It is clear that 3 induction levels of 0, 10 and 100 ng/mL Dox are sufficient for proper investigation in further experiments.*

4.2.4. Examination of the RNA-binding proteins fusions

The synthetic RBP cassette (pUC57-sRBP) was ordered from GenScript as a plasmid ready to use, encoding for both fusion proteins rTetR-PCP-CFP (sDRBP) and MCP-YFP-HSF1 (RBP-AD), separated by the self-cleavage peptide P2A. Hence, the proteins are transcribed in the same mRNA but translated independently into 2 proteins (*i.e* double-cassette). Each protein fusion was rationally designed based on existing building-blocks, with no guarantee the proteins functionality is maintained during the conjugation. Moreover, my system does not have a proper positive control, thus I had to test these fusions by indirect means, such as fluorescent and binding assays, as I will describe below.

4.2.4.1. Expression efficiency of the sRBP fusions in CHO cells

First, I tested the expression efficiency of the fusions in CHO cells by carrying flow-cytometry experiment to measure the fluorescence intensity of CFP and YFP, markers for the expression of the sDRBP and the RBP-AD. The analysis revealed very low fluorescence of both CFP and YFP. To troubleshoot the initial design of the proteins cassette I cloned the double-cassette (rTetR-PCP-CFP & MCP-YFP-HSF1) into its derivatives consisting of the single fusions rTetR-PCP-CFP, MCP-YFP-HSF1, and their most basic RBP fusions PCP-CFP, MCP-YFP. Overall, 4 additional plasmids were generated.

CHO cells were then transfected with each of the 5 plasmid variants, and CFP and YFP fluorescence were measured using flow-cytometry. The collected data was used to calculate the weighted fluorescence intensity of each sample (**Figure 18**). The results show that the intensities of CFP (marked in blue) in the single fusions are higher than the fluorescence measured from the same fusion when expressed from the double-cassette. On the other hand, the intensity of YFP (marked in green) was improved only for the MCP-YFP-HSF1 variant and not for MCP-YFP. Moreover, in the double-cassette sample, higher fluorescence of YFP was observed in comparison to CFP, even though the YFP is encoded downstream to the CFP and the P2A sequence (see **Figure 3B**).

Since the weighted fluorescence across all samples were relatively low, I also plotted (**Figure 18**-inset) the YFP fluorescence measured from another two plasmids encoding for the same YFP expressed from ubC promoter: pubC-YFP and its derivate pubC-rTetR-HSF1(P2A)YFP (see methods section 3.2). No similar

construct for CFP was available. The results revealed that the YFP fluorescence can go as high as $1.3 \cdot 10^6$ A.U in the pubC-rTetR-HSF1(P2A)YFP plasmid, 4 times higher than the YFP measured in the MCP-YFP-HSF1 plasmid (327,840 A.U).

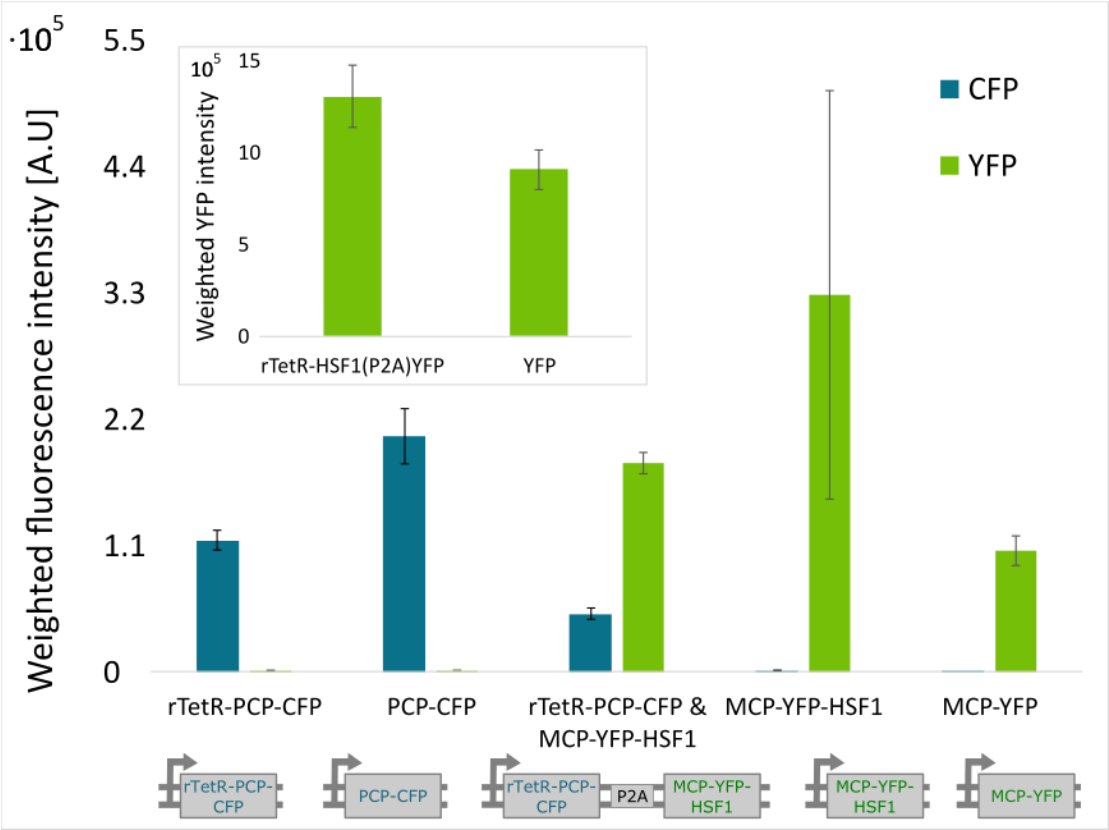

Figure 18: *Flow-cytometry analysis of RNA-binding proteins fusions cassette.*

CHO cells were transfected with either double-cassette encoding for both rTetR-PCP-CFP and MCP-YFP-HSF1 fusion protein or with single fusion plasmid. CFP and YFP fluorescence intensities (blue and green bars, respectively) were measured using flow-cytometry. Error bars obtained from analysis of 2 replicates. ***Inset:*** *YFP weighted fluorescence as measured from plasmids encoding the ubC promoter expressing either the rTetR-HSF1(P2A)YFP or YFP.*

To further understand the differences observed in **Figure 18**, I looked on the raw data of frequency of fluorescent cells and the median intensity of each sample (presented in **Table 3**), and found that while the frequencies are relatively similar across all samples (18.9-32.5%), the intensities measured from the single fusion plasmids (rTetR-PCP-CFP, PCP-CFP, MCP-YFP-HSF1, MCP-YFP) are lower by one order of magnitude (5462, 6790, 11,108, 5586 A.U, respectively), as compared to those of pubC-rTetR-HSF1(P2A)YFP (40,216 A.U) and pubC-YFP (43,978 A.U).

Table 3: *Frequency of cells expressing fluorescence of CFP or YFP and the measured median intensity*

Plasmid name	Frequency of fluorescent cells	Median intensity (A.U)
rTetR-PCP-CFP	20.9%	5462
PCP-CFP	30.2%	6790
MCP-YFP-HSF1	28.5%	11108
MCP-YFP	18.9%	5586
rTetR-HSF1(P2A)YFP	32.5%	40216
YFP	20.9%	43978

Overall, the results indicate that the double-cassette required for the screening system is being successfully transfected into CHO cells as indicated from the frequency of fluorescent cells, but the intensity values indicate low expression efficiencies. Therefore, further characterization and improvement of these parts is necessary.

4.2.4.2. Fusion DNA-binding component binds the tetO sites

Second, I sought to examine whether the DNA-binding component in the rTetR-PCP-CFP fusion retained its ability to bind DNA. To do so, I exploited the theory behind the so called "dominant-negative effect", occurs when a mutant gene product can still interact with the same elements (*e.g.* DNA-binding) as the wild-type product, but lacks some reporting aspect of its function (*e.g.* transcription activation).

In my molecular setup, the “wild-type” product is equivalent to rTetR-P65-HSF1 that can bind the DNA and activate transcription, while the “mutant” proteins are rTetR or rTetR-PCP which can only bind the DNA without activating transcription. Therefore, full occupancy of the tetO binding-sites on pTRE-mCherry plasmid by the transactivator rTetR-P65-HSF1 will lead to maximum transcription activation and mCherry levels. However, when the transactivator is simultaneously expressed with a "mutant" that can bind the DNA but lacks the activation function (such as rTetR or rTetR-PCP) it will result with less mCherry activation. See **Figure 19A** for illustration of the experiment.

CHO-mCherry cells were co-transfected with a combination of 2 plasmids carry one of the rTetR variants (rTetR-P65-HSF1, rTetR-PCP or rTetR) or an empty pUC19 plasmid as control (not marked in the plot). The results presented in **Figure 19B** demonstrate conclusive dominant-negative effect as expected. The proteins rTetR and rTetR-PCP can't activate transcription thus the mCherry intensities in the three bottom samples (rTetR-PCP, rTetR and rTetR-PCP +rTetR) are low and not responsive to Doxycycline induction. In contrary, when the transactivator rTetR-P65-HSF1 is expressed, we observe inducible mCherry activation, with maximum mCherry for rTetR-P65-HSF1 alone and decreased levels when co-expressed with either rTetR or rTetR-PCP.

Consequently, I was able to show that the rTetR-PCP fusion under investigation is able to bind DNA at tetO sequences as desired.

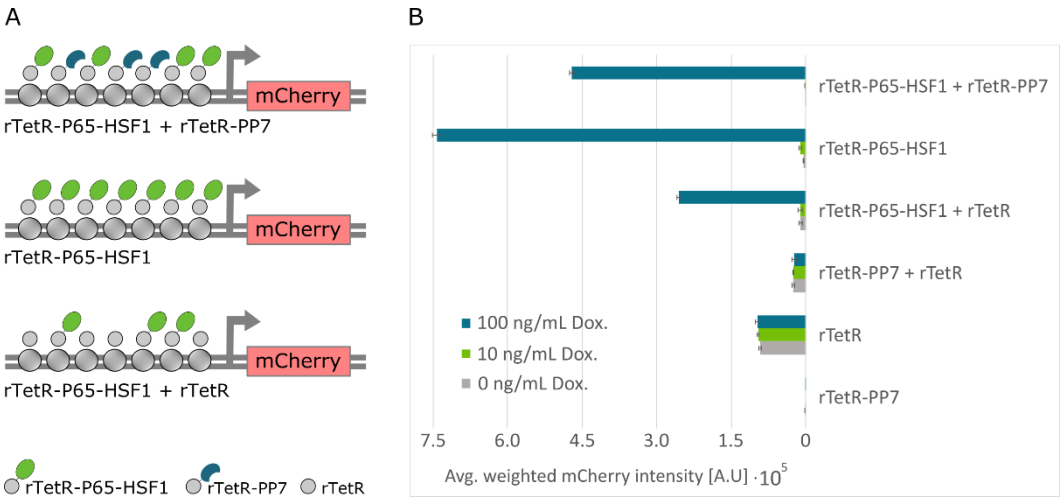

Figure 19: *Validating DNA-binding of rTetR-PCP fusion using the "dominant-negative effect".*

(A) *Illustration of the experiment rational – optional setups for occupancy of tetO binding-sites by rTetR variants (rTetR-P65-HSF1 (grey circle-green oval), rTetR-PCP (grey circle-blue moon) or rTetR (grey circle)). Each setup will result in different mCherry activation level according to the DNA-binding ability of the variants.* ***(B)*** *Experiment results – when the transactivator rTetR-P65-HSF1 is expressed alone (2nd bars from the top) we see maximal activation of mCherry when fully induced (blue bar). In contrast, when the transactivator is co-expressed with one of the rTetR "mutants" (rTetR-PCP or rTetR, 1st and 3rd bars from the top, respectively) we observe decrease in mCherry. When cells were transfected with only rTetR "mutants" (three lower bars) we see low mCherry expression regardless of induction.*

4.2.5. slncRNA library – sequencing and genomic integration

The synthetic long non-coding RNA (slncRNA) library was ordered as an oligo-pool consists of many RNA sequences with variable MS2 RNA-binding sites. During the synthesis and cloning process of the library it is inevitable that some sequences variant will be lost. Therefore, it is very important to follow and evaluate the complexity of the library using deep sequencing.

Ultimately, the slncRNA library intended to integration into the HAC of CHO cells, to ensure each single cell will express only one copy of the RNA library. As preliminary experiment I examined the HAC-based system, which was described elsewhere[98], by using a GFP reporter to evaluate the efficiency of the transfection and the characteristics of the resulted cell population.

4.2.5.1. slncRNA library sequencing post-PCR and post-cloning

The slncRNA library was ordered as an oligo-pool from TWIST bioscience and was first amplified by PCR to generate a double-stranded DNA (dsDNA) library. Subsequently, the dsDNA library was sequenced using NGS (at the Technion Genomic Center, TGC) to compare between the ordered sequences and the actual variants in the library and to facilitate follow-up on the library complexity during the cloning process. Analysis of the sequencing reads was done by MATLAB code to generate the histogram presented in **Figure 20** showing an average of 98.2±50.0 reads per variant. The results indicate that the library distribution is sufficient, although some variants did not appear in the sequencing (bar at 0 reads is approximately 600), which means that either these variants were not synthesized or that the sequencing depth of this run didn't allow us to observe them.

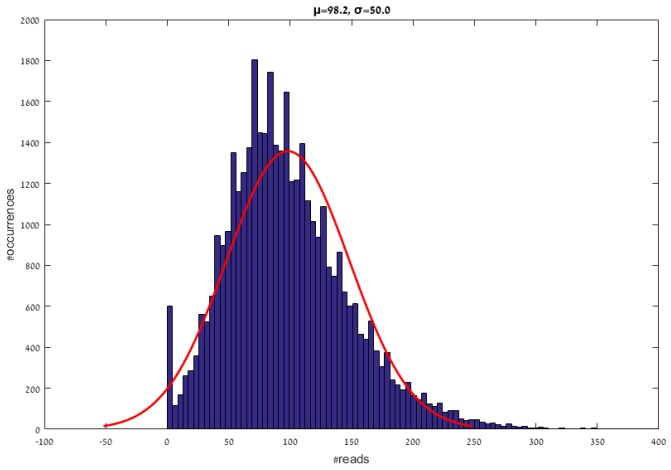

Figure 20: *Post-PCR sequencing of the slncRNA oligo-pool*

Sequencing results were analyzed using MATLAB code. Reads were mapped to the list of variants originally ordered from TWIST and the number of mapped reads per variant was counted to generate the histogram. The average of reads per variant is 98.2 with a standard deviation of 50.0.

Next, the dsDNA library was subjected to digestion with restriction enzymes, cleaning and ligation with the final plasmid (pNeo-attB(ΦC31)-CMV-library-3'box). Subsequently, I transformed the plasmids into E.cloni cells, which were plated on agar plate to form colonies overnight. The plasmids were then purified from the cells using a DNA extraction kit and the library region on the plasmids was amplified using two specific primers adding Illumina overhangs adapter sequences (see **Table 4**). The PCR products were sent to the TGC for micro-Miseq sequencing. The sequencing results revealed that only 174 variants (out of 39,500 originally ordered) were presented in the DNA sample sent to sequencing (results are not shown). It remained unclear what could have led to these results. I speculate that the cloning process caused extreme bias to the slncRNA library complexity, and further examinations are required.

Table 4: *Oligo sequences used for amplifying the library after cloning, adding Illumina overhang adapter sequences (marked regions)*

Primer name	Primer sequence
F-lib.illu.ovhang	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG GACTCTAGGTCATATACCAC
R-lib.illu.ovhang	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG CAACAATTGCATTCATTCCTAG

4.2.5.2. Preliminary examination of genomic integration into the HAC of CHO cells with GFP

Prior to integration of the slncRNA library into the human artificial chromosome (HAC) of CHO-mCherry cells I wanted to examine the ΦC31 recombination efficiency by using the source plasmid of pNeo-attB(ΦC31)-CMV-eGFP encoding for enhanced green fluorescence protein (GFP), as described in the original work of Yamaguchi and Kazuki[87]. GFP plasmid was transfected with the appropriate integrase expression plasmid (pCMV-ΦC31) and cells were selected for 14 days in G418 (Neomycin).

Transfected cells were analyzed by flow cytometry to evaluate the number of GFP expressing clones. The results in **Figure 21** show relatively homogeneous GFP-positive population, comprised 50.5% of the total cell population. The negative control is CHO cells before transfection.

Additionally, control samples with either the GFP or recombinase plasmid alone died upon G418 selection, indicating that the recombinant DNA does not go random integration and that this site-specific integration methodology is very specific.

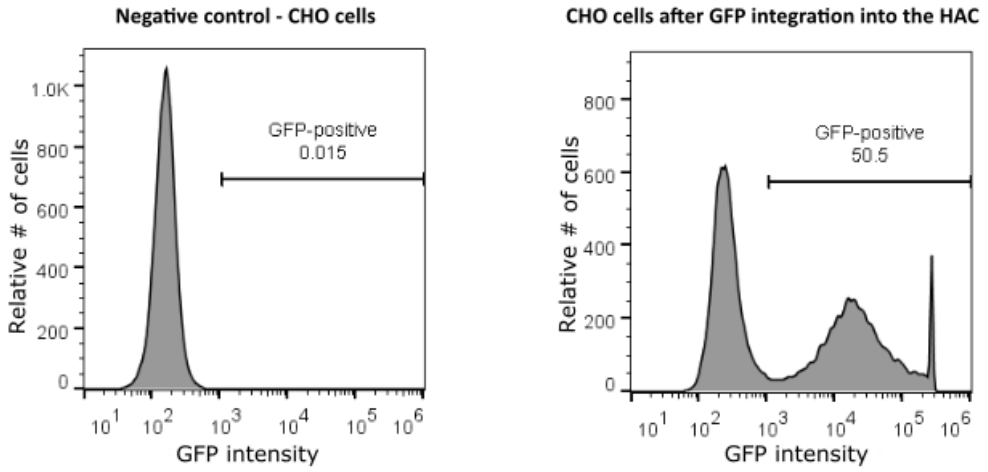

Figure 21: *CHO cell population after GFP integration into the HAC.*

The enhanced green fluorescent protein (eGFP) was inserted into the HAC of CHO cells using ΦC31 recombinase. After 2 weeks of G418 selections the cells were analyzed using flow cytometry which showed that 50.5% of the cells are GFP positive, with relatively homogeneous GFP expression (right panel). For control, CHO cells before integration were also analyzed (left panel).

5. Discussion

5.1. Study RNA structures using SHAPE-Seq

SHAPE-Seq is a relatively new, next generation sequencing approach to probe the structure of an RNA molecule via selective modification of non-interacting nucleotides. By applying SHAPE-Seq analysis on representative mRNA constructs of previous work, we were able to gain further insight on the molecular mechanisms govern the observed translational regulation.

5.1.1. Observation of an extended protected region by PCP

By using our extension for *in vitro* SHAPE-Seq protocol with recombinant protein addition to the RNA sample, we were able to investigate the RNA structures in the presence and absence of the corresponding RBP both *in vitro* and *in vivo*. For both *in vitro* and *in vivo* experiments on PP7-wt δ=6 construct, the analysis revealed that the RBP-binding effect spanned a much wider segment of RNA than previously reported both for phage coat proteins *in vitro*[99] and for other proteins with their cognate RNA target using SHAPE-MaP[73]. There are several scenarios, which may explain this result. In one scenario, PCP may form a large multi-protein complex that is anchored to the binding site, which in turn can lead to a wide protected segment on the RNA. Alternatively, PCP binding may trigger refolding of flanking regions to form structures with fewer non-interacting nucleotides leading to the reduced reactivity result in those regions in the *in vitro* setting.

In the *in vivo* setting a cascade of structural events may be triggered by the refolding or protection of the flanking segments in the immediate vicinity of the binding site. Since these segments include the ribosome binding site, any protection or structuring effect is likely to inhibit initiation and subsequent elongation. This will make the mRNA devoid of ribosomes, which will in turn lead to restructuring of mRNA segments further away from the hairpin resulting in the translationally inactive and highly structured induced state inferred from the reactivity data.

Our newly developed SHAPE-Seq protocol with recombinant protein *in vitro* is an innovative addition to the existing SHAPE methods, which enable broader study of RNA structures when bound to their corresponding RBP. This extended

protocol can strengthen *in vivo* structural observations associated with RNA-protein interaction, and may be applied to similar studies on RNA-protein complexes which require complementary data for *in vitro* settings.

5.1.2. Revealing different structures for PP7-wt and PP7-USs δ=-29 *in vivo*

While translation repression by RBP is a known phenomenon, translation stimulation has never been observed, particularly when the only difference between both constructs is a deletion of only 2 nucleotides in the binding-site sequence.

Integration of the expression level data, SHAPE-Seq data and follow-up structural analysis suggest that a "densely" structured 5' UTR is associated with an inhibited-translation-initiation state. Inhibited translation is alleviated by RBP binding, which seems to stabilize the binding-site hairpin while simultaneously weakening flanking structures in the 5' UTR, leading to translation stimulation. Consequently, the up-regulation phenomenon that we observed is a transition from a strongly-repressing densely structured 5' UTR to a weakly-repressing loosely structured 5' UTR that occurs upon RBP binding.

In the two cases studied in detail here, we demonstrated that upon induction the RBP triggers structural changes in the RNA molecule. This result is not surprising for several reasons: first, the size of the RBPs are comparable to typical structural feature on RNAs, and thus they are likely to affect the stability of nearby structural elements. Second, it is believed that RNA structures fluctuate between closely related ensemble of structures[100–103], and thus binding of an RBP can easily shift the energetic equilibrium of this ensemble leading to a different cumulative translation rate. Third, interaction with the translational machinery can substantially alter the underlying structure.

Others[46,104,105] have shown that mRNAs, which are strongly translated are predominantly non-structured. However, bound RBPs in the 5' UTR near the RBS are likely to hinder translation initiation. This slowdown can, in turn, trigger re-structuring of the RNA molecule leading to a further slowdown of translation, and to a radically different reactivity signature for the RBP bound and unbound states as was observed here.

Consequently, the reactivity data and structural analysis indicates that the deletion of the two nucleotides which encode the PP7-USs binding site together with the translational machinery, are sufficient to trigger large-scale structural changes across the 5' UTR, which in turn lead to the divergent expression levels at the non-induced level. Conversely, the binding of PCP-mCerulean is sufficient for the stabilization of the binding site, which in turn stabilizes the satellite structures in the flanking regions leading to an indistinguishable expression level in the induced states.

5.2. Establishing reporter system for slncRNA engineering

In this part of my work I established and tested the parts constitute the screening assay for functional slncRNA. First, I created a reporting cell-line characterized with strong activation response of a mCherry gene, stably and uniformly expressed from CHO cells genome. Second, I designed the RBP fusions required for the implementation of the system, even though they still require further characterization and possibly optimization.

5.2.1. Efficient gene activation by novel synthetic transactivators

To create a highly efficient and robust reporting system, a strong acting effector is needed. Based on the natural cooperative recruitment process of transcription factors, few studies[85,86,106] have recently showed the power of using multiple activation domains complexes to generate effective transcription initiation at a specific genomic loci. Therefore, I sought to compare between conservative effectors such as VP64 and P300 to their novel synthetic counterparts, VPR (VP64-P65-Rta) and P65-HSF1. Accordingly, I found that the two synthetic effectors indeed displayed an improved transactivation, resulting in significantly higher expression of the reporter gene (**Figure 13**). This interesting observation reflect on the power of synthetic biology to develop novel parts with advanced activities, and suggest that more combinations of transactivation domains are there to be discovered.

5.2.2. Construction of stable reporting cell-line by random integration

Random integration of heterogeneous DNA is the most common method used to produce transgenic cells. DNA is introduced to the cells by transfection, and stable pool of cells is generated by antibiotic selection followed by functional screening to identify individual clones that have the correct phenotype.

Normally, random integration will lead to heterocellular transgene expression due to variability in the integration sites and local chromatin state[107,108]. Previous work done in CHO cells have demonstrated broad variation of GFP expression levels after random integration[98]. In contrast, the random integration discussed in this thesis resulted in relatively homogeneous expression of mCherry, with two distinct populations of negative and positive cells, and a small sub-population in between (**Figure 15**). I hypothesized that the observed expression profile originates from the low expression rates of mCherry from the minimal CMV promoter (when not induced), diminishing the effect of variability in the integration position.

Subsequently, I sought to isolate a single clone in order to proceed with a neat, homogeneous population with the desired functionality. Using flow cytometry sorting I isolated cells with low mCherry expression into 96-well plate. After 1 month of culturing, 18 clones (~19%) survived and were expanded for sufficient amount of cells to enable functionality screening. The reason for such survival fraction could stem from cellular damage due to the crude sorting procedure (*e.g.* pressure changes during droplet formation) or from poor cell viability before sorting. Sorting process can be improved by using viability dyes so only viable cells will be collected. Moreover, some mammalian cell-lines show poor growth as single cells due to lack of physical contact and/or inadequacy of growth factors. Previous research[109] had shown that the optimal cell concentration for single cell isolation is 4-6 cells per well (in Terasaki plate), resulted in 20%-35% of wells with live single cell. On the other hand, to avoid single cell death, it is possible to use conditioned media to enhance the growth of single cell culture. Conditioned medium is a used medium obtained from proliferating cells secreting growth factors needed for the survival of single cells.

Out of the 18 clones that were expanded, only 2 clones, named herein B4 and C1, presented the expected response of mCherry activation by rTetR-P65-HSF1 in the presence of the inducer doxycycline. But a deeper analysis showed that the fold-induction of clone C1 is very limited (**Figure 16**), that is to say that only clone B4 is adequate to the purposes of this study. I speculate that the low fold-induction of clone C1 is an outcome of a relatively closed chromatin state at the integration site, making the tetO sites inaccessible to the binding of the rTetR transactivator fusion.

To conclude, although the process of generating and selecting transgenic cells using a random integration can be labor intense and time consuming, I managed to isolate a single clone with the desired functionality, with success rates of ~1% (1 out of 96).

5.2.3. Expression and functionality of the RNA-binding protein fusions

Synthetic RNA-binding protein (sRBP) fusions were designed from existing building-blocks, customly synthesized and ordered as a single gene fragment on minimal backbone. First of all, I examined the expression efficiency of the sRBP fusions by employing flow-cytometry experiment to measure the fluorescence of CFP and YFP as expression markers fused to rTetR-PCP and MCP-P65-HSF1, respectively. The results revealed mildly inefficient expression of the proteins (**Figure 18** and **Table 3**), originate mainly from poor fluorescence intensities measured for these fusions (expressed from double-cassette or as singlets). This may suggest that while the DNA is being well transfected into the cells (as indicated from the similar frequencies of fluorescent cells), the transcription rate and/or the folding are inadequate. It was shown before that the ubC promoter is poorly effective in mammalian cells[110] and particularly in CHO cells[111]. The reason for choosing this promoter and not the strong CMV[110,111] is because I wanted to avoid over-using it in both the sRBPs and the slncRNA plasmid (see illustration in **Figure 3**). Moreover, translation of these fusion proteins, especially tripartite fusions, may result in misfolding of the fluorescent component.

Consequently, it is clear that the expression of the sRBP fusions needs optimization, starting with switching to a stronger promoter such as SV40[110,111]. Additionally, alternative design of the fusions may improve the folding and increase the fluorescence levels.

Next, lacking a real positive control in my system, I had to employ indirect approaches to test the functionality of the fusion proteins, rTetR-PCP and MCP-P65-HSF1, in terms of DNA and RNA-binding.

Concluded from **Figure 19**, rTetR-PP7 fusion protein has retained its ability to bind DNA, as indicated from the decrease in mCherry levels when both rTetR-P65-HSF1 and rTetR-PCP compete on the tetO binding-sites, lead to less activation by the P65-HSF1 component. Thereby, I have shown a sophisticated

way to verify DNA-binding capacity by exploiting the theory behind the "dominant-negative effect" (detailed at section 5.2 of the results).

Lastly, the RNA-binding components of the fusion proteins were tested using a fluorescence microscopy assay, in which fluorescent spots in the cell are tracked as indicators of RBPs clusters on an RNA cassette encoding the matching RNA binding-sites. The assay results are not presented in this thesis since no clear observation was established. The failure of the experiment stem from the low percentages (13.8%) of cells expressing the fluorescent sRBP, making it difficult to observe these cells under the fluorescence microscope and reach to meaningful conclusions.

5.3. Site-specific integration of oligo-pool into an artificial chromosome

A challenging aspect of oligonucleotide libraries screening in mammalian cells is to generate a cell pool stably expressing single copy variant. The most common methods to do so is by viral transduction[112,113] or site-specific gene recombination systems[98,114].

In this thesis, I described, for the first time, the integration of an oligo library into a HAC-based system developed by Yamaguchi and others[98]. Initially, I tested the system with GFP integration into the HAC and got 50.5% GFP-positive cells (**Figure 21**), while the original work report 89.9%. It is possible that the differences stem from changes in the transfection procedure, mainly due to different transfection reagent (I used PolyJet and not Lipofectamine). The results also demonstrate a homogeneous cell population in GFP expression, which is an important advantage of site-specific recombination.

To conclude, the described integration method is rather easy and straightforward, allowing us to generate a cell population expressing a DNA of choice within 2-3 weeks, and most importantly in single copy at a precise location in the artificial chromosome. This is an alternative and innovative fashion to carry high-throughput studies of DNA libraries in mammalian cells.

6. Conclusions and outlook

During my research I had the chance to investigate the fascinating world of RNA, both from the structural and functional perspective. Although many techniques are now available for RNA investigation, there are still many more aspects to discover. I believe that by combining structural probing techniques such as SHAPE-Seq with high-throughput analysis of functional slncRNA variants, we can reveal some of the rules governing RNA folding and functionality. For example, slncRNA with binding motifs may perform differently giving the spacer regions in between, and we need to reveal what is the required architecture of these spacers in terms of length and secondary structure (*e.g*. single-stranded or hairpin). Other important feature of RNA to be considered are nuclear localization sequences, size limitations and stability. All of these questions and more may be answered from a thorough, multidisciplinary research as presented in this thesis.

Bibliography

1. Win MN, Smolke CD. Higher-order cellular information processing with synthetic RNA devices. *Science*. 2008;322(5900):456-460.

2. Green AA, Silver PA, Collins JJ, Yin P. Toehold Switches: De-Novo-Designed Regulators of Gene Expression. *Cell*. 2014;159(4):925-939.

3. Xie Z, Wroblewska L, Prochazka L, Weiss R, Benenson Y. Multi-Input RNAi-Based Logic Circuit for Identification of Specific Cancer Cells. *Science (80-)*. 2011;333(6047):1307-1311.

4. Wroblewska L, Kitada T, Endo K, et al. Mammalian synthetic circuits with RNA binding proteins for RNA-only delivery. *Nat Biotechnol*. 2015;33(8):839-841.

5. Harvey I, Garneau P, Pelletier J. Inhibition of translation by RNA-small molecule interactions. *RNA*. 2002;8(4):452-463.

6. Suess B, Hanson S, Berens C, Fink B, Schroeder R, Hillen W. Conditional gene expression by controlling translation with tetracycline-binding aptamers. *Nucleic Acids Res*. 2003;31(7):1853-1858.

7. Desai SK, Gallivan JP. Genetic Screens and Selections for Small Molecules Based on a Synthetic Riboswitch That Activates Protein Translation. *J Am Chem Soc*. 2004;126(41):13247-13254.

8. Buxbaum AR, Haimovich G, Singer RH. In the right place at the right time: visualizing and understanding mRNA localization. *Nat Rev Mol Cell Biol*. 2015;16(2):95-109.

9. Green AA, Kim J, Ma D, Silver PA, Collins JJ, Yin P. Complex cellular logic computation using ribocomputing devices. *Nature*. 2017;548(7665):117-121. doi:10.1038/nature23271.

10. Hentze MW, Caughman SW, Rouault TA, et al. Identification of the iron-responsive element for the translational regulation of human ferritin mRNA. *Science*. 1987;238(4833):1570-1573.

11. St Johnston D. Moving messages: the intracellular localization of mRNAs. *Nat Rev Mol Cell Biol*. 2005;6(5):363-375.

12. Saito H, Kobayashi T, Hara T, et al. Synthetic translational regulation by an L7Ae-kink-turn RNP switch. *Nat Chem Biol*. 2010;6(1):71-78.

13. Lewis CJT, Pan T, Kalsotra A. RNA modifications and structures cooperate to guide RNA–protein interactions. *Nat Rev Mol Cell Biol*. 2017;18(3):202-210.

14. Khalil AS, Collins JJ. Synthetic biology: applications come of age. *Nat Rev Genet*. 2010;11(5):367-379.

15. Isaacs FJ, Dwyer DJ, Collins JJ. RNA synthetic biology. *Nat Biotechnol Vol*. 2006;24(5).

16. Werstuck G, Green MR. Controlling gene expression in living cells through small molecule-RNA interactions. *Science*. 1998;282(5387):296-298. http://www.ncbi.nlm.nih.gov/pubmed/9765156. Accessed December 25, 2018.

17. Hutvagner G, Zamore PD. A microRNA in a Multiple-Turnover RNAi Enzyme Complex. *Science (80-)*. 2002;297(5589):2056-2060.

18. Rinaudo K, Bleris L, Maddamsetti R, Subramanian S, Weiss R, Benenson Y. A universal RNAi-based logic evaluator that operates in mammalian cells. *Nat Biotechnol*. 2007;25(7):795-801.

19. Chen AH, Silver PA. Designing biological compartmentalization. *Trends Cell Biol*. 2012;22(12):662-670.

20. Ausländer S, Stücheli P, Rehm C, Ausländer D, Hartig JS, Fussenegger M. A general design strategy for protein-responsive riboswitches in mammalian cells. *Nat Methods*. 2014;11(11):1154-1160.

21. Sachdeva G, Garg A, Godding D, Way JC, Silver PA. In vivo co-localization of enzymes on RNA scaffolds increases metabolic production in a geometrically dependent manner. *Nucleic Acids Res*. 2014;42(14):9493-9503.

22. Pardee K, Green AA, Takahashi MK, et al. Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell*. 2016;165(5):1255-1266.

23. Brown D, Brown J, Kang C, Gold L, Allen P. Single-stranded RNA recognition by the bacteriophage T4 translational repressor, regA. *J Biol Chem*. 1997;272(23):14969-14974.

24. Schlax PJ, Xavier KA, Gluick TC, Draper DE. Translational Repression of the *Escherichia coli* α Operon mRNA. *J Biol Chem*. 2001;276(42):38494-38501.

25. Romaniuk PJ, Lowary P, Wu HN, Stormo G, Uhlenbeck OC. RNA binding site of R17 coat protein. *Biochemistry*. 1987;26(6):1563-1568.

26. Cerretti DP, Mattheakis LC, Kearney KR, Vu L, Nomura M. Translational regulation of the spc operon in Escherichia coli. Identification and structural analysis of the target site for S8 repressor protein. *J Mol Biol*. 1988;204(2):309-329. http://www.ncbi.nlm.nih.gov/pubmed/2464692. Accessed December 25, 2018.

27. Sacerdot C, Caillet J, Graffe M, et al. The Escherichia coli threonyl-tRNA synthetase gene contains a split ribosomal binding site interrupted by a hairpin structure that is essential for autoregulation. *Mol Microbiol*. 1998;29(4):1077-1090.

28. Lim F, Peabody DS. RNA recognition site of PP7 coat protein. *Nucleic Acids Res*. 2002;30(19):4138-4144.

29. Hattman S, Newman L, Murthy HM, Nagaraja V. Com, the phage Mu mom translational activator, is a zinc-binding protein that binds specifically to its cognate mRNA. *Proc Natl Acad Sci U S A*. 1991;88(22):10027-10031.

30. Wulczyn FG, Kahmann R. Translational stimulation: RNA sequence and structure requirements for binding of Com protein. *Cell*. 1991;65(2):259-269.

31. Kinney JB, Murugan A, Callan CG, Cox EC. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc Natl Acad Sci*. 2010;107(20):9158-9163.

32. Sharon E, Kalma Y, Sharp A, et al. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol*. 2012;30(6):521-530.

33. Shen SQ, Myers CA, Hughes AEO, Byrne LC, Flannery JG, Corbo JC. Massively parallel cis-regulatory analysis in the mammalian central nervous system.

34. Levy L, Anavy L, Solomon O, et al. A Synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial σ(54) Promoters. *Cell Rep*. 2017;21(3):845-858.

35. Weingarten-Gabbay S, Elias-Kirma S, Nir R, et al. Comparative genetics: Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science (80-)*. 2016;351(6270).

36. Gott JM, Wilhelm LJ, Uhlenbeck OC. RNA binding properties of the coat protein from bacteriophage GA. *Nucleic Acids Res*. 1991;19(23):6499-6503.

37. Peabody DS. The RNA binding site of bacteriophage MS2 coat protein. *EMBO J*. 1993;12(2):595-600.

38. Lim F, Peabody DS. RNA recognition site of PP7 coat protein. *Nucleic Acids Res*. 2002;30(19):4138-4144.

39. Lim F, Spingola M, Peabody DS. The RNA-binding site of bacteriophage Qbeta coat protein. *J Biol Chem*. 1996;271(50):31839-31845.

40. Winkler WC, Breaker RR. REGULATION OF BACTERIAL GENE EXPRESSION BY RIBOSWITCHES. *Annu Rev Microbiol*. 2005;59(1):487-517.

41. Lucks JB, Mortimer SA, Trapnell C, et al. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc Natl Acad Sci U S A*. 2011;108(27):11063-11068. doi:10.1073/pnas.1106501108.

42. Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*. 2014;505(7485):701-705.

43. Ding Y, Kwok CK, Tang Y, Bevilacqua PC, Assmann SM. Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nat Protoc*. 2015;10(7):1050-1066.

44. Flynn RA, Zhang QC, Spitale RC, Lee B, Mumbach MR, Chang HY. Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. *Nat Protoc*. 2016;11(2):273-290.

45. Zubradt M, Gupta P, Persad S, Lambowitz AM, Weissman JS, Rouskin S. DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods*. 2017;14(1):75-82.

46. Kertesz M, Wan Y, Mazor E, et al. Genome-wide measurement of RNA secondary structure in yeast. *Nature*. 2010;467(7311):103-107.

47. Peattie DA, Gilbert W. Chemical probes for higher-order structure in RNA. *Proc Natl Acad Sci U S A*. 1980;77(8):4679-4682.

48. Johnsson P, Lipovich L, Grandér D, Morris K V. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta*. 2014;1840(3):1063-1071.

49. Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*. 2009;458(7235):223-227.

50. Pang KC, Frith MC, Mattick JS. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet*. 2006;22(1):1-5.

51. Li R, Zhu H, Luo Y. Understanding the functions of long non-coding RNAs through their higher-order structures. *Int J Mol Sci*. 2016;17(5).

52. Rinn JL, Chang HY. Genome Regulation by Long Noncoding RNAs. *Annu Rev Biochem*. 2012;81(1):145-166.

53. Wutz A, Rasmussen TP, Jaenisch R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat Genet*. 2002;30(2):167-174.

54. Swiezewski S, Liu F, Magusin A, Dean C. Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature*. 2009;462(7274):799-802.

55. Tripathi V, Ellis JD, Shen Z, et al. The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Mol Cell*. 2010;39(6):925-938.

56. Rinn JL, Kertesz M, Wang JK, et al. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell*. 2007;129(7):1311-1323.

57. Kino T, Hurt DE, Ichijo T, Nader N, Chrousos GP. Noncoding RNA Gas5 Is a Growth Arrest- and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Sci Signal*. 2010;3(107):ra8-ra8.

58. Bernstein E, Allis CD. RNA meets chromatin. *Genes Dev*. 2005;19(14):1635-1655.

59. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. *Science*. 1969;165(3891):349-357. http://www.ncbi.nlm.nih.gov/pubmed/5789433. Accessed January 9, 2019.

60. Paul J, Duerksen JD. Chromatin-associated RNA content of heterochromatin and euchromatin. *Mol Cell Biochem*. 1975;9(1):9-16.

61. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. *Mol Cell*. 2012;43(6):904-914.

62. Khalil AM, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci*. 2009;106(28):11667-11672.

63. Guttman M, Donaghey J, Carey BW, et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*. 2011;477(7364):295-300.

64. Tsai M-C, Manor O, Wan Y, et al. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science (80-)*. 2010;329(5992):689-693.

65. Delebecque CJ, Lindner AB, Silver PA, Aldaye FA. Organization of Intracellular Reactions with Rationally Designed RNA Assemblies. *Science (80-)*. 2011;333(6041):470-474.

66. Shechner DM, Hacisuleyman E, Younger ST, Rinn JL. Multiplexable, locus-specific targeting of long RNAs with CRISPR-Display. *Nat Methods*. 2015;12(7):664-670.

67. Zalatan JG, Lee ME, Almeida R, et al. Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds. *Cell*. 2015;160(0):339-350.

68. Mali P, Aach J, Stranges PB, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*. 2013;31(9):833-838.

69. Garneau JE, Dupuis M-È, Villion M, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*. 2010;468(7320):67-71.

70. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. 2012;337(6096):816-821.

71. Duncan CDS, Weeks KM. Nonhierarchical ribonucleoprotein assembly suggests a strain-propagation model for protein-facilitated RNA folding. *Biochemistry*. 2010;49(26):5418-5425.

72. Watters KE, Abbott TR, Lucks JB. Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucleic Acids Res*. 2016;44(2):e12.

73. Smola MJ, Calabrese JM, Weeks KM. Detection of RNA-Protein Interactions in Living Cells with SHAPE. *Biochemistry*. 2015;54(46):6867-6875.

74. Shukla CJ, McCorkindale AL, Gerhardinger C, et al. High-throughput identification of RNA nuclear enrichment sequences. *EMBO J*. January 2018:e98452.

75. Medina G, Juárez K, Valderrama B, Soberón-Chávez G. Mechanism of Pseudomonas aeruginosa RhlR transcriptional regulation of the rhlAB promoter. *J Bacteriol*. 2003;185(20):5976-5983.

76. Katz N, Cohen R, Solomon O, et al. An in vivo Binding Assay for RNA-Binding Proteins Based on Repression of a Reporter Gene. *ACS Synth Biol*. November 2018:acssynbio.8b00378.

77. Katz N, Cohen R, Solomon O, et al. RBP-RNA interactions in the 5 UTR lead to structural changes that alter translation. *bioRxiv*. April 2018:174888.

78. Spitale RC, Crisalli P, Flynn R a, Torre EA, Kool ET, Chang HY. RNA SHAPE analysis in living cells. *Nat Chem Biol*. 2013;9(1):18-20.

79. Aviran S, Trapnell C, Lucks JB, et al. Modeling and automation of sequencing-based characterization of RNA structure. *Proc Natl Acad Sci*. 2011;108(27):11069-11074.

80. Spitale RC, Flynn RA, Zhang QC, et al. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*. 2015;519(7544):486-490.

81. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009;6(5):343-345.

82. Kim JH, Lee S-R, Li L-H, et al. High Cleavage Efficiency of a 2A Peptide Derived from Porcine Teschovirus-1 in Human Cell Lines, Zebrafish and Mice. *Zebrafish Mice PLoS ONE*. 2011;6(4):18556-70204.

83. Beerli RR, Segal DJ, Dreier B, Barbas CF. Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc Natl Acad Sci U S A*. 1998;95(25):14628-14633.

84. Ogryzko V V, Schiltz RL, Russanova V. The Transcriptional Coactivators p300 and CBP Are Histone Acetyltransferases. *Cell*. 1996;87:953-959..

85. Chavez A, Scheiman J, Vora S, et al. Highly efficient Cas9-mediated transcriptional programming. *Nat Methods*. 2015;12(4):326-328.

86. Konermann S, Brigham MD, Trevino AE, et al. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*. 2015;517(7536):583-588.

87. Yamaguchi S, Kazuki Y, Nakayama Y, Nanba E, Oshimura M, Ohbayashi T. A Method for Producing Transgenic Cells Using a Multi-Integrase System on a Human Artificial Chromosome Vector. Najbauer J, ed. *PLoS One*. 2011;6(2):e17267.

88. Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J. 5S Ribosomal RNA Database. *Nucleic Acids Res*. 2002;30(1):176-178.

89. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie*. 1994;125(2):167–188.

90. Washietl S, Hofacker IL, Stadler PF, Kellis M. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Res*. 2012;40(10):4261-4272.

91. Zarringhalam K, Meyer MM, Dotu I, Chuang JH, Clote P. Integrating Chemical Footprinting Data into RNA Secondary Structure Prediction. Gibas C, ed. *PLoS One*. 2012;7(10):e45160.

92. Katherine E. Deigana, Tian W. Lia, David H. Mathewsb,1, and Kevin M. Weeksa 1 aDepartment. Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci*. 2009.

93. Ouyang Z, Snyder MP, Chang HY. SeqFold: Genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res*. 2013;23(2):377-387.

94. Watters KE, Abbott TR, Lucks JB. Simultaneous Characterization of Cellular RNA Structure and Function with in-cell SHAPE-Seq Title for Mobile Devices : Simultaneous Cellular RNA Structure and Function. :1-25.

95. Kunert R, Vorauer-Uhl K. Strategies for efficient transfection of CHO-cells with plasmid DNA. *Methods Mol Biol*. 2012;801:213-226.

96. Li R. Transient transfection of CHO cells using linear polyethylenimine is a simple and effective means of producing rainbow trout recombinant IFN-γ protein. *Cytotechnology*. 2014;67(6):987-993.

97. Nissim L, Perli SD, Fridkin A, Perez-Pinera P, Lu TK. Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Mol Cell*. 2014;54(4):698-710.

98. Yamaguchi S, Kazuki Y, Nakayama Y, Nanba E, Oshimura M, Ohbayashi T. A Method for Producing Transgenic Cells Using a Multi-Integrase System on a Human Artificial Chromosome Vector. Najbauer J, ed. *PLoS One*. 2011;6(2):e17267.

99. Bernardi A, Spahr PF. Nucleotide sequence at the binding site for coat protein on RNA of bacteriophage R17. *Proc Natl Acad Sci U S A*. 1972;69(10):3033-3037.

100. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*. 1990;29(6-7):1105-1119.

101. Ding Y, Chan CY, Lawrence CE. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res*. 2004;32(Web Server):W135-W141.

102. Kutchko KM, Sanders W, Ziehr B, et al. Multiple conformations are a conserved and regulatory feature of the RB1 5' UTR. *RNA*. 2015;21(7):1274-1285.

103. Halvorsen M, Martin JS, Broadaway S, Laederach A. Disease-Associated Mutations That Alter the RNA Structural Ensemble. Gojobori T, ed. *PLoS Genet*. 2010;6(8):e1001074.

104. Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*. 2014;505(7485):696-700. doi:10.1038/nature12756.

105. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-Sequence Determinants of Gene Expression in Escherichia coli. *Science (80-)*. 2009;324(5924):255-258.

106. Tanenbaum ME, Gilbert LA, Qi LS, Weissman JS, Vale RD. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell*. 2014;159(3):635-646.

107. Pilbrough W, Munro TP, Gray P. Intraclonal Protein Expression Heterogeneity in Recombinant CHO Cells. Kudla G, ed. *PLoS One*. 2009;4(12):e8432.

108. Ramírez A, Milot E, Ponsa I, et al. Sequence and chromosomal context effects on variegated expression of keratin 5/lacZ constructs in stratified epithelia of transgenic mice. *Genetics*. 2001;158(1):341-350. h

109. Yaron JR, Ziegler CP, Tran TH, Glenn HL, Meldrum DR. *A Convenient, Optimized Pipeline for Isolation, Fluorescence Microscopy and Molecular Analysis of Live Single Cells*. Vol 16.; 2014.

110. Qin JY, Zhang L, Clift KL, et al. Systematic Comparison of Constitutive Promoters and the Doxycycline-Inducible Promoter. *PLoS One*. 2010;5(5).

111. Wang X-Y, Zhang J-H, Zhang X, Sun Q-L, Zhao C-P, Wang T-Y. Impact of Different Promoters on Episomal Vectors Harbouring Characteristic Motifs of Matrix Attachment Regions. *Sci Rep*. 2016;6:26446.

112. Weingarten-Gabbay S, Elias-Kirma S, Nir R, et al. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science (80-)*. 2016;351(6270):aad4939.

113. Schmidt T, Schmid-Burgk JL, Hornung V. Synthesis of an arrayed sgRNA library targeting the human genome. *Sci Rep*. 2015;5(1):14987.

114. Matreyek KA, Stephany JJ, Fowler DM. A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res*. 2017;45(11):e102-e102.

תקציר

במשך שנים רבות, מניפולציות ובקרה על ביטוי גנטי התאפשרו רק באמצעות קומץ של פרומוטורים ופקטורי שעתוק חלבוניים מוכרים ומאופיינים היטב. עם זאת, לאחרונה אנו עדים לפיתוח של יותר ויותר שיטות בקרה המבוססות על מולקולות רנ"א, מתוך הסתכלות ולמידה של מערכות טבעיות שעושות שימוש ברנ"א כבקר. כיום ידוע לנו כי הגנום האנושי משועתק בצורה נרחבת לצורות שונות של רנ"א שאינו מקודד לחלבונים אשר מבצע מגוון פעולות בתא, כגון בקרה על שעתוק של גנים דרך השתקת או הפעלת מנגוני השעתוק או דרך שינוי ועריכה של הדנ"א ברמת הכרומטין. ישנן מספר סברות על אופן פעולתו של הרנ"א וכיצד הוא מזהה את אתר המטרה שלו בגנום. אחד המנגנונים העיקריים בהם רנ"א משפיע על הדנ"א הוא דרך גיוס חלבונים פונקציונליים (למשל פקטורי שעתוק) על-גבי פיגום רנ"א אליו נקשרים החלבונים הדרושים לביצוע פעולה ייחודית במקום מסוים בגנום. זיהוי האתר הגנומי מתרחש דרך קישור ישיר בין בסיסי הרנ"א לבסיסי הדנ"א או דרך תיווך של חלבונים המסוגלים לזהות את אתר המטרה בדנ"א ובמקביל לקשור גם את מולקולת הרנ"א. בנוסף, מחקרים הראו כי בעוד שרצפי הבסיסים של רנ"א לא-מקודד אינם נשמרו לאורך האבולוציה, המבנה השניוני של המולקולות הללו שמור ברמה גבוהה, מה שמעיד על כך שיכולת פעולתו של הרנ"א קשורה באופן ישיר למבנה הדו-ממדי של התעתיק. עדיין אין ביכולתנו להבין את החוקיות העומדת מאחורי מבנים אלו ולכן על אף שמולקולות הרנ"א מהוות פוטנציאל אדיר בתחום של בקרת הגנום, אין ביכולתנו לתכנן ולעצב רנ"א סינטטי אשר יבצע פעולות ייחודיות כרצוננו. לשם כך אנו צריכים להמשיך ולחקור את הקשר שבין רצף, מבנה ופונקציונליות של מולקולות רנ"א לא -קודד באופן יותר מערכתי.

במהלך עבודה זו למדתי מנגנוני בקרה של רנ"א משני כיווני מחקר: הבנת מנגנוני בקרה של תרגום רנ"א שליח חיידקי מנקודת מבט מבנית, והנדסת רנ"א לא-מקודד סינטטי לצורך הפעלת שעתוק גנטי. בחלק הראשון של המחקר שלי עסקתי בחקר המבנה של מולקולות רנ"א שליח חיידקי המקודד לאתרי קישור לחלבונים קושרי רנ"א, אשר הציג תופעות בקרה שלאחר השעתוק שהשפיעו על רמות הביטוי של הגן המקודד ברנ"א השליח (הפחתת והגברת פעילות). לשם כך, השתמשתי בשיטה הנקראת SHAPE-Seq (Selective 2′Hydroxyl acylation Analyzed by Primer Extension followed by sequencing) העושה שימוש בשינויים כימיים על-גבי מולקולות רנ"א וריצוף בדור החדש (NGS) במטרה לזהות את המבנה הדו-ממדי והקשרים הבין-מולקולריים שבין מולקולות הרנ"א לבין חלבונים, דנ"א או רנ"א אחר. בפרויקט זה הצלחנו להראות כי האפקט של הפחתת הפעילות נובע ממעבר בין מצב תרגום פעיל המאופיין בחוסר מבניות של מולקולת הרנ"א לבין מצב תרגום עצור המאופיין במבניות גבוהה של מולקות הרנ"א שגורמת לעיכוב פעולת התרגום של הריבוזום. בהמשך הראנו כי אפקט הגברת הפעילות ככל הנראה נובע ממבנה מאוד סגור שחוסם את פעולת התרגום, מבנה אשר משתנה

בעת קישור של חלבון קושר הרנ"א התואם לאתרי הקישור על הרנ"א כך שהתרגום מתאפשר. בחלק השני של המחקר עסקתי בפרויקט תכנון של ספריית רצפי רנ"א לא-מקודד סינטטי המיועדים לבצע שפעול שעתוק גנטי באמצעות מיזוגים של חלבונים קושרי רנ"א. לשם סריקה של רצפי הרנ"א ומציאת משתנים (variants) פונקציונליים פיתחתי מערכת סריקה המבוססת על שפעול של גן מדווח. במהלך עבודתי על פרויקט זה הצלחתי לייצר קו תאים יציב המבטא גן מדווח המבוסס על גן פלורסנטי מושרה מסוג mCherry. קו תאים זה מאופיין ברמת ביטוי בסיסית נמוכה אשר משופעלת באופן חד רק בנוכחות מפעיל השעתוק והמשרן. בנוסף, נקטתי בגישה חדשנית למחקר של מאגר רצפי דנ"א מסוג אוליגונוקלאוטידים בתאים אנימליים באמצעות הכנסה של ספריית האוליגונוקלאוטידים לכרומוזום מלאכותי בתאי אוגר סיני. על אף שמטרת העל של החלק השני במחקר שלי לא הושלמה, אני מאמינה שהעבודה המוצגת בעבודת גמר זו עשויה לקדם עבודת המשך בעתיד בתחום של רנ"א בקר סינטטי.

למידת רגולציה של תרגום באמצעות מבני רנ"א וחקירה של רנ"א לא-מקודד סינטטי רגולטורי

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר מגיסטר למדעים

בהנדסת ביוטכנולוגיה ומזון

רוני כהן

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

אדר ב' תשע"ט חיפה אפריל 2019

למידת רגולציה של תרגום באמצעות מבני רנ"א וחקירה של רנ"א לא-מקודד סינטטי רגולטורי

רוני כהן