Protein-RNA interactions: Synthetic Biology applications

Noa Katz

Protein-RNA interactions: Synthetic Biology applications

In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Noa Katz

Submitted to the Senate of the Technion - Israel Institute of Technology

Av, 5780, Haifa July 2020

The research thesis was carried out under the supervision of Assistant Professor Dr. Roee Amit in the Department of Biotechnology and Food Engineering of the Technion - Israel Institute of Technology.

The generous financial help of the Technion - Israel Institute of Technology is gratefully acknowledged.

Table of Contents

List of Figures

Abstract

During my PhD I focused on protein-RNA interactions. First, we set out to improve a universal method for live RNA imaging based on protein-RNA binding. RNA imaging cassettes are typically made of repetitive hairpin-structured sites, hence hindering their retention, synthesis, and functionality. The goal was to generate RNA binding sites that have different nucleotide sequences from the native site but retain high affinity to the protein. We first developed an assay for quantifying protein affinity in a cellular environment, based on a competition between the ribosome and the protein for binding to the RNA. We programmed and utilized a liquid-handling robot to carry out experiments.

Next, we wanted to test enough binding sites to generate binding sequences de-novo for labs worldwide. To do so, we developed the induction-based Sort-Seq technique together with our assay system, and tested the affinity of 20,000 mutated sites simultaneously to three proteins. We applied a neural network to expand this space of binding sites, which allowed us to identify the structural and sequence features critical for binding. Finally, we designed new non-repetitive binding site cassettes and validated their functionality in mammalian cells. Consequently, we provide the scientific community with a tool for designing non-repetitive binding sites cassettes, thus substantially shortening the time from design to imaging, while potentially allowing for robust measurements and quantitative data. So far, labs from Harvard, Berkeley, Bio-Frontiers Institute, and Caltech, have expressed interest in our work.

The second part focused on structure-function relationship of RNA. Using a system similar to the binding assay, we demonstrated that deletion of two bases in the binding site alters the structure of the entire RNA molecule. Consequently, the same protein that used to be down-regulating translation upon RNA binding is now up-regulating. This inversion in function due to two bases difference in sequence was surprising and strengthened the notion that RNA structure-function relationship is an open and exciting subject.

List of symbols and abbreviations

DNA - deoxyribonucleic acid
RBP - RNA binding proteins
RBS - ribosome binding site
ATG - start codon
ORF - open reading frame
1D - one-dimensional, 3D - three-dimensional
Amp - ampicillin
ATP - adenosine triphosphate
BA - bioassay media
bp - base pair
C4-HSL - N-butanoyl-L-homoserine lactone
CDS - coding sequence
CMV - cytomegalovirus
CO2 - carbon dioxide
DMEM - Dulbecco Eagle's Minimum Essential Medium
PBS - Dulbecco's phosphate buffered saline
EDTA - ethylenediaminetetraacetic acid
YFP - yellow fluorescent protein
FACS - flow cytometry activated cell sorter
FBS - fetal bovine serum
FL - fluorescence
FP - fluorescent protein
GFP - green fluorescent protein
Kan - kanamycin
lncRNA - long non-coding RNA
mL - milliliter
mM, M - millimolar, molar
mRNA - messenger RNA
NGS - next-generation sequencing
NLS - nuclear localization signal
o/n - overnight
OD - optical density
oligos - oligonucleotides
PCR - polymerase chain reaction
RT - room temperature
RNA - ribonucleic acid

Part 1: Introduction

2.1. RNA Binding Proteins (RBPs)

During the past few years, our knowledge of gene regulation by RNA-binding proteins has greatly increased[1]. It is now evident that RBPs influence the structure and interactions of RNA molecules and play critical roles in their biogenesis, stability and protection, function, transport and cellular localization. Eukaryotic cells encode a large number of RBPs, estimated to be thousands in vertebrates, with each having a unique RNA-binding activity and protein-protein interaction characteristics[2].

The noteworthy diversity of RBPs, which appears to have increased during evolution, has allowed eukaryotic cells to utilize them in an enormous array of combinations unique for each RNA molecule[2]. This diversity is made possible as a result of the modular structure of RBPs, most of which usually contain more than one RNA binding module[3]. The nature of RNA molecules allows for the interaction between the RBP and its substrate to have a structural aspect; their recognition stems from both the sequence of the RNA as well as the formed structure[4]. This allows for more sophisticated regulation, such as that utilized by riboswitches (see next section).

However, despite their crucial significance, much is still unknown about RBPs and their function. It is much harder to study RNA binding than DNA binding for various reasons, mostly due to the dynamic nature of substrate, the RNA, and the difficulty in identifying the RNA target, since most RBPs likely have multiple targets.

2.2. Phage Coat Proteins as RNA Binding Proteins (RBPs)

2.2.1. Phage coat proteins

The phage MS2 is a single stranded RNA coliphage, its capsid protein recognizes a 19-nucleotide stem-loop structure of RNA in the phage genome that includes the SD sequence and the initiation codon of the phage replicase gene[5]. By binding it, the coat protein causes translational repression by the proposed mechanism of secondary structure stabilization via the binding of the coat protein[6]. Such similar mechanism of repression is conserved in other known single stranded RNA phages[7,8], such as the Qβ, PP7, and GA.

It is believed that the exact structure of the binding site is needed for recognition by the coat proteins while only the identity of a certain nucleotides in the operators is crucial for binding; therefore, exchanging nucleotides at insignificant positions will still allow binding at a fairly high affinity. For example, essential nucleotides include the bulged purine in the middle of the

base-paired region for all three binding sites (Figure 1.1). Structural studies have confirmed that these nucleotides are the ones involved in RNA-protein interaction[9].

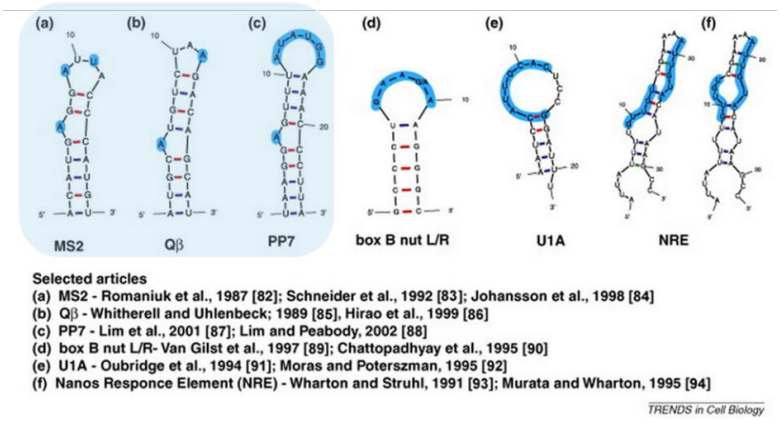


Figure 1.1: RBP binding sites [9]. Dark blue regions represent significant nucleotides for binding

2.2.2. Determining Binding Affinity

The first *in-vivo* studies utilized synthetic constructs of *replicase –lacZ* fusions, which contain the translation initiation site of the phage replicase gene in place of the first nine codons of *lacZ.* This experimental design meant that the translation initiation of *lacZ* was dependent on the site specific to that replicase[10,11]; when a coat protein was bound to the RNA molecule, translation initiation was inhibited. In this manner, it was possible to measure the affinities of various coat proteins to their wild type RNA operators, as well as to each other's wild type operators and various mutant operators.

Recently, a study was published combining both *in-vitro* and *in-vivo* assays to build a system in which an archaeal ribosomal protein regulates the translation of a designed mRNA *in-vitro* and in human cells[12]. The plan was based on the creation of a synthetic riboswitch that utilizes a protein. The work presents two configurations for building such a riboswitch: the binding site placed downstream to the RBS and ATG of the reporter protein, or upstream to the RBS and downstream to the promoter of the reporter protein (Figure 1.2). In each approach the researchers varied the amount of nucleotides between the RBS and the binding site. Their goal was to test their effectiveness, i.e the fold repression effect in response to the protein, of each of the configurations, and choose the best one for the future construction of their translational regulators in eukaryotic cells. Their results indicate that the best approach was to place the binding site downstream of the RBS, in the Open Reading Frame (ORF), with the minimal amount of nucleotides between. Such a design yielded a 70% repression in the *in-vitro* system and up to 80% repression *in-vivo*, as compared to the control lacking the effector protein.

One year ago, a paper was published implementing a high throughput analysis of the MS2 binding site and MS2 protein using a sophisticated *in-vitro* set-up [13]. In this case, the researchers utilized a high-throughput sequencing instrument to quantitatively measure binding and dissociation of the fluorescently labeled *MS2-PCP* to over ten million RNA targets generated on a flow cell surface via *in situ* transcription and intermolecular tethering of RNA to DNA. Among their results, they observe that sequence-specific mutations in the binding site cause significant changes in both association and dissociation rates that influence the overall RBP affinity to these sites. Several of their tested binding sites are also tested in the work presented here.

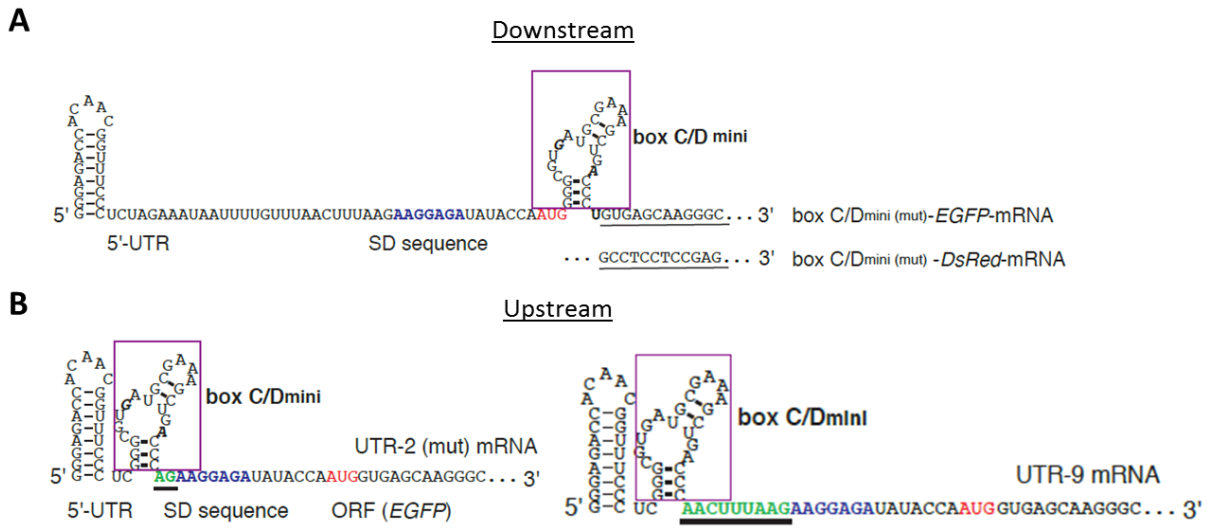


Figure 1.2: Two Approaches to Build the Riboswich[12] A: Downstream approach, the binding site (purple) is right after the AUG (red) B: Upstream approach, the binding site is 2 bases (left) or 9 bases (right) upstream the RBS

2.3. Regulation of Translation

The ribosome is a ribonucleoprotein that is responsible for protein synthesis or translation of mRNA in all live cells. The function of the prokaryotic ribosome is generally divided into three steps[14]: Initiation, which involves the assembly of the two subunits onto the mRNA to be translated. Via base pairing, the RNA component (16S rRNA) in the 30S ribosomal subunit binds the mRNA at the Shine-Dalgarno (SD) sequence, which is typically 4 or 5 bases in length. This creates a double stranded RNA structure in a way that places the AUG, the initiation codon, in the P site of the ribosome. Elongation, in which amino acids are added via a peptide bond to the growing carboxyl end of the chain; once the bond is formed, both the empty tRNA and the next tRNA translocate to the P and E sites along with the mRNA, while a new tRNA

moves to the A site. During Termination, the ribosome encounters one of three termination codons- UAA, UAG, and UGA- and proteins named "release factors" trigger the release of the peptide from the ribosome.

The mechanisms to control translation are numerous. Protein mediated mechanisms include protein binding to translation activating factors, regulation of the start codon selection and subunit joining, and phosphorylation of specific initiation required proteins. Mechanisms which directly involve the mRNA include protein binding to UTRs which either promote or decrease translation, and structure mediated repression such as translation repression by miRNA (eukaryotes) or trans-acting RNA (prokaryotes), and base-paired structures of the mRNA itself.

Creating such base-paired secondary structures as to hinder the ability of the ribosome to bind to the mRNA in the SD sequence, or Ribosome Binding Site (RBS), has been shown to effect translation efficiency[15]. An evolutionary study presents the clearest evidence for this mechanism: the Coliphage MS2 was used to prove that expansion or abbreviation of the RBS provoked compensatory changes in the strength of a hairpin structure that encompasses the ribosome binding site, thus preserving the overall expression levels [16]. Other studies tried to strengthen the RBS-ribosome interaction to overcome the masking of the RBS by a secondary structure[17,18]. Additionally, secondary structures of the mRNA in the region between the SD and the AUG have also been shown to effect translation [19]. Moreover, such structures also take place in the expression of prokaryotic genes via polycistronic transcripts, where translation of a downstream cistron is coupled to that of the preceding cistron. In the more sophisticated control mechanisms, the ribosome needs to pause at a particular point during translation of the upstream cistron to enable initiation of the downstream one; such a mechanism is usually coupled with a downstream cistron that has a usable RBS which is temporarily obscured by secondary structure[20–23].

One distinct example for a structural RNA-based gene regulatory system that occur naturally in bacteria and eukaryotes is the Riboswitch. It is a regulatory segment of an mRNA that binds a small molecule, causing conformational changes in nascent structured mRNA, which results in the repression or activation of gene expression at the translational level [24]. Thus, an mRNA that contains a riboswitch is directly involved in regulating its own activity in response to the concentration of its effector molecule. Synthetic riboswitches have been successfully designed and constructed to regulate translation in bacteria and eukaryotic cells utilizing small molecules, such as tetracycline or theophylline, as the ligands. Yet, the development of a synthetic riboswitch that uses a protein expressed in the cell as the input ligand has rarely been attempted [12].

2.4. Synthetic Circuits

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 [25–28], even-though RNA-based regulatory components have many advantages. Several RNA components have been shown to be functional in multiple organisms [29–33]. RNA can respond rapidly to stimuli, enabling a faster regulatory response as compared with transcriptional regulation [34,35,12,36]. From a structural perspective, RNA molecules can form a variety of biologically functional secondary and tertiary structures [27], which enables modularity. For example, distinct sequence domains within a molecule [36,37] may target different metabolites or nucleic acid molecules [38,39]. All of these characteristics make RNA an appealing target for engineered-based applications[40–42,26,43,44,27,45,46].

Perhaps the most well-known class of RNA-based regulatory modules are riboswitches[38,47–50]. Riboswitches are noncoding mRNA segments that regulate the expression of adjacent genes via structural change, effected by a ligand or metabolite. However, response to metabolites cannot be easily used as the basis of a regulatory network, as there is no convenient feedback or feed-forward mechanism for connection with additional network modules. Implementing network modules using RBPs could enable an alternative multicomponent connectivity for gene-regulatory networks that is not based solely on transcription factors.

Regulatory networks require both inhibitory and up-regulatory modules. The vast majority of known RBP regulatory mechanisms are inhibitory[51–56]. A notable exception is the phage RBP Com, whose binding was demonstrated to destabilize a sequestered ribosome binding site (RBS) of the Mu phage *mom* gene, thereby facilitating translation[57,58]. Several studies have attempted to engineer activation modules utilizing RNA-RBP interactions, based on different mechanisms: recruiting the eIF4G1 eukaryotic translation initiation factor to specific RNA targets via fusion of the initiation factor to an RBP[59,60], adopting a riboswitch-like approach[44], and utilizing an RNA-binding version of the TetR protein[61]. However, despite these notable efforts, RBP-based translational stimulation is still difficult to design in most organisms.

2.5. The design-build-test (DBT) bottleneck in Synthetic Biology

For the past two decades, synthetic biologists have built a portfolio of increasingly sophisticated biological circuits that are able to perform logical functions inside living cells[62–65]. Such circuits are made from “biological parts” which are biochemical analogs of electronic components that are routinely used for the design of electrical circuits (see previous section). Unfortunately, unlike their electronic counterparts, connecting biological parts to form circuits

often fails. This is mostly due to the fact that many parts are short sequences of DNA or RNA, and connecting them introduces unpredictable and undesirable effects[66]. As a result, many iterations of trial and error are often needed before a successful design is achieved. This is termed the design, build, test (DBT) cycle in synthetic biology and is considered to be a major bottleneck for progress in the field. Specifically, the field is lacking computational methods that allow users to reliably design their system of choice without going through multiple time-consuming DBT cycles.

The challenge of formulating such algorithms is rooted in the large space of biomolecules that make-up the biological parts, and the variety of interactions that are possible between them. This translates to a plethora of molecular mechanisms, each governed by differing kinetics, thermodynamic parameters, and free-energy considerations. Consequently, modelling these systems necessitates case-specific kinetic and/or thermodynamic modelling approaches to devise a reliable design algorithm. In recent years, several studies have demonstrated such algorithms for diverse RNA-, DNA- and protein-based applications, with varying degrees of success[67–69]. Notable examples include the Cello algorithm and the Ribosome-binding-site calculator, which are limited to bacterial chassis[70,71] at the present time.

Reliable algorithms are especially needed for the design of RNA-centric functional modules for various applications. In a recent study, the authors we demonstrated model-based functional design of non-repetitive sgRNA cassettes for targeting multiple metabolic genes in bacteria[72]. Another RNA-based system where a reliable design algorithm can help bring about the full potential of the technology is the encoding of multiple repeats of phage coat protein binding elements on an RNA molecule of choice.

2.6. RNA Imaging Systems

The use of fluorescent proteins in tracking gene expression has been demonstrated twenty years ago [73] and has been utilized ever since to follow protein dynamics in living cells. It is only in the last decade, however, that scientists have been following RNA transcripts in a similar fashion. The first report of a system to follow RNA in living cells came out in 1998 [74], describing two constructs: the first, an RNA bacteriophage capsid protein fused to a Green fluorescent protein (GFP) sequence, and the second, multiple copies of the phage coat protein's binding sites fused into the mRNA of a reporter gene (Figure 1.3). The use of a phage derived RNA binding protein prevents the possibility of attachment to non-specific DNA, as these RBPS recognize a unique structure as their binding site.

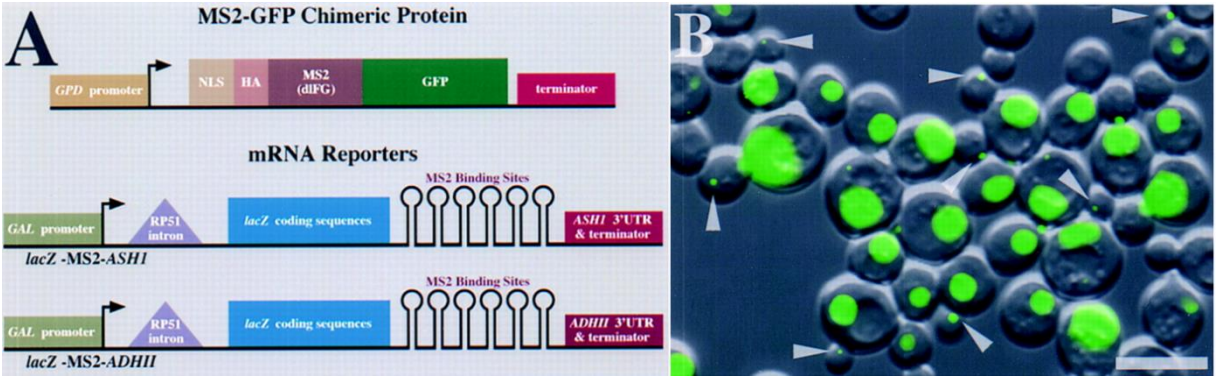


Figure 1.3: First report of a system to follow RNA in living cells [75] A: Schematic describing the constructs used in this approach. The system is comprised of two components, a reporter mRNA and a GFP-MS2 fusion protein. The GFP-MS2 was expressed under the control of the constitutive GPD promoter, while the reporter mRNA was under the control of the GAL promoter. The reporter mRNA contains six binding sites for the coat protein of the bacterial phage MS2. The 3'UTRs were either from the ASH1 gene, to induce mRNA localization at the bud tip, or from the ADHII gene, as control. In addition, a nuclear localization signal (NLS) followed by an HA tag was introduced at the N terminus of the fusion protein, so that that only the GFP protein that is bound to its target mRNA would be present in the cytoplasm. B: Live cells expressing the GFP-MS2 fusion protein and the lacZ-MS2-ASH1 reporter mRNA. Arrows indicate some of the particles, usually in the bud. Bar, 5 µm.

The reason for inserting multiple copies of the binding site is to achieve higher fluorescence intensity on the mRNA than in the cell so as to distinguish between bound and unbound fusion proteins (FPs). However, the number of binding site repeats varies; the shorter ones are 6 MS2 binding sites[75], 12 MS2 binding sites[76], and 24 MS2 and 24 PP7 binding sites[77], while the longest one is a cassette of 96 repeats of the MS2 binding site[78]. The vast potential of this technology for live-tracking of transcription in living cells has opened up the possibility of live mRNA research not possible with any other method, since the perturbation of the cells is minimal[79]. Its first application was in yeast cells, but since then the system has been shown to work in bacteria[78,80], amoebae[81] mammalian cells[82] and Drosophila embryos[83]. Such cassettes have also been utilized in studies for gene editing applications[84,85].

However, a limited understanding of CP-binding *in vivo* has forced cassette designs into incorporating repeated hairpin-like sequence elements, making them cumbersome to synthesize using current oligo-based technology. Subsequent steps, including cloning and genome maintenance, are also badly affected by the repeat nature of the cassette. Finally, repeat sequence elements are notoriously unstable[86], thus damaging protein binding to the cassette and causing occupancy-related experimental noise. Consequently, these limitations hinder the utility of these cassettes for robust quantitative measurements[87] as well as expansion to more complex multi-genic applications.

Part 2: Methods

2.1. Single clone: lab-work

2.1.1. Design and construction of binding-site plasmids

Binding-site cassettes (see Supplementary Table 1) were ordered either as double-stranded DNA minigenes from Gen9 or as cloned plasmids (minigene + vector) from Twist Biosciences. Each minigene was ~500 bp long and contained the parts in the following order: EagI restriction site, ~40 bases of the 5' end of the Kanamycin (Kan) resistance gene, pLac-Ara promoter, ribosome binding site (RBS), an RBP binding site, 80 bases of the 5' end of the mCherry gene, and an ApaLI restriction site. As mentioned, each cassette contained either a wild-type or a mutated RBP binding site (see Supplementary Table 1), at varying distances downstream to the RBS. All binding sites were derived from the wild-type binding sites of the coat proteins of one of the four bacteriophages MS2, PP7, GA and Qβ. For insertion into the binding-site plasmid backbone, they were double-digested with EagI-HF and ApaLI (New England Biolabs [NEB]). The digested minigenes were then cloned into the binding-site backbone containing the rest of the mCherry gene, terminator, and a Kanamycin resistance gene, by ligation and transformation into *E. coli* TOP10 cells (ThermoFisher Scientific). Purified plasmids were stored in 96-well format, for transformation into *E. coli* TOP10 cells containing one of four fusion-RBP plasmids (see below).

2.1.2. Design and construction of fusion-RBP plasmids

RBP sequences lacking a stop codon were amplified via PCR of either Addgene or custom-ordered templates (Genescript or IDT, see Supplementary Table 2). All RBPs presented (MCP, PCP, GCP, and QCP) were cloned into the RBP plasmid between restriction sites KpnI and AgeI, immediately upstream of an mCerulean gene lacking a start codon, under the pRhlR promoter (containing the *rhlAB* las box [88]) and induced by C_4-HSL. The backbone contained an Ampicillin (Amp) resistance gene. The resulting fusion-RBP plasmids were transformed into *E. coli* TOP10 cells. After Sanger sequencing, positive transformants were made chemically-competent and stored at -80°C in 96-well format.

2.1.3. Transformation of binding-site plasmids

Binding-site plasmids stored in a 96-well format were simultaneously transformed into chemically-competent bacterial cells containing one of the RBP-mCeulean plasmids. After transformation, cells were plated using an 8-channel pipettor on 8-lane plates (Axygen)

containing LB-agar with relevant antibiotics (Kan and Amp). Double transformants were selected, grown overnight, and stored as glycerol stocks at -80°C in 96-well plates (Axygen).

2.1.4. Single clone expression level assay

Dose-response fluorescence experiments were performed using a liquid-handling system in combination with a Liconic incubator and a TECAN Infinite F200 PRO platereader. Each measurement was carried out in duplicates. Double-transformant strains were grown at 37°C and 250 rpm shaking in 1.5 ml LB in 48-well plates with appropriate antibiotics (Kan and Amp) over a period of 16 hours (overnight). In the morning, the inducer for the rhlR promoter C_4-HSL was pipetted manually to 4 wells in an inducer plate, and then diluted by the robot into 24 concentrations ranging from 0 to 218 nM. While the inducer dilutions were being prepared, semi-poor medium consisting of 95% bioassay buffer (for 1 L: 0.5 g Tryptone [Bacto], 0.3 ml Glycerol, 5.8 g NaCl, 50 ml 1M MgSO4, 1ml 10xPBS buffer pH 7.4, 950 ml DDW) and 5% LB was heated in the incubator, in 96-well plates. The overnight strains were then diluted by the liquid-handling robot by a factor of 100 into 200 µL of pre-heated semi-poor medium, in 96-well plates suitable for fluorescent measurement. The diluted inducer was then transferred by the robot from the inducer plate to the 96-well plates containing the strains. The plates were shaken at 37°C for 6 hours. Note, that induction was only used for the rhlR promoter, which controls the expression of the RBP-mCerulean fusion. The pLac/Ara promoter controlling the mCherry reporter gene functioned as a constitutive promoter of suitable strength in our strains and did not require IPTG or Arabinose induction.

Measurement of OD, and mCherry and mCerulean fluorescence were taken via a platereader every 30 minutes. Blank measurements (growth medium only) were subtracted from all fluorescence measurements. For each day of experiment (16 different strains), a time interval of logarithmic growth was chosen (T_0 to T_{final}) according to the measured growth curves, between the linear growth phase and the stationary (T_0 is typically the third measured time point). Six to eight time points were taken into account, discarding the first and last measurements to avoid errors derived from inaccuracy of exponential growth detection. Strains that showed abnormal growth curves or strains where logarithmic growth phase could not be detected, were not taken into account and the experiment was repeated. See Fig. S2 for experimental schematic and a sample data set.

2.1.5. RNA extraction and reverse-transcription for qPCR measurements

Starters of *E. coli* TOP10 containing the relevant constructs on plasmids were grown in LB medium with appropriate antibiotics overnight (16 hr). The next morning, the cultures were diluted 1:100 into fresh semi-poor medium and grown for five hours. For each isolation, RNA

was extracted from 1.8 ml of cell culture using standard protocols. Briefly, cells were lysed using Max Bacterial Enhancement Reagent followed by TRIzol treatment (both from 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. RNA quality was assessed by running 500 ng on 1% agarose gel. After extraction, RNA was subjected to DNAse (Ambion/Life Technologies) and then reverse-transcribed using MultiScribe Reverse Transcriptase and random primer mix (Applied Biosystems/Life Technologies). For qPCR experiments, RNA was isolated from three individual colonies for each construct.

2.1.6. qPCR measurements

Primer pairs for mCherry and normalizing gene *idnT* were chosen using the Primer Express software and aligned using BLAST [89] (NCBI) with respect to the *E. coli* K-12 substr. DH10B (taxid:316385) genome (which is similar to TOP10) to avoid off-target amplicons. qPCR was carried out on a QuantStudio 12K Flex machine (Applied Biosystems/Life Technologies) using SYBR-Green. Three technical replicates were measured for each of the three biological replicates. A C_T threshold of 0.2 was chosen for all genes.

2.2. Single clone: analysis

2.2.1. Single clone expression level analysis

The average normalized fluorescence of mCerulean, and rate of production of mCherry, were calculated for each inducer concentration using the routine developed in [90], as follows:

mCerulean average normalized fluorescence: for each inducer concentration, mCerulean measurements were normalized by OD. Normalized measurements were then averaged over the N logarithmic-growth timepoints in the interval [T0, Tfinal], yielding:

$$mCerulean = \frac{1}{N}\sum_{t=T_0}^{T_{final}} \frac{mCerulean(t)}{OD(t)} \quad (1)$$

mCherry rate of production: for each inducer concentration, mCherry fluorescence at T_0 was subtracted from mCherry fluorescence at T_{final}, and the result was divided by the integral of OD during the logarithmic growth phase:

$$mCherry\ rate\ of\ production = \frac{mCherry(T_{final}) - mCherry(T_0)}{\int_{T_0}^{T_{final}} dtOD(t)} \quad (2)$$

Finally, we plotted mCherry rate of production [91] as a function of averaged normalized mCerulean expression, creating dose response curves as a function of RBP-mCerulean fluorescence. Our choice for computing rate of production for mCherry stems from our belief that this observable best quantifies the regulatory effect, which is a function of the absolute number of inducer protein present (i.e RBP-mCerulean) at a any given moment in time. Data points with higher than two standard deviations calculated over mCerulean and mCherry fluorescence at all the inducer concentrations of the same strain) between the two duplicates were not taken into account and plots with 25% or higher of such points were discarded and the experiment repeated.

2.2.2. Dose response fitting routine and K_d extraction

Final data analysis and fit were carried out on plots of rate of mCherry production as a function of averaged normalized mCerulean fluorescence at each inducer concentration. Such plots represent production of the reporter gene as a function of RBP presence in the cell. The fitting analysis and K_d extraction were based on the following two-state thermodynamic model:

$$mCherry\ rate\ of\ production = P_{bound}k_{bound} + P_{unbound}k_{unbound} \quad (3)$$

Here, the mCherry mRNA is either bound to the RBP or unbound, with probabilities P_{bound} and $P_{unbound}$ and ribosomal translation rates k_{bound} and $k_{unbound}$, respectively. The probabilities of the two states are given by:

$$P_{bound} = \frac{\left([x]/K_d\right)^n}{1+\left([x]/K_d\right)^n} \quad (4)$$

and

$$P_{unbound} = \frac{1}{1+\left([x]/K_d\right)^n} \quad (5)$$

where $[x]$ is RBP concentration, K_d is an effective dissociation constant, and n is a constant that quantifies RBP cooperativity; it represents the number of RBPs that need to bind the binding site simultaneously for the regulatory effect to take place. Substituting the probabilities into Eq. 3 gives:

$$mCherry\ rate\ of\ production = \frac{\left([x]/K_d\right)^n}{1+\left([x]/K_d\right)^n}k_{bound} + \frac{1}{1+\left([x]/K_d\right)^n}k_{unbound} \quad (6)$$

For the case in which we observe a down-regulatory effect, we have significantly less translation for high [*x*], which implies that $k_{bound} \ll k_{unbound}$ and that we may neglect the contribution of the bound state to translation. For the case in which we observe an up-regulatory affect for large [*x*], we have $k_{bound} \gg k_{unbound}$, and we neglect the contribution of the unbound state.

The final models used for fitting the two cases are summarized as follows:

$$mCherry\ rate\ of\ production \simeq \begin{cases} \dfrac{k_{unbound}}{1+\left(\left[x\right]/K_d\right)^n} + C & downregulatory\ effect \\ \dfrac{\left(\left[x\right]/K_d\right)^n k_{bound}}{1+\left(\left[x\right]/K_d\right)^n} + C & upregulatory\ effect \end{cases} \quad (7)$$

where *C* is the fluorescence baseline. Only fit results with $R^2 > 0.6$ were taken into account. For those fits, K_d error was typically in the range of 0.5-20%, for a 0.67 confidence interval.

2.3. Oligo-library: lab-work

2.3.1. Construction of the oligo library

We designed 10,000 mutated versions of the WT binding sites to the phage CPs of PP7, MS2 and Qβ, and positioned them at two positions within the ribosomal initiation region (Figure 4.1). Each of the designed 10k sites were positioned either one or two nucleotides downstream to the mCherry start codon, resulting in 20k different configurations. We then ordered the following oligo library (OL) from Agilent: 100k oligos (Table S1), each 210bp long containing the following components: BamHI restriction site, barcode (five for each variant), constitutive promoter (cPr), Ribosome Binding Site (RBS), mCherry start codon, one or two bases (denoted by delta), the variant binding site, ~60 bases of the mCherry gene, and an ApaLI restriction site. We then cloned the OL using restriction-based cloning strategy. Briefly, the 100k-variant ssDNA library from Agilent was amplified in a 96-well plate using PCR (see Table S2 for primers), purified, and merged into one tube. Following purification, dsDNA was cut using BamHI-hf and ApaLI and cleaned. Resulting DNA fragments were ligated to the target plasmid containing an mCherry open reading frame and a terminator, using a 1:1 ratio. Ligated plasmids were transformed to E. cloni® cells (Lucigen) and plated on 37 large agar plates with Kanamycin antibiotics in order to conserve library complexity. Approximately two million colonies were scraped and transferred to an Erlenmeyer for growth. After O/N growth, plasmids were extracted using a maxiprep kit (Agilent), their concentration was measured, and they were stored in an Eppendorf tube in -20.

2.3.2. Construction of RBP-GFP fusions

RBP sequences lacking a stop codon were amplified via PCR of either Addgene or custom-ordered templates (Genescript or IDT, see Table S3). MCP, PCP and QCP were cloned into the RBP plasmid between restriction sites KpnI and AgeI, immediately upstream of a GFP gene lacking a start codon, under the pRhlR promoter (containing the rhlAB las box38) and induced by C4−HSL. The backbone contained an Ampicillin (Amp) resistance gene. The resulting fusion-RBP plasmids were transformed into *E.coli* TOP10 cells. After Sanger sequencing, positive transformants were made chemically competent and stored at −80°C in 96-well format.

2.3.3. Double Transformation of OL and RBP-GFP plasmids.

Note: steps 3 to 5 were conducted three times, one for each RBP-GFP fusions.

OL DNA was transformed into ~300 chemically competent bacterial cell in 100ul aliquots containing one of the RBP-mCeulean plasmids in 96-well format. After transformation, cells were grown in 2L liquid LB with twice the concentration of the antibiotics – Kanamycin and Ampicillin – overnight at 37°C and 250rpm. After growth glycerol stocks were made by centrifugation, re-suspension in 30ml LB, mix 1.2ml with 400ul 80% glycerol – 20% LB solution and store in -80°C.

2.3.4. Induction-based Sort-Seq OL assay

One full glycerol stock of the library was dissolved in 500ml of LB with antibiotics and grown overnight at 37°C and 250rpm. In the morning, the bacterial culture was diluted 1:50 into 100ml of semi-poor medium consisting of 95% bioassay buffer (BA: for 1L - 0.5g Tryptone [Bacto], 0.3ml Glycerol, 5.8g NaCl, 50ml 1M MgSO4, 1ml 10xPBS buffer pH 7.4, 950ml DDW) and 5% LB. The inducer, N-butanoyl-L-homoserine Lactone (C4-HSL), was pipetted manually to a final concentration of one out of six final concentrations: 0uM, 0.02uM, 0.2uM, 2uM, 20uM, and 200uM. Cells were grown at 37°C and 250rpm to mid-log phase (OD600 of ~0.6) as measured by a spectrophotometer and taken to the FACS for sorting.

During sorting by the FACSAria II (BD Biosciences) cell sorter each inducer level culture was sorted into eight bins of increasing mCherry levels spanning the entire fluorescence range except for 5% at the higher end (bin 1 - low mCherry to bin 8 - high mCherry), and set GFP levels (for example, the 0mM culture were sorted according to zero GFP fluorescence, the 0.02uM culture to slightly positive GFP fluorescence, and so on). Sorting was done at a flow rate of ~20,000 cells per second. 300k cells were collected in each bin for the entire 6x8 bin matrix. After sorting, the binned bacteria were transferred to 10ml LB+KAN+AMP growth culture and shaken at 37°C and 250rpm overnight. In the morning, cells were prepared for

sequencing (see below) and glycerol stocks were made by mixing 1ml of bacterial solution with 500ul 80% glycerol – 20% LB solution and stored in -80°C.

2.3.5. Sequencing

Cells were lysed (TritonX100 0.1% in 1XTE: 15µl, culture: 5µl, 99°C for 5 min and 30°C for 5 min) and the DNA from each bin was subjected to PCR with a different 5' primer containing a specific bin-inducer level barcode. PCR products were verified in an electrophoresis gel and cleaned using PCR Clean-Up kit. Equal amounts of DNA (2ng) from 16 bins were joined to one 1.5ml microcentrifuge tube for further analysis, to a total of three tubes. This procedure was conducted three times, one for each RBP-GFP fusions.

Each one of the three samples were sequenced on an Illumina HiSeq 2500 Rapid Reagents V2 50bp 465 single-end chip. 20% PhiX was added as a control. This resulted in ~540 million reads, about 180 million reads per RBP.

2.4. Oligo-library- analysis

Note: the following analysis procedure was conducted three times, one for each RBP.

2.4.1. Read normalization and filtration

Read number was normalized by percentage of bacteria in each bin from the total library, given by the FACS during sorting. This is done in order to be able to compare between numbers of reads of the same variant in different bins.

$$\text{Eq. 1: } N_{reads}(i,j,k) = R_{reads}(i,j,k) \times \%cells(j,k), \quad \begin{matrix} i = 1{:}\,100{,}000 \\ j = 1{:}\,6 \\ k = 1{:}\,8 \end{matrix}$$

where $N_{reads}(i,j,k)$ and R_{reads} are the number of normalized and raw reads per variant, bin, and inducer concentration respectively. *%cells (j,k)* corresponds to the percentages of the cells of variant *i* in each bin per inducer concentration during sorting from the entire library as supplied by the sorter.

Two cut-offs were introduced on the variant read counts: (i) only inducer levels that had above 30 reads for all eight bins were taken into account; and (ii) only variants that had more than 300 reads in total for the entire 6-by-8 matrix were taken into account.

2.4.2. Estimation of mean mCherry levels (μ) per inducer concentration

For each inducer concentration *j*, we have an 8-bin histogram for which we need to calculate the mCherry averaged fluorescence ($\mu(i,j)$). First, for every variant we renormalize N_{reads} by the

total number of reads obtained for that inducer level (each column in the read matrix and color bar, Data Figure 4.2 A-top).

$$\text{Eq. 2: } \widetilde{N}_{reads}(i,j,k) = \frac{N_{reads}(i,j,k)}{\sum_{k=1}^{8} N_{reads}(i,j,k)}, \quad \begin{matrix} i = 1{:}\,100{,}000 \\ j = 1{:}\,6 \\ k = 1{:}\,8 \end{matrix}$$

Next, we convert the bin index (j=1:8) to mCherry fluorescence (*Bin(i,j,k)*). This is done by retrieving the maximum mCherry fluorescence value that was assigned to each bin by the sorter. Then, we compute the cumulative renormalized reads by adding all the normalized reads successively from the lowest to the highest fluorescent bin as follows:

$$\text{Eq. 3: } \widetilde{N}^{cum}_{reads}(i,j,k) = \sum_{l=1}^{k} \widetilde{N}_{reads}(i,j,l), \quad \begin{matrix} i = 1{:}\,100{,}000 \\ j = 1{:}\,6 \\ k = 1{:}\,8 \end{matrix}$$

Finally, to compute μ, we fit the cumulative renormalized read values to a cumulative Gaussian as follows:

$$\text{Eq. 4: } \widetilde{N}^{cum}_{reads}(i,j,k) = 0.5 + 0.5\,\mathrm{erf}\left(\frac{Bin(i,j,k) - \mu(i,j)}{\sigma(i,j)\sqrt{2}}\right), \quad \begin{matrix} i = 1{:}\,100{,}00 \\ j = 1{:}\,6 \\ k = 1{:}\,8 \end{matrix}$$

where *σ(i,j) i*s the standard deviation for mCherry fluorescence extracted from the fitting procedure (see Figure 4.2 A-bottom for sample calculation). Note, only induction levels that had a goodness-of-fit higher than 0.5 were taken into account in the final analysis.

2.4.3. μ normalization and filtration

Since each inducer concentration experiment was carried out in different conditions (e.g. duration of incubation on ice, O/N shaking, binning time) and at a different time (different days), mCherry levels assigned for each bin varied greatly as a function of experiment as well as over-all fluorescence recorded. Therefore, to quantify this systematic error, we first computed a normalized mean fluorescence level (μ_{norm}) per variant as follows:

$$\text{Eq. 5: } \mu_{norm}(i,j) = \frac{\mu(i,j)}{max\{\mu(i,j); j=1:6\}}, \quad \begin{matrix} i = 1{:}\,100{,}000 \\ j = 1{:}\,6 \end{matrix}.$$

To ascertain the scope of the problem presented by the systematic error, we plot in Data S2-B a heat-map of μ_{norm} values consisting of 3000 variants for PCP. Here, low fluorescence was recorded for induction level 1, 4, and 6, while higher levels were recorded for induction levels 2,3, and 5 respectively. These results are consistent with the fact that the induction experiments of level 1,4, and 6 were carried out on the same day, while those of 2,3, and 5 on a separate day.

Next, to accommodate for these systematic discrepancies in our data, for each inducer level we extracted the μ_{norm} for all the negative control variants that were introduced into the OL (220 variants for PCP, 160 variants for MCP and QCP). We then computed the average μ_{norm} for all negative controls per inducer level to obtain $\mu_{neg(j)}$. Finally, we rescaled all $\mu_{norm(i,j)}$ values by $\mu_{neg(j)}$ to eliminate the systematic error from the average fluorescence level as follows:

$$\text{Eq. 6: } \tilde{\mu}_{norm}(i,j) = \frac{\mu_{norm(i,j)}}{\mu_{neg}(j)}, \quad \begin{matrix} i = 1{:}\,100{,}000 \\ j = 1{:}\,6 \end{matrix}.$$

Figure 4.2 shows that this rescaling operation successfully compensated for the systematic error. Note, that since the experiment is based on detecting a repression effect as a function of inducer, we filtered out the variants that displayed averaged mCherry levels at the three lowest concentrations below 15% of the averaged mCherry levels at the three lowest concentrations of the positive control.

2.4.4. Calculating the responsiveness score (R_{score})

To characterize binding to our variants, we compute an empirical score which quantifies how similar a given variant's mCherry levels were to either the positive or negative controls. The score, termed the "responsiveness" (R_{score}), is proportional to the binding affinity K_d (see SI for derivation) provided that the R_{score} obtained for the various negative and positive controls are distributed in a Gaussian fashion.

To derive an expression for the R_{score}, we first compute two n-dimensional probability density functions defining the probability in an n-dimensional space to find either the coat-protein binding or non-binding positive and negative controls, respectively.

$$\text{Eq. 7: } pdf(pos,n) = \frac{exp\left(-\frac{1}{2}(\tilde{\mu}_{norm}(pos,n)-mean(\tilde{\mu}_{norm}(pos,n)))^T\Sigma^{-1}(\tilde{\mu}_{norm}(pos,n)-mean(\tilde{\mu}_{norm}(pos,n)))\right)}{\sqrt{(2\pi)^3|\Sigma|}}, \quad \begin{matrix} pos = positive\ controls \\ n = n_1, n_2, \ldots, n_N \end{matrix}$$

$$\text{Eq. 8: } pdf(neg,n) = \frac{exp\left(-\frac{1}{2}(\tilde{\mu}_{norm}(neg,n)-mean(\tilde{\mu}_{norm}(pos,n)))^T\Sigma^{-1}(\tilde{\mu}_{norm}(neg,n)-mean(\tilde{\mu}_{norm}(pos,n)))\right)}{\sqrt{(2\pi)^3|\Sigma|}}, \quad \begin{matrix} neg = negative\ controls \\ n = n_1, n_2, \ldots, n_N \end{matrix}$$

Where the set $\{n_j\}$ corresponds to N independent parameters by which one can describe the fluorescence measurement of each variant, and $\sum$ is the co-variance matrix. For example, one such set is the six dimensional set corresponding to the fluorescence measurements for each inducer level.

Using these probability density functions, we can compute the probability that an n-dimensional vector (*i*) belongs to each of these distributions, as follows:

$$\text{Eq.9: } \begin{matrix} p(i,pos) \equiv p\left(\tilde{\mu}_{reg}(i,n)|pdf(pos,n)\right) \\ p(i,neg) \equiv p\left(\tilde{\mu}_{reg}(i,n)|pdf(neg,n)\right) \end{matrix},$$

which allows us to define the responsiveness score (R_{score}) as follows:

$$\text{Eq. 10: } R_{score}(i) \equiv log\left(\frac{p(i,pos)}{p(i,neg)}\right).$$

A higher R_{score} indicates a more likely grouping to the coat-protein binding positive control, while a lower score indicates a more likely grouping to the non-binding negative control.

In the analysis carried out in this paper, we chose to reduce the parameter space to a 3-dimensional space consisting of the following components: the *slope (m)* and *goodness-of-fit* (R^2) to a simple linear fit of the rescaled fluorescence $\tilde{\mu}_{norm}(i,j)$ to inducer concentration values. The third component is a standard deviation (*std)* of $\tilde{\mu}_{norm}(i,j)$ computed at the three highest concentration induction bins. We term this new vector:

$$\text{Eq. 11: } \left\{\tilde{\mu}_{norm}(i,j), \begin{matrix} i = 1{:}\,100{,}000 \\ j = 1{:}\,6 \end{matrix}\right\} \rightarrow \left\{\tilde{\mu}_{reg}(i,n), \begin{matrix} i = 1{:}\,100{,}000 \\ n = m, R^2, std \end{matrix}\right\}.$$

Based on the 3-dimensional space- R^2, m, and std- we conducted a multivariant Gaussian fit for the positive and negative control populations (see Figure 4.3), which in turn allowed us to compute the 3-dimensional *pdf(pos,n)* and *pdf(neg,n)*. Finally, we computed the R_{score} for each non-control variant by averaging the score over as many bar-codes which past our filters (each variant appeared in our library 5 times). The results of this computation are presented in the heatmaps of Figure 4.3 and 4.4, which are arranged in accordance with decreasing R_{score}

2.4.5. Calculating *ΔΔG* for high-affinity variants

Up to this point, we have developed the R_{score} to sort the different variants, but did not dive into what it means physically or from a binding perspective. The approach relied on mapping the behavior of the native (wt) binding site and non-binding negative control in some three-dimensional parameter space, and computing the likelihood that a given variant would belong to one or the other group. The R_{score} is the log of the ratio of the two computations. In principle, R_{score} can be computed from any number of probability density functions. We could have used the original 6D space consisting of the 6 inducer concentrations, or chose any other combination. In the computation below, we will map the 6D space to a 1D space of binding affinities that can be in principle computed from each 6-vector using a Hill function fit. In the case of such a mapping, we can replace eqn. 7 and 8 in the paper with the following terms:

Eq. 12:
$$pdf(pos,n) = \frac{1}{\sigma_{pos}\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{K_d^n - K_d^{pos}}{\sigma_{pos}}\right)^2\right), \begin{matrix} pos = positive\ controls \\ n = n_1, n_2, \dots, n_N \end{matrix}$$
$$pdf(neg,n) = \frac{1}{\sigma_{neg}\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{K_d^n - K_d^{neg}}{\sigma_{neg}}\right)^2\right), \begin{matrix} neg = negative\ controls \\ n = n_1, n_2, \dots, n_N \end{matrix}$$

In such a case, the probability for a given variant to have a K_d similar to the native and negative control distributions is given by:

Eq.13:
$$p(i,pos) \equiv p\left(K_d^i | pdf(pos,n)\right)$$
$$p(i,neg) \equiv p\left(K_d^i | pdf(neg,n)\right)$$

We can then compute the R_{score}(i) similar to Eq. 10, which allows us to write:

Eq. 14: $$R_{score}(i) = log\left[\left(\frac{\sigma_{neg}}{\sigma_{pos}}\right)\exp\left(-\frac{1}{2}\left(\frac{K_d^i - K_d^{pos}}{\sigma_{pos}}\right)^2 + \frac{1}{2}\left(\frac{K_d^i - K_d^{neg}}{\sigma_{neg}}\right)^2\right)\right]$$

If we assume for simplicity that $\sigma_{pos} \sim \sigma_{neg} \sim \sigma$ we get:

Eq. 15: $$R_{score}(i) = \frac{K_d^{pos} - K_d^{neg}}{\sigma^2} K_d^i + \frac{(K_d^{neg})^2 - (K_d^{pos})^2}{\sigma^2}$$

which implies that the R_{score}(i) for a given variant is proportional to its K_d.

Finally, we note that the expressions derived in equations 14 and 15 have the following general form to a reasonable first approximation:

Eq. 16: $$R_{score}(i) = a + bK_d^i + O((K_d^n)^2) \cong a + bK_d^i$$

This then allows us to convert any R_{score} value to binding affinity provided that we have a reasonable approximation to a and b.

Given the fact that:

Eq. 17: $$\Delta G = -k_B T \ln K_d,$$

the binding energy can be estimated from Rscore values. We next used the previous study17, which derived the ΔΔG for MCP with over 100k variants, 609 of them were present in our OL variants. We screened for the high affinity variants by setting thresholds of ΔΔG > -6.667 and Rscore > 3.5, which left us with 37 data points. In order to derive the ΔΔG for PCP and QCP using the same equation, we normalized the Rscore values by the mean calculated value for the ms2-wt strain. We then implemented a linear regression and derived a and b. Using these

values, we were able to calculate ΔΔG for every high-affinity variant with all three RBPs. The results of this computation are given in Table S1.

$$\text{Eq. 18: } \Delta\Delta G(i) = ln\frac{\frac{R.score(i)}{R.score(wt)} - a}{b}, i = 1{:}\,100{,}000$$

2.5. Mammalian Cassette microscopy experiments

2.5.1. Construction of mammalian expression plasmids

We ordered three plasmids from addgene containing PCP-3xGFP (#75385), MCP-3xBFP (#75384), and N22-3xmCherry (#75387), and used them to create the following two plasmids: MCP-3xmCherry and QCP-3xBFP. In brief, using two restriction enzymes, BamHI and MluI, we restricted the plasmids and conducted PCR with the same restriction sites added as primers on both MCP and QCP. After PCR purification, we restricted the product with the same two enzymes and ligated them to the matching plasmids. Then, we performed transformation to Top10 *E.coli* cells and screened for positive clones. All plasmids used in the microscopy experiments were sequence-verified via Sanger sequencing.

RNA binding site cassettes were ordered from IDT as g-blocks (see Table S4 for sequences). Restricted and ligated them to a vector downstream of a CMV promoter using the restriction enzyme EcoRI. Then, we performed transformation to Top10 *E.coli* cells and screened for positive clones. All plasmids used in the microscopy experiments were sequence-verified via Sanger sequencing.

2.5.2. Mammalian Microscopy assay

Cell culture:

The Human Bone Osteosarcoma Epithelial Cell line was incubated and maintained in 100x20mm cell culture dishes under standard cell culture conditions at 37°C in humidified atmosphere containing 5% CO_2 and were passaged at 80-85% confluence. Cells were washed once with 1x PBS, and subsequently treated with 1mL trypsin/EDTA (ethylenediaminetetraacetic acid, Biological Industries) followed by incubation at 37°C for 3-5 minutes. DMEMcomplete, complemented with 10% FBS and final concentrations of 100U penicillin plus 100µg streptomycin, was added and transferred into fresh DMEMcomplete in subcultivation ratios of 1:10.

Fluorescent microscopy experiments:

Before the experiment, U2OS cells were seeded on 60mm glass-bottom imaging dishes. Transient transfection was performed with Polyjet (Invivogen) transfection reagent according to the manufacture's instructions. Typical DNA for transfection was 150ng from RBP-3xFP and 850ng from the cassette plasmid. After inoculation for 24-48 hours, the growth medium was removed and replaced with Leibovitz L15 medium with 10% FBS. During microscopy, the sample was kept at 37°C.

Microcopy was carried out on a Nikon Ti-E eclipse epifluorescent microscope. Images were taken with a 40X oil immersion objective and the following excitation lasers: 585nm for mCherry, 490nm for GFP, 400nm for BFP. The images were recorded with the Xion EMCCD camera. The microscope was controlled with NIS Elements imaging software. Time-lapse movies of a single Z-plane were recorded with, 1500ms exposure time and time intervals between frames were 30 seconds.

2.6. Shape-Seq

2.6.1. Experimental setup

LB medium supplemented with appropriate concentrations of Amp and Kan was inoculated with glycerol stocks of bacterial strains harboring both the binding-site plasmid and the RBP-fusion plasmid and grown at 37°C for 16 hours while shaking at 250 rpm. Overnight cultures were diluted 1:100 into SPM. 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 SPM. 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) [92], 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 (both from 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 [93]. 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 [94], including reverse transcription (steps 40-51), adapter ligation and purification (steps 52-57) as well as dsDNA sequencing library preparation (steps 68-76). 1000 ng of RNA were converted to cDNA using the reverse transcription primers (for details of primer and adapter sequences used in this work see Table S3) for mCherry (#1) or 5S rRNA (#2) that are specific for either the mCherry transcripts (PP7-USs δ=-29, PP7-wt δ=-29). The RNA was mixed with 0.5 µM primer (#1) or (#2) 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 minutes, 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 (#3) (see Table S3) was ligated to the cDNA using a CircLigase 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 (#3), 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.

2.6.2. SHAPE-Seq library preparation and sequencing

To produce the dsDNA for sequencing 10ul of purified cDNA from the SHAPE procedure (see above) were PCR amplified using 3 primers: 4nM mCherry selection (#4) or 5S rRNA selection primer (#5), 0.5µM TruSeq Universal Adapter (#6) and 0.5µM TrueSeq Illumina indexes (one of #7-26) (Table S3) with PCR reaction mix (1x Q5 HotStart reaction buffer, 0.1 mM dNTPs, 1 U Q5 HotStart Polymerase [NEB]). A 15-cycle PCR program was used: initial denaturation at 98°C for 30 s followed by a denaturation step at 98°C for 15 s, primer annealing at 65°C for 30 s and extension at 72°C for 30 s, 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. See Table S4 for read counts for all experiments presented in the manuscript.

Part 3: Findings

1. Katz, N.*, Cohen, R.*, Solomon, O., Kaufmann, B., Atar, O., Yakhini, Z., Goldberg, S., and Amit, R. **An in Vivo Binding Assay for RNA-Binding Proteins Based on Repression of a Reporter Gene**. *ACS Synth. Biol*. 2018, 7, 12, 2765-2774. A*ccepted for journal supplementary cover art*.

2. Katz, N.*, Cohen, R.*, Atar, O., Goldberg, S., Amit, R. **An Assay for Quantifying Protein-RNA Binding in Bacteria**. *J. Vis. Exp.* (148), e59611, doi:10.3791/59611 (2019).

3. Katz, N., Cohen, R., Solomon, O., Kaufmann, B., Atar, O., Yakhini, Z., Goldberg, S., and Amit, R. (2019). **Synthetic 5' UTRs Can Either Up- or Downregulate Expression upon RNA-Binding Protein Binding**. *Cell Systems*. 10.1016/j.cels.2019.04.007, 93-106.e8.

ACS SyntheticBiology
Cite This: ACS Synth. Biol. XXXX, XXX, XXX−XXX
Letter
pubs.acs.org/synthbio

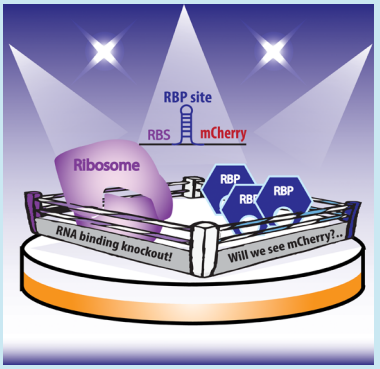
An *in Vivo* Binding Assay for RNA-Binding Proteins Based on Repression of a Reporter Gene

Noa Katz,[†,#] Roni Cohen,[†,#] Oz Solomon,[†,§] Beate Kaufmann,[†] Orna Atar,[†] Zohar Yakhini,[‡,§] Sarah Goldberg,[†] and Roee Amit*[,†,∥]

[†]Department of Biotechnology and Food Engineering, Technion − Israel Institute of Technology, Haifa 32000, Israel
[‡]Department of Computer Science, Technion − Israel Institute of Technology, Haifa 32000, Israel
[§]School of Computer Science, Interdisciplinary Center, Herzeliya 46150, Israel
[∥]Russell Berrie Nanotechnology Institute, Technion − Israel Institute of Technology, Haifa 32000, Israel

Supporting Information

ABSTRACT: We study translation repression in bacteria by engineering a regulatory circuit that functions as a binding assay for RNA binding proteins (RBP) *in vivo*. We do so by inducing expression of a fluorescent protein−RBP chimera, together with encoding its binding site at various positions within the ribosomal initiation region (+11−13 nt from the AUG) of a reporter module. We show that when bound by their cognate RBPs, the phage coat proteins for PP7 (PCP) and Qβ (QCP), strong repression is observed for all hairpin positions within the initiation region. Yet, a sharp transition to no-effect is observed when positioned in the elongation region, at a single-nucleotide resolution. Employing *in vivo* Selective 2′-hydroxyl acylation analyzed by primer extension followed by sequencing (SHAPE-seq) for a representative construct, established that in the translationally active state the mRNA molecule is nonstructured, while in the repressed state a structured signature was detected. We then utilize this regulatory phenomena to quantify the binding affinity of the coat proteins of phages MS2, PP7, GA, and Qβ to 14 cognate and noncognate binding sites *in vivo*. Using our circuit, we demonstrate qualitative differences between *in vitro* to *in vivo* binding characteristics for various variants when comparing to past studies. Furthermore, by introducing a simple mutation to the loop region for the Qβ-wt site, MCP binding is abolished, creating the first high-affinity QCP site that is completely orthogonal to MCP. Consequently, we demonstrate that our hybrid transcriptional−post-transcriptional circuit can be utilized as a binding assay to quantify RNA−RBP interactions *in vivo*.



KEYWORDS: *RNA binding protein (RBP), MS2, PP7, phage coat protein, binding assay, post-transcriptional regulation, SHAPE-seq, translation repression, synthetic circuit*

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. RNA hairpins have been studied in three distinct positions: either immediately downstream of the AUG,[1] upstream of the Shine−Dalgarno sequence,[2] or as structures that entrap Shine−Dalgarno motifs, as is the case for the PP7 and MS2 phage coat-protein binding sites. 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.

In recent years, advances in next generation sequencing (NGS) technology combined with selective nucleic acid probing approaches have facilitated focused study of specific RNA structures *in vivo*. These chemical-modification approaches[3−7] can generate a "footprint" of the dynamical structure of a chosen RNA molecule *in vivo*, while in complex with ribosomes and/or other RBPs. In parallel, 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 characterizing parts that regulate transcription,[8−11] their potential for deciphering post-transcriptional regulatory mechanisms have been demonstrated in a recent study that interrogated IRES sequences in mammalian cells.[12]

Building on these advances and on the development of a translational repression circuit that was used to characterize the binding characteristics of the RBP L7Ae in both bacteria and mammalian cells,[13] we engineering a hybrid transcriptional−post-transcription circuit that was designed to be a general platform for characterizing RBP binding *in vivo*. Using the circuit, we measured the regulatory output of a small library of synthetic constructs in which we systematically varied the

Received: September 10, 2018
Published: November 8, 2018

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

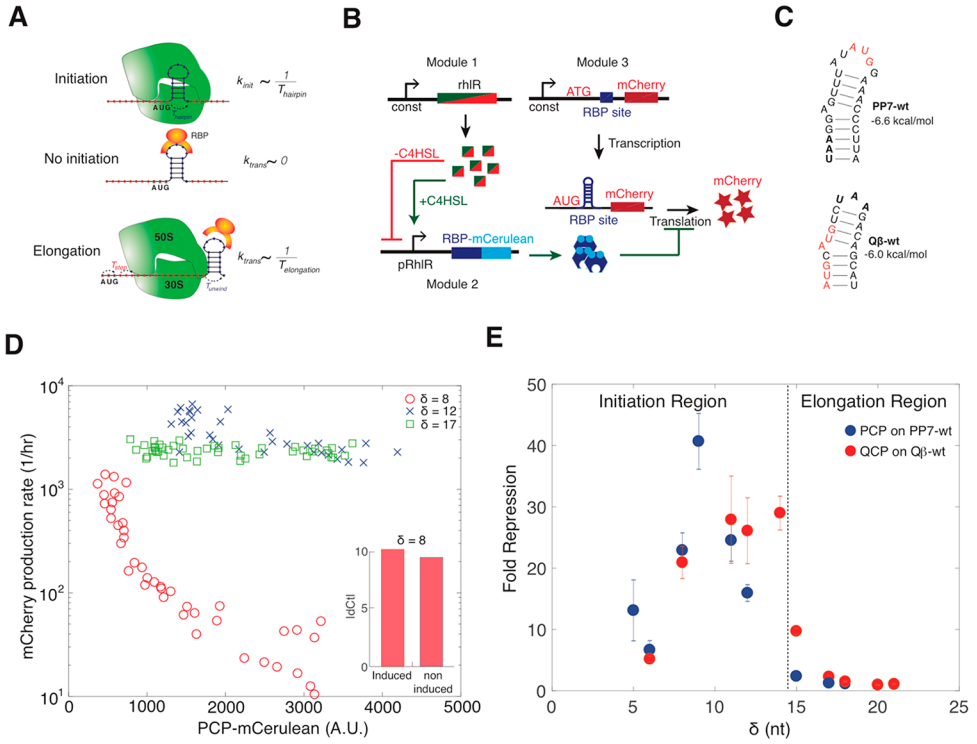


Figure 1. Translational regulation by a RBP-hairpin complex in the ribosomal initiation region. (A) A schematic for the hypothesized repression mechanism. The position of the hairpin within the ribosomal initiation region dictates the rate initiation $T_{hairpin}$, which in turn may control the rate of translation (top). When bound by an RBP (middle) the hairpin−RBP complex is able to disrupt initiation, thus inhibiting translation. If the hairpin is positioned downstream of the initiation region (bottom), initiation and subsequent elongation is likely to occur, leading to unwinding of the RBP-hairpin complex by the ribosome. (B) Gene regulatory circuit: (left-top) transducer plasmid−module 1: rhlR expression cassette; (left-bottom) transducer plasmid−module 2: RBP-mCerulean expression cassette under the control of pRhlR; (right-top) reporter plasmid−module 3: mCherry reporter expression under the control of a constitutive promoter; and (right-bottom) resultant mRNA encoding a folded RBP binding site with the ribosomal initiation region. When the binding site is occupied by the RBP, translation repression ensues. (C) The two hairpins used in this experiment were the native (wt) binding sites for the PP7and Qβ coat proteins. Stop codons and start codons inside the binding sites are highlighted, in bold and red. Note, positions where stop codons are in-frame were not tested, and so are most of the start codons. For those start codons that are in-frame—Qb at the second position in each frame—no different response was generated compared with the other strains, supporting a lack of detectable effect for the second in-frame AUG. (D) Dose−response functions for PCP with a reporter mRNA encoding PP7-wt at three positions: δ = 8 (red), δ = 12 (blue), and δ = 17 (green) nt. Inset: quantitative RT-PCR results for mRNA levels for the PP7-wt δ = 8 with and without induction. (E) Fold-repression measurements for PCP (blue) and QCP (red) as a function of hairpin position δ. Fold repression is computed by the ratio of the mCherry rate of production at no induction to the rate of production at full induction. Note, for three constructs (PCP with δ = 14, and QCP with δ = 5 and δ = 9) the basal levels without induction were too low for fold-repression measurements.

position and type of RBP binding sites. In addition, we applied Selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq)[14,15,6] to a single variant, to further characterize RBP-based regulatory mechanisms in bacteria. Our findings indicate that the chosen hairpin-binding RBPs (coat proteins from the bacteriophages GA,[16] MS2,[17] PP7,[18] and Qβ[19]), generate a strong repression response when bound to the translation initiation region. This inhibitory response is associated with RNA-restructuring that spans a large segment of the RNA, including both the RBP binding site and the RBS. We employed this strong repression phenomenon as an *in vivo* binding assay for RBP−RNA interactions. Using our synthetic regulatory circuit as a binding assay, we quantitatively characterized RBP binding affinity to a set of mutated binding sites in a high-throughput manner, thereby increasing our understanding of RBP−RNA binding *in vivo* and enabling the engineering of more complex RNA-based applications.

RESULTS AND DISCUSSION

RBPs Repress Translation When Bound within δ < 15 from the AUG. We hypothesized (Figure 1A) that a hairpin may be tolerated within the ribosomal initiation region facilitating translation if sufficiently unstable, but once bound by an RBP, initiation will be inhibited leading to a translational repression effect. To test this hypothesis, we designed a trimodule transcriptional and post-transcriptional gene regulatory circuit that was encoded on two plasmids (transducing and reporting) that were simultaneously transformed into *E. coli* (Figure 1B). The transducing plasmid (Figure 1B-top) encoded a rhlR gene under the control of a constitutive promoter on the first module, and either the phage coat protein for PP7 (PCP) or Qβ (QCP) fused to mCerulean, under the control of a pRhlR promoter inducible by *N*-butanoyl-L-homoserine lactone (C_4−HSL) on the second module. The reporter plasmid initially encoded the two wild-type binding sites (PP7-wt and Qβ-wt) for PCP and QCP at several positions downstream to the AUG of an *mCherry* reporter gene. The two native binding sites (Figure 1C) are characterized by hairpins of a varying length, which are interrupted by a single unpaired nucleotide or "bulge", and comprise a loop of either size 3 nt (Qβ-wt) or 6 nucleotides (PP7-wt). We constructed 12 variants for each binding site

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

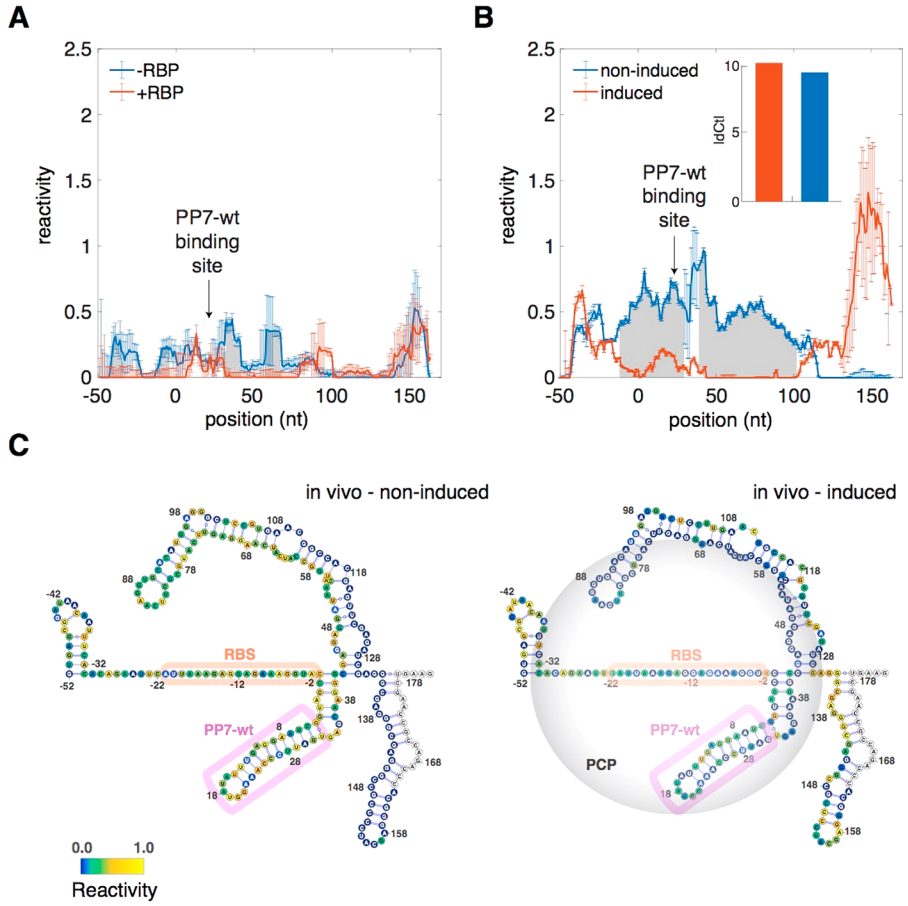


Figure 2. SHAPE-seq analysis of the PP7-wt binding site in the absence and in the presence of RBP. (A) *In vitro* reactivity. Scores for the SHAPE-seq reactions carried out on refolded mCherry reporter mRNA molecules containing a PP7-wt binding site at $\delta = 6$ with (red) and without (blue) a recombinant PCP present in the reaction buffer. (B) *In vivo* reactivity. Scores for the SHAPE-seq reactions carried out *in vivo* on the PP7-wt $\delta = 6$ construct with the PCP-mCerulean protein noninduced (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 noninduced (B) cases, respectively. Error bars were computed using boot-strap resampling and subsequent averaging over two biological replicates. See also Figure S4 and associated discussion for comparison of results using our reactivity definition with another reactivity analysis using a model-based approach. (C) Structural schematics of the segment of the PP7-wt $\delta = 6$ 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 noninduced (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. Noncolored bases correspond to position of the reverse transcriptase primer.

type starting at $\delta = 5$ exploring every single position until $\delta = 21$, except for those where an internal hairpin stop codon and most of the internal start codons were in frame (see note in figure caption). Each transducer−reporter plasmid pair was transformed into *E. coli* TOP10 and grown in 24 different C_4−HSL concentrations, in duplicate. Optical density, mCherry, and mCerulean fluorescence levels were measured at multiple time points for each inducer concentration. From these data, mCherry production rates[20,21] were computed over a 2−3 h window (see Supporting Methods and Figure S1) for each inducer level, and mCerulean levels were averaged over the same time frames. In Figure 1D we plot a series of dose−response curves obtained for PCP on three constructs containing the PP7-wt binding site, positioned at $\delta = 8$ (red), 12 (blue), and 17 (green) nt. To first rule out that the repression response stems from different number of RNA transcripts or degradation-related effects, we checked that the RNA levels at both states were similar using quantitative real-time PCR (Figure 1D-inset). For the hairpin located at $\delta = 8$, the mCherry production rate is reduced by nearly 2 orders of magnitude as a function of RBP concentration, while the hairpin positioned at $\delta = 12$ produced a weakly repressing dose−response function, and no RBP-induced repression was observed at $\delta = 17$.

Next, we computed the fold repression, defined as the ratio of mCherry production rate at no induction to full induction (*i.e.*, low to high mCerulean fluorescence levels), measured for the PCP on PP7-wt constructs. We plot the results for PCP in Figure 1E (blue circles). The figure shows that strong repression is triggered by PCP induction for all available positions in the region demarked by $\delta < 15$ (dashed line). However, fold repression by PCP rapidly diminishes for $\delta \geq 15$, and seems to disappear for $\delta \geq 17$ positions for all constructs. To show that this repression phenomenon was not limited to the PCP−PP7-wt interaction, we tested the translation repression effect generated by the QCP-mCerulean protein when induced in the presence of a reporter gene encoding the Qβ-wt binding site at various positions. We plot

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

the results for the QCP-induced fold repression in Figure 1E (red). The results show a similar fold-repression response behavior for QCP to that observed for PCP with strong repression observed for $\delta < 15$, and a rapid decline for $\delta > 15$ positions. Consequently, our data indicates that the region immediately downstream to the AUG and up to $\delta \sim 15$ seems to be susceptible to interference with translation making it a "hot spot" for potential translational repression mechanisms.

***In Vitro* SHAPE-seq Reveal an Extended Protected Region by PCP.** To provide a structural perspective on the inhibition mechanism triggered by the RBP binding to their hairpin binding sites, we employed SHAPE-seq. Specifically, we used acylimidazole reagent 2-methylnicotinic acid imidazolide (NAI), which modifies the 2′ OH of non- or less-structured, accessible RNA nucleotides as found in single-stranded RNA molecules.[14] We hypothesized that SHAPE-Seq data can provide a protection footprint (as in Smola *et al.*[22]) that develops when the RBP is bound to its cognate binding site. SHAPE-seq is a next generation sequencing approach (see Materials and Methods and Figure S2 for details), whereby an insight into the structure of an mRNA molecule can be obtained *via* selective modification of "unprotected" RNA segments. "Unprotected" segments mean single-stranded nucleotides that do not participate in any form of interaction, such as Watson−Crick base-pairing 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 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 unbound (*i.e.*, single-stranded). The single nucleotide propensity for modification is then calculated to a value that is referred to as "reactivity" score, which is computed from the ratio of the normalized modified to unmodified read count (see Supporting Information for details).

In our version of the reactivity score, any negative values are set to 0, indicating that the nucleotides at those particular positions do not get modified. We used boot-strapping statistics (as in refs 23, 24) and *Z*-factor analysis (as in refs 22, 25, 26; see Supporting Information for definition) to identify the regions on the RNA molecule where the observed differences between the signals at +RBP and −RBP are statistically significant (equal to or more than three sigmas). Finally, to eliminate method bias, we repeated the reactivity analysis on all our data sets using a model-based analysis approach.[23,24] In all cases studied the reactivity results from both methods being in good agreement (see Figure S4 and associated discussion).

In Figure 2A, we present the results for the reactivity analysis carried out on the *in vitro* SHAPE-seq data for the PP7-wt $\delta = 6$ construct with (red line, +RBP) and without (blue line, −RBP) the presence of a recombinant PCP protein in the reaction solution. Reactivities are presented as a running average over a 10 nt window to eliminate high frequency noise (for further details about the analysis pipeline, see Supporting Information and Figure S3). The *in vitro* modification experiments were carried out after refolding of the RNA followed by 30 min incubation at 37 °C with or without the recombinant PCP, and subsequently modified by the SHAPE reagent (*i.e.*, NAI). 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 noninteracting 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. This indicates that the nucleotides that flank the binding site (positions 6−30 nt) are sequestered and are unmodified or unreactive. We used *Z*-factor analysis to determine the sequence segments (gray shade) where a statistically significant reduction in reactivity, between the + and −RBP cases, can be observed. These segments span a range $\sim \pm 50$ nts from the position of the binding site, consistent with a previous RNase-based *in vitro* study.[27] In contrast, for the positions spanning the range 70−180 nt, the reactivities for both + and − cases are indistinguishable. Together, the reactivity analysis indicates that the RBP is protecting a wide-swath of RNA, which spans the 5′ UTR, the initiation, and a portion of the elongation region. This protection is alleviated for positions that are distal from the binding site by >50 nts, resulting in a realigned reactivity signature indicating that a similar underlying structure for the RNA molecule is maintained for both reaction conditions.

***In Vivo* SHAPE-seq Measurements Are Consistent with *in Vitro* Measurements.** To confirm the observations of the *in vitro* SHAPE-seq protection footprint, we carried out an *in vivo* SHAPE-seq experiments (see Materials and Methods for differences from the *in vitro* protocol) on the PP7-wt $\delta = 6$ construct at two induction states (Figure 2B): 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. We plot in Figure 2B the reactivity results for both the induced (red) and noninduced (blue) cases. For the noninduced case, we observe a strong reactivity signal (>0.5) over the range spanning −45−110 nts, 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. Interestingly, both for the *in vivo* induced and the +RBP *in vitro* cases (orange signals), the region in the signal corresponding to the protein occupied binding site (arrow point down) seems to be slightly more sensitive to modifications in comparison with the adjacent regions. 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 noninduced reactivity was significantly larger than the induced-reactivity. This shaded region flanks the binding site by ∼50 nts both upstream and downstream and is consistent with an interpretation of a wide-swath of PCP protected RNA *in vivo*.

A closer examination of the *in vivo* SHAPE-seq data reveals two major differences from the *in vitro* SHAPE-seq First, the noninduced 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 noninduced signal becomes sharply nonreactive over this range. To gain a structural perspective for the extent of these differences, we plot in Figure 2C two structures. The structures were

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

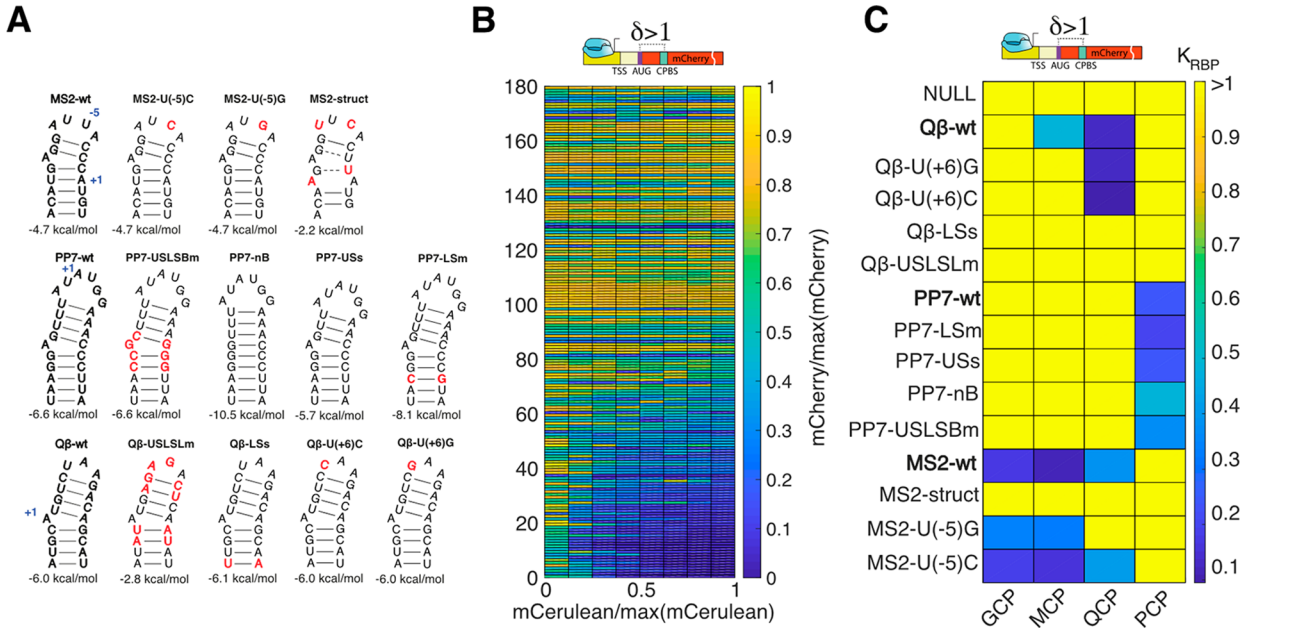


Figure 3. Repression effect can be used to estimate an effective dissociation constant K_{RBP}. (A) Structural schematic for the 14 binding sites used in the binding affinity study. Red nucleotides indicate mutations from the original wt binding sequence. Abbreviations: US/LS/L/B = upper stem/lower stem/loop/bulge, m = mutation, s = short, struct = significant change in binding site structure. (B) Dose responses for 180 variants whose basal rate of production levels were >50 au/h. Each response is divided by its maximal mCherry level, for easier comparison. Variants are arranged in order of increasing fold up-regulation. (C) Normalized K_{RBP} for variants that generated a detectable down-regulatory effect for at least one position. Dark blue corresponds to low K_{RBP}, while yellow indicates high K_{RBP}. If there was no measurable interaction between the RBP and binding site, K_{RBP} was set to 1.

computed using RNAfold[28] for the sequence of this molecule and overlaid by its *in vivo* noninduced (left structure) or induced (right structure) reactivity scores (depicted by a heatmap). 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). The structures reveal that the reactivity for the noninduced case is inconsistent with the structural prediction. This observation is suggestive of a structure-destabilizing role that an initiating 30S subunit may be generating in the 5′UTR and initiation region. A structural role for the ribosome can be further inferred by the complete lack of reactivity observed deeper in the elongation region of the noninduced case, which is consistent with the presence of a chain of translating ribosomes that may be protecting the RNA from modifications. This is supported by the recovery of the reactivity signal in the elongation region for the induced case, where translation is for the most part abolished. Consequently, the SHAPE-seq analysis *in vivo* reveals significant structural differences between the induced and noninduced cases that are consistent with their RBP-bound states, resultant translational level, and the observed post-transcriptional repression.

Effective Dissociation Constant of RBPs Is Insensitive to Binding-Site Position. Given the strong RBP-induced repression phenomenon observed for the $\delta < 15$ region, we hypothesized that we can use this effect to further characterize the binding of the RBPs to structured binding sites. To do so, we constructed a set of mutated binding sites with various structure-modifying and nonstructure-modifying mutations [compare Figure 3A: bold letters highlighting the native sites for MCP (MS2-wt, top-left), PCP (PP7-wt, middle-left), and QCP (Qβ-wt, bottom-left)]. The mutated binding sites for MCP and PCP were taken from refs 29, 30, and 18, respectively (Figure 3A), while the ones for QCP were devised by us. All mutations are highlighted in red letters. We then constructed two to four new constructs for each mutated binding site that differed in binding-site position downstream to the AUG. In addition, we constructed a set of control plasmids that lacked a hairpin within the N-terminus of the mCherry reporter gene. Altogether, we constructed 27 additional hairpin-reporter plasmids and 10 no-hairpin controls (see Table S1). The new constructs, and the ones previously tested (Figure 1B, 61 in total), were cotransformed with all four RBP plasmids to yield 232 RBP−binding site strains (*i.e.*, not all potential binding site−RBP pairs were covered). Our goal with this design was to test not only the binding affinity to the native RBPs, but also the relative affinity to the other RBPs, thus obtaining an estimate for the selectivity of RBP binding.

We plot the dose−response curves of 180 out of the 232 strains as a heatmap in Figure 3B (strains with basal mCherry rate of production <50 au/h were excluded). In all cases, the data for both the mCherry rate of production and mean mCerulean levels are normalized by the respective maximal value. The dose response functions are arranged in accordance with fold-regulation of the response, with the most repressive variants positioned at the bottom, and the least repressive at the top. The data show that there is a substantial subset of strains, which exhibit strong repression for at least one hairpin position (~50 variants), with the strongest mCherry signal occurring at the lowest mCerulean level. To obtain an estimate for the effective binding affinity for each down-regulating variant, we fitted each dose−response curve that exhibited a typical repression response (see Figure S1) with a Hill-function-based model (see Supporting Methods), which assumes a simple relationship between the concentration of RBP measured by its fluorescence, the dissociation constant,

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

and the output expression rate. Finally, we normalized the resulting dissociation constant by the maximal mCerulean expression for the matching RBP to facilitate comparison of the results for the different proteins, yielding an effective dissociation constant (K_{RBP}, see Table S5). Typical error in estimation of the effective dissociation constant was 5−20%, and by averaging K_{RBP} of each RBP−binding site pair over multiple positions (values of δ) we obtained estimated errors of ∼10%.

In Figure 3C, we plot the averaged K_{RBP} for different RBP−binding site combinations as a heatmap, only for those sites (Figure 3A) for which all four RBPs were tested ("null" corresponds to an average K_{RBP} computation made on several of the non-binding-site controls). The data show that the effective dissociation constants measured for native sites with their cognate RBPs were low and approximately equal, indicating that native sites are evolutionarily optimized for binding (blue squares). Mutated sites which retained binding affinity displayed slightly larger dissociation constants (light-blue/turquoise), while the K_{RBP} values of RBP-binding site combinations that did not generate a binding signature were set to the maximum normalized value 1 (K_{RBP} > 1, yellow). When examining the data more closely, we found that PCP is completely orthogonal to the MCP/QCP/GCP group, with no common binding sites. Conversely, we observed crosstalk between the different members of the MCP/QCP/GCP group, with increased overlap between MCP and GCP, which is consistent with previous studies.[16]

A closer look at the mutant binding sites reveals that structure-conserving mutations to native binding sites in the loop area [Qβ-U(+6)G, Qβ-U(+6)C, MS2-U(−5)C and MS2-U(−5)G] or stem (PP7-USLSBm and PP7-LSs) did not seem to affect binding of the cognate protein. However, the interaction with a noncognate RBP is either diminished or eliminated altogether as is the case for MCP with Qβ-U(+6)G and Qβ-U(+6)C, and for QCP with MS2-U(−5)G. In addition, putative structure-altering (MS2-struct, where the lower stem is abolished) and destabilizing (Qb-USLSLm, where the GC base-pairs are converted to UA base pairs in the lower stem) mutations significantly affected binding. Finally, structure-altering mutations, which retain apparent binding site stability (PP7-nB and PP7-USs), also seemed to retain at least a partial binding affinity to the native RBP. Altogether, these results suggest that binding sites positioned within the δ < 15 nt region can tolerate multiple mutations as long as certain key structural features necessary for binding and hairpin stability (*e.g.*, loop size) are conserved, as was previously observed *in vitro*.[18,30−32]

DISCUSSION

Synthetic biology approaches have been increasingly used in recent years to map potential regulatory mechanisms of transcriptional and translational regulation, in both eukaryotic and bacterial cells. In this work, we built on the work of ref 13 to design a hybrid transcriptional and post-transcriptional regulatory circuit to quantitatively study RBP-based regulation in bacteria using a combined synthetic biology and SHAPE-seq approach. Using our library of RNA regulatory variants, we were able to identify and characterize a position-dependent repression of translation when the hairpin was bound by an RBP. The extent of the repression effect was strongly dependent on position, and diminished for δ > 15. The localization of a strong inhibition effect to region nearby the AUG for at least two different RBP-hairpin pairs suggests that this region may be particularly susceptible for repression effects. Previous works[33,34] have provided evidence that the ribosomal initiation region extends from the RBS to about 9−11 nucleotides downstream of the AUG (δ = 12 to δ = 14 as in our coordinate system). Furthermore, these authors also showed that structured stems of 6 bp or longer in the N-terminus can silence expression up to +11−13 from the AUG, but show negligible silencing when positioned further downstream. Thus, the region where the strong regulatory effects were detected in our experiments likely overlaps with the presumed ribosomal initiation region. This suggests that translation initiation may be susceptible to regulation, which can be an important guideline for RNA-based synthetic biology circuit design.

The sensitivity of the initiation region to translation regulation is further supported by SHAPE-seq reactivity analysis using both a signal-to-noise and a model-based approach. For both *in vitro* and *in vivo* experiments, 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*[27] and for other proteins with their cognate RNA target using SHAPE-MaP.[22] There are several scenarios, which may explain this result. In one scenario, PCP may form a large multiprotein complex that is anchored to the binding site, which in turn can lead to a wide protected segment on the RNA. Such a scenario can stem from the capsid-forming characteristics of PCP, even though *PCP-delF-G* was the version used in all experiments, which lack the component that is associated with multidimerization. Alternatively, PCP binding may trigger refolding of flanking regions to form structures with fewer noninteracting 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.

The strong fold repression effect generated by the RBP within the initiation region allowed us to characterize the specific *in vivo* interaction of each RBP−binding site pair by an effective K_{RBP}, which we found to be independent of binding site location. Interestingly, the *in vivo* K_{RBP} measured for some of the binding sites relative to their native site, differ from past *in vitro* and *in situ* measurements. In particular, PP7-nB,PP7-USs, and MS2-U(−5)G exhibited little or no binding in the *in vitro* setting,[18,30] yet displayed strong binding in our assay, while MS2-U(−5)C exhibited a reverse behavior—very high affinity *in vitro* and lower affinity in our assay.[30] Finally, MS2-struct showed no binding in our assay, but exhibited an affinity higher to that of the wild type in an *in situ* setting.[29] These discrepancies may be due to structural constraints, as our *in vivo* RNA constructs were significantly longer than what was used previously *in vitro* and included a 700 nt reporter gene. Another reason for these differences may stem from variations in structure of RNA molecules that emerges from their presence inside cells. Our SHAPE-seq analysis revealed that for at least the one construct that was characterized, a transla-

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

tionally active mRNA molecule is less structured *in vivo* as compared with its counterpart *in vitro*. This phenomenon was also previously observed in other studies.[35−37] Such structural differences may lead to intramolecular interactions that yield stable folded states *in vivo* that are more amenable to binding as compared with the short constructs that were used in the *in vitro* experiment, and vice versa.

Finally, we found that both MCP and QCP can bind binding sites with different loop sizes than the wild-type binding sites with relatively high affinity. While they do not seem to be sensitive to the sequence content for a loop whose size is equal to the cognate loop (*i.e.*, 4 nt for MCP and 3 nt for QCP), sequence sensitivity is observed for noncognate loop sizes for both RBPs. This implies that either [GCP, QCP, and PCP] or [MCP, QCP, and PCP], are capable of binding mutually orthogonal binding sites that differ in structure, opening the door for smart design of mutated binding sites for applications where either set of the three RBPs can be used simultaneously. Our work thus establishes a blueprint for an *in vivo* assay for measuring the dissociation constant of RBPs with respect to their candidate binding sites in a more natural *in vivo* setting. This assay can be used to discover additional binding sites for known RBPs, which could be utilized in synthetic biology applications where multiple nonidentical or orthogonal binding sites are needed.

MATERIALS AND METHODS

Design and Construction of Binding-Site Plasmids. Binding-site cassettes (see Table S1) were ordered either as double-stranded DNA minigenes from Gen9 or as cloned plasmids (minigene + vector) from Twist Biosciences. Each minigene was ∼500 bp long and contained the parts in the following order: EagI restriction site, ∼40 bases of the 5′ end of the Kanamycin (Kan) resistance gene, pLac-Ara promoter, ribosome binding site (RBS), an RBP binding site, 80 bases of the 5′ end of the mCherry gene, and an ApaLI restriction site. As mentioned, each cassette contained either a wild-type or a mutated RBP binding site (see Table S1), at varying distances downstream to the RBS. All binding sites were derived from the wild-type binding sites of the coat proteins of one of the four bacteriophages MS2, PP7, GA and Qβ. For insertion into the binding-site plasmid backbone, they were double-digested with EagI-HF and ApaLI (New England Biolabs [NEB]). The digested minigenes were then cloned into the binding-site backbone containing the rest of the mCherry gene, terminator, and a Kanamycin resistance gene, by ligation and transformation into *E. coli* TOP10 cells (ThermoFisher Scientific). Purified plasmids were stored in 96-well format, for transformation into *E. coli* TOP10 cells containing one of four fusion-RBP plasmids (see below).

Design and Construction of Fusion-RBP Plasmids. RBP sequences lacking a stop codon were amplified *via* PCR of either Addgene or custom-ordered templates (Genescript or IDT, see Table S2). All RBPs presented (MCP, PCP, GCP, and QCP) were cloned into the RBP plasmid between restriction sites *Kpn*I and AgeI, immediately upstream of an mCerulean gene lacking a start codon, under the pRhlR promoter (containing the *rhlAB* las box[38]) and induced by C_4−HSL. The backbone contained an Ampicillin (Amp) resistance gene. The resulting fusion-RBP plasmids were transformed into *E. coli* TOP10 cells. After Sanger sequencing, positive transformants were made chemically competent and stored at −80 °C in 96-well format.

Transformation of Binding-Site Plasmids. Binding-site plasmids stored in a 96-well format were simultaneously transformed into chemically competent bacterial cells containing one of the RBP-mCeulean plasmids. After transformation, cells were plated using an 8-channel pipettor on 8-lane plates (Axygen) containing LB-agar with relevant antibiotics (Kan and Amp). Double transformants were selected, grown overnight, and stored as glycerol stocks at −80 °C in 96-well plates (Axygen).

RNA Extraction and Reverse-Transcription for qPCR Measurements. Starters of *E. coli* TOP10 containing the relevant constructs on plasmids were grown in LB medium with appropriate antibiotics overnight (16 h). The next morning, the cultures were diluted 1:100 into fresh semipoor medium and grown for 5 h. For each isolation, RNA was extracted from 1.8 mL of cell culture using standard protocols. Briefly, cells were lysed using Max Bacterial Enhancement Reagent followed by TRIzol treatment (both from 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. RNA quality was assessed by running 500 ng on 1% agarose gel. After extraction, RNA was subjected to DNase (Ambion/Life Technologies) and then reverse-transcribed using MultiScribe Reverse Transcriptase and random primer mix (Applied Biosystems/Life Technologies). For qPCR experiments, RNA was isolated from three individual colonies for each construct.

qPCR Measurements. Primer pairs for mCherry and normalizing gene *idnT* were chosen using the Primer Express software and aligned using BLAST[39] (NCBI) with respect to the *E. coli* K-12 substr. DH10B (taxid:316385) genome (which is similar to TOP10) to avoid off-target amplicons. qPCR was carried out on a QuantStudio 12K Flex machine (Applied Biosystems/Life Technologies) using SYBR-Green. Three technical replicates were measured for each of the three biological replicates. A C_T threshold of 0.2 was chosen for all genes.

***In Vivo* SHAPE-seq.** LB medium supplemented with appropriate concentrations of Amp and Kan was inoculated with glycerol stocks of bacterial strains harboring both the binding-site plasmid and the RBP-fusion plasmid (see Table S3 for details of primers and barcodes, and Figure S2), and grown at 37 °C for 16 h while shaking at 250 rpm. Overnight cultures were diluted 1:100 into semipoor medium. Each bacterial sample was divided into a noninduced 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 semipoor medium supplemented with a final concentration of 30 mM 2-methylnicotinic acid imidazole (NAI) suspended in anhydrous dimethyl sulfoxide (DMSO, Sigma-Aldrich),[6,14] 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. Column-based RNA isolation (RNeasy mini kit, QIAGEN) was performed for the strain harboring PP7-wt δ = 6. Samples were divided into the following subsamples (Figure S2A):

1. induced/modified (+C_4−HSL/+NAI)
2. noninduced/modified (−C_4−HSL/+NAI)
3. induced/nonmodified (+C_4−HSL/+DMSO)

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

4. noninduced/nonmodified (−C_4−HSL/+DMSO).

Subsequent steps of the SHAPE-seq protocol, that were applied to all samples, have been described elsewhere,[15] including reverse transcription (steps 40−51), adapter ligation and purification (steps 52−57) as well as dsDNA sequencing library preparation (steps 68−76). In brief, 1000 ng of RNA were converted to cDNA using the reverse transcription primers (for details of primer and adapter sequences used in this work see Table S3). The RNA was mixed with 0.5 μM primer for mCherry (#1) 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; 1× 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 17 000*g* and resuspended in 22.5 μL ultrapure water. Next, 1.7 μM of 5′ phosphorylated ssDNA adapter (#2) (see Table S3) was ligated to the cDNA using a CircLigase (Epicenter) reaction mix (1× CircLigase reaction buffer, 2.5 mM $MnCl_2$, 50 μM ATP, 100 Units 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 17 000g) and resuspended in 20 μL ultrapure water. To remove nonligated adapter (#2), resuspended cDNA was further purified using the Agencourt AMPure XP beads (Beackman Coulter) by mixing 1.8× 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 ultrapure water. cDNAs were subjected to PCR amplification to construct dsDNA library as detailed below.

RBP Protection Assay Using *in Vitro* SHAPE-seq. *In vitro* modification was carried out on noninduced, DMSO-treated samples (Figure S3A) and has been described elsewhere.[6] Briefly, 1500 ng of isolated RNA 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 allowing the RNA molecule to refold. Next, we added 15.6 pmol (based on 1:2 molar ratio between RNA:PP7 protein) of highly purified recombinant PP7 protein (Genscript) 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. Samples were then transferred to ice to stop the SHAPE reaction and precipitated by addition of 300 μL ice-cold 100% ethanol, 10 μL Sodium Acetate 3M, 0.5 μL ultrapure glycogen (Thermo scientific) and 70 μL DEPC-treated water. Samples were incubated at −80 °C for 15 min followed by centrifugation at 4 °C, 17 000*g* for 15 min. Supernatant was removed and samples were air-dried for 5 min at room temperature and resuspended in 10 μL of RNase-free water.

SHAPE-Seq 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: 4 nM mCherry selection (#3) (primer extends 4 nucleotides into mCherry transcript to avoid the enrichment of ssDNA-adapter products), 0.5 μM TruSeq Universal Adapter (#4) and 0.5 μM TrueSeq Illumina indexes (one of #5−16) (Table S3) with PCR reaction mix (1× Q5 HotStart reaction buffer, 0.1 mM dNTPs, 1 U Q5 HotStart Polymerase [NEB]). A 15-cycle PCR program was used: initial denaturation at 98 °C for 30 s followed by a denaturation step at 98 °C for 15 s, primer annealing at 65 °C for 30 s and extension at 72 °C for 30 s, 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 1.8× volume of Agencourt AMPure XP beads to the PCR/ExoI mix and purified according to manufacturer's protocol. Samples were eluted in 20 μL ultrapure 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 sequenced using the Illumina HiSeq 2500 sequencing system using 2 × 100 bp paired-end reads.

Analysis Routines and Models. See the Supporting Information.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.8b00378.

Detailed modeling and analysis routines, Figures S1−S4 (PDF)
Table S1 (XLSX)
Table S2 (XLSX)
Table S3 (XLSX)
Table S4 (XLSX)
Table S5 (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*Tel: 972-77-8871894. Fax: 972-4-8293399. E-mail: roeeamit@technion.ac.il.

ORCID

Roee Amit: 0000-0003-0580-7076

Author Contributions

#N.K. and R.C. contributed equally.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to acknowledge the Technion's LS&E staff (Tal Katz-Ezov and Anastasia Diviatis) for help with sequencing of the SHAPE-seq fragments. This project received funding the I-CORE Program of the Planning and Budgeting

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Committee and the Israel Science Foundation (Grant No. 152/11), Marie Curie Reintegration Grant No. PCIG11-GA-2012-321675, and by the European Union's Horizon 2020 Research And Innovation Programme under Grant Agreement No. 664918 - MRG-Grammar.

REFERENCES

(1) Cerretti, D. P., Mattheakis, L. C., Kearney, K. R., Vu, L., and Nomura, M. (1988) Translational regulation of the spc operon in Escherichia coli. *J. Mol. Biol. 204*, 309−325.

(2) Sacerdot, C., Caillet, J., Graffe, M., Eyermann, F., Ehresmann, B., Ehresmann, C., Springer, M., and Romby, P. (1998) 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. 29*, 1077−1090.

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

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

(5) Ding, Y., Kwok, C. K., Tang, Y., Bevilacqua, P. C., and Assmann, S. M. (2015) Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nat. Protoc. 10*, 1050−1066.

(6) Flynn, R. A., Zhang, Q. C., Spitale, R. C., Lee, B., Mumbach, M. R., and Chang, H. Y. (2016) Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. *Nat. Protoc. 11*, 273−290.

(7) Zubradt, M., Gupta, P., Persad, S., Lambowitz, A. M., Weissman, J. S., and Rouskin, S. (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat. Methods 14*, 75−82.

(8) Kinney, J. B., Murugan, A., Callan, C. G., and Cox, E. C. (2010) Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U. S. A. 107*, 9158−9163.

(9) Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012) Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol. 30*, 521−530.

(10) Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S.-I., Cooper, G. M., et al. (2012) Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol. 30*, 265−270.

(11) Levy, L., Anavy, L., Solomon, O., Cohen, R., Brunwasser-Meirom, M., Ohayon, S., Atar, O., Goldberg, S., Yakhini, Z., and Amit, R. (2017) A Synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial σ54Promoters. *Cell Rep. 21*, 845−858.

(12) Weingarten-Gabbay, S., Elias-Kirma, S., Nir, R., Gritsenko, A. A., Stern-Ginossar, N., Yakhini, Z., Weinberger, A., and Segal, E. (2016) Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. *Science 351*, aad4939.

(13) Saito, H., Kobayashi, T., Hara, T., Fujita, Y., Hayashi, K., Furushima, R., and Inoue, T. (2010) Synthetic translational regulation by an L7Ae−kink-turn RNP switch. *Nat. Chem. Biol. 6*, 71−78.

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

(15) Watters, K. E., Abbott, T. R., and Lucks, J. B. (2016) Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucleic Acids Res. 44*, No. e12.

(16) Gott, J. M., Wilhelm, L. J., and Uhlenbeck, O. C. (1991) RNA binding properties of the coat protein from bacteriophage GA. *Nucleic Acids Res. 19*, 6499−6503.

(17) Peabody, D. S. (1993) The RNA binding site of bacteriophage MS2 coat protein. *EMBO J. 12*, 595−600.

(18) Lim, F., and Peabody, D. S. (2002) RNA recognition site of PP7 coat protein. *Nucleic Acids Res. 30*, 4138−4144.

(19) Lim, F., Spingola, M., and Peabody, D. S. (1996) The RNA-Binding Site of Bacteriophage Qβ Coat Protein. *J. Biol. Chem. 271*, 31839−31845.

(20) Zeevi, D., Sharon, E., Lotan-Pompan, M., Lubling, Y., Shipony, Z., Raveh-Sadka, T., Keren, L., Levo, M., Weinberger, A., and Segal, E. (2011) Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. *Genome Res. 21*, 2114−2128.

(21) Keren, L., Zackay, O., Lotan-Pompan, M., Barenholz, U., Dekel, E., Sasson, V., Aidelberg, G., Bren, A., Zeevi, D., Weinberger, A., et al. (2013) Promoters maintain their relative activity levels under different growth conditions. *Mol. Syst. Biol. 9*, 701.

(22) Smola, M. J., Calabrese, J. M., and Weeks, K. M. (2015) Detection of RNA-Protein Interactions in Living Cells with SHAPE. *Biochemistry 54*, 6867−6875.

(23) Choudhary, K., Ruan, L., Deng, F., Shih, N., and Aviran, S. (2016) SEQualyzer: interactive tool for quality control and exploratory analysis of high-throughput RNA structural profiling data. *Bioinformatics 33*, btw627.

(24) Aviran, S., Lucks, J. B., and Pachter, L. (2011) RNA structure characterization from chemical mapping experiments. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp 1743−1750, IEEE.

(25) Zhang, J.-H., and Chung and Oldenburg (1999) A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J. Biomol. Screening 4*, 67−73.

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

(27) Bernardi, A., and Spahr, P.-F. (1972) Nucleotide Sequence at the Binding Site for Coat Protein on RNA of Bacteriophage R17. *Proc. Natl. Acad. Sci. U. S. A. 69*, 3033−3037.

(28) Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem. 125*, 167−188.

(29) Buenrostro, J. D., Araya, C. L., Chircus, L. M., Layton, C. J., Chang, H. Y., Snyder, M. P., and Greenleaf, W. J. (2014) Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol. 32*, 562−568.

(30) Johansson, H. E., Dertinger, D., LeCuyer, K. A., Behlen, L. S., Greef, C. H., and Uhlenbeck, O. C. (1998) A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein. *Proc. Natl. Acad. Sci. U. S. A. 95*, 9244−9249.

(31) Spingola, M., and Peabody, D. S. (1997) MS2 coat protein mutants which bind Qbeta RNA. *Nucleic Acids Res. 25*, 2808−2815.

(32) Witherell, G. W., and Uhlenbeck, O. C. (1989) Specific RNA binding by Q.beta. coat protein. *Biochemistry 28*, 71−76.

(33) Paulus, M., Haslbeck, M., and Watzele, M. (2004) RNA stem−loop enhanced expression of previously non-expressible genes. *Nucleic Acids Res. 32*, No. e78.

(34) Espah Borujeni, A., Cetnar, D., Farasat, I., Smith, A., Lundgren, N., and Salis, H. M. (2017) Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Res. 45*, 5437−5448.

(35) Watters, K. E., Yu, A. M., Strobel, E. J., Settle, A. H., and Lucks, J. B. (2016) Characterizing RNA structures in vitro and in vivo with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Methods 103*, 34−48.

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(36) Weissman, J., Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature 505*, 701.
(37) Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., C Bevilacqua, P., and M Assmann, S. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature 505*, 696.
(38) Medina, G., Juárez, K., Valderrama, B., and Soberón-Chávez, G. (2003) Mechanism of Pseudomonas aeruginosa RhlR Transcriptional Regulation of the rhlAB Promoter. *J. Bacteriol. 185*, 5976−5983.
(39) Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol. 215*, 403−410.

DOI: 10.1021/acssynbio.8b00378
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Video Article

An Assay for Quantifying Protein-RNA Binding in Bacteria

Noa Katz*[1], Roni Cohen*[1], Orna Atar[1], Sarah Goldberg[1], Roee Amit[1,2]

[1]Department of Biotechnology and Food Engineering, Technion-Israel Institute of Technology

[2]Russell Berrie Nanotechnology Institute, Technion-Israel Institute of Technology

*These authors contributed equally

Correspondence to: Noa Katz at katznoa@gmail.com

URL: https://www.jove.com/video/59611
DOI: doi:10.3791/59611

Keywords: Genetics, Issue 148, RNA binding protein (RBP), MS2, PP7, phage coat protein, binding assay, post-transcriptional regulation, translation repression, synthetic circuit, RBP binding affinity, RNA circuit, reporter gene, RBP interaction

Date Published: 6/12/2019

Citation: Katz, N., Cohen, R., Atar, O., Goldberg, S., Amit, R. An Assay for Quantifying Protein-RNA Binding in Bacteria. *J. Vis. Exp.* (148), e59611, doi:10.3791/59611 (2019).

Abstract

In the initiation step of protein translation, the ribosome binds to the initiation region of the mRNA. Translation initiation can be blocked by binding of an RNA binding protein (RBP) to the initiation region of the mRNA, which interferes with ribosome binding. In the presented method, we utilize this blocking phenomenon to quantify the binding affinity of RBPs to their cognate and non-cognate binding sites. To do this, we insert a test binding site in the initiation region of a reporter mRNA and induce the expression of the test RBP. In the case of RBP-RNA binding, we observed a sigmoidal repression of the reporter expression as a function of RBP concentration. In the case of no-affinity or very low affinity between binding site and RBP, no significant repression was observed. The method is carried out in live bacterial cells, and does not require expensive or sophisticated machinery. It is useful for quantifying and comparing between the binding affinities of different RBPs that are functional in bacteria to a set of designed binding sites. This method may be inappropriate for binding sites with high structural complexity. This is due to the possibility of repression of ribosomal initiation by complex mRNA structure in the absence of RBP, which would result in lower basal reporter gene expression, and thus less-observable reporter repression upon RBP binding.

Video Link

The video component of this article can be found at https://www.jove.com/video/59611/

Introduction

RNA binding protein (RBP)-based post-transcriptional regulation, specifically characterization of the interaction between RBPs and RNA, has been studied extensively in recent decades. There are multiple examples of translational down-regulation in bacteria originating from RBPs inhibiting, or directly competing with, ribosome binding[1,2,3]. In the field of synthetic biology, RBP-RNA interactions are emerging as a significant tool for the design of transcription-based genetic circuits[4,5]. Therefore, there is an increase in demand for characterization of such RBP-RNA interactions in a cellular context.

The most common methods for studying protein-RNA interactions are the electrophoretic mobility shift assay (EMSA)[6], which is limited to in vitro settings, and various pull-down assays[7], including the CLIP method[8,9]. While such methods enable the discovery of de novo RNA binding sites, they suffer from drawbacks such as labor-intensive protocols and expensive deep sequencing reactions and may require a specific antibody for RBP pull-down. Due to the susceptible nature of RNA to its environment, many factors can affect RBP-RNA interactions, emphasizing the importance of interrogating RBP-RNA binding in the cellular context. For example, we and others have demonstrated significant differences between RNA structures in vivo and in vitro[10,11].

Based on the approach of a previous study[12], we recently demonstrated[10] that when placing pre-designed binding sites for the capsid RBPs from the bacteriophages GA[13], MS2[14], PP7[15], and Qβ[16] in the translation initiation region of a reporter mRNA, reporter expression is strongly repressed. We present a relatively simple and quantitative method, based on this repression phenomenon, to measure the affinity between RBPs and their corresponding RNA binding sites in vivo.

Protocol

1. System Preparation

1. **Design of binding-site plasmids**
    1. Design the binding site cassette as depicted in **Figure 1**. Each minigene contains the following parts (5' to 3'): EagI restriction site, #40 bases of the 5' end of the kanamycin (Kan) resistance gene, pLac-Ara promoter, ribosome binding site (RBS), AUG of the mCherry gene, a spacer (δ), an RBP binding site, 80 bases of the 5' end of the mCherry gene, and an ApaLI restriction site.

NOTE: To increase the success rate of the assay, design three binding-site cassettes for each binding site, with spacers consisting of at least one, two, and three bases. See Representative Results section for further guidelines.

2. **Cloning of binding site plasmids**
    1. Order the binding-site cassettes as double-stranded DNA (dsDNA) minigenes. Each minigene is #500 bp long and contains an EagI restriction site and an ApaLI restriction site at the 5' and 3' ends, respectively (see step 1.1.1).
    NOTE: In this experiment, mini-genes with half of the kanamycin gene were ordered to facilitate screening for positive colonies. However, Gibson assembly[17] is also suitable here, in which case the binding site can be ordered as two shorter complementary single-stranded DNA oligos.
    2. Double-digest both the mini-genes and the target vector with EagI-HF and ApaLI by the restriction protocol[18], and column purify[19].
    3. Ligate the digested minigenes to the binding-site backbone containing the rest of the mCherry reporter gene, terminator, and a kanamycin resistance gene[20].
    4. Transform the ligation solution into *Escherichia coli* TOP10 cells[21].
    5. Identify positive transformants via Sanger sequencing.
        1. Design a primer 100 bases upstream to the region of interest (see **Table 1** for primer sequences).
        2. Miniprep a few bacterial colonies[22].
        3. Prepare 5 µL of a 5 mM solution of the primer and 10 µL of the DNA at 80 ng/µL concentration.
        4. Send the two solution to a convenient facility for Sanger sequencing[23].
    6. Store purified plasmids at -20 °C, and bacterial strains as glycerol stocks[24], both in the 96-well format. DNA will then be used for transformation into *E. coli* TOP10 cells containing one of four fusion-RBP plasmids (see step 1.3.5).

3. **Design and construction of the RBP plasmid**
    NOTE: Amino acid and nucleotide sequences of the coat proteins used in this study are listed in **Table 2**.
    1. Order the required RBP sequence lacking a stop codon as a custom-ordered dsDNA minigene lacking a stop codon with restriction sites at the ends (**Figure 1**).
    2. Clone the tested RBP lacking a stop codon immediately downstream of an inducible promoter and upstream of a fluorescent protein lacking a start codon (**Figure 1**), similar to steps 1.2.2-1.2.4. Make sure that the RBP plasmid contains a different antibiotic resistance gene than the binding-site plasmid.
    3. Identify positive transformants via Sanger sequencing, similar to step 1.2.5 (see **Table 1** for primer sequences).
    4. Choose one positive transformant and make it chemically-competent[25]. Store as glycerol purified plasmids at -20 °C and glycerol stocks of bacterial strains[24] at -80 °C in 96-well plates.
    5. Transform the binding-site plasmids (from step 1.2.6) stored in 96-well plates into chemically-competent bacterial cells already containing an RBP-mCerulean plasmid[21]. To save time, instead of plating the cells on Petri dishes, plate them using an 8-channel pipettor on 8-lane plates containing Luria-Bertani (LB)[26] agar with relevant antibiotics (Kan and Amp). Colonies should appear in 16 h.
    6. Select a single colony for each double transformant and grow overnight in LB medium with the relevant antibiotics (Kan and Amp) and store as glycerol stocks[24] at -80 °C in 96-well plates.

2. Experiment Setup

NOTE: The protocol presented here was performed using a liquid-handling robotic system in combination with an incubator and a plate reader. Each measurement was carried out for 24 inducer concentrations, with two duplicates for each strain + inducer combination. Using this robotic system, data for 16 strains per day with 24 inducer concentrations was collected. However, if such a device is unavailable, or if fewer experiments are necessary, these can easily be done by hand using an 8-channel multi-pipette and adapting the protocol accordingly. For example, preliminary results for four strains per day with 12 inducer concentrations and four time-points were acquired in this manner.

1. Prepare, in advance, 1 L of bioassay buffer (BA) by mixing 0.5 g of tryptone, 0.3 mL of glycerol, 5.8 g of NaCl, 50 mL of 1 M $MgSO_4$, 1 mL of 10x phosphate-buffered saline (PBS) buffer pH 7.4, and 950 mL of double distilled water (DDW). Autoclave or sterile filter the BA buffer.
2. Grow the double-transformant strains at 37 °C and 250 rpm shaking in 1.5 mL LB with appropriate antibiotics (kanamycin at a final concentration of 25 µg/mL and ampicillin at a final concentration of 100 µg/mL), in 48-well plates, over a period of 18 h (overnight).
3. **In the morning, make the following preparations.**
    1. Inducer plate. In a clean 96-well plate, prepare wells with semi-poor medium (SPM) consisting of 95% BA and 5% LB[26] in the incubator at 37 °C. The number of wells corresponds to the desired number of inducer concentrations. Add C4-HSL to the wells in the inducer plate that will contain the highest inducer concentration (218 nM).
    2. Program the robot to serially dilute medium from each of the highest-concentration wells into 23 lower concentrations ranging from 0 to 218 nM. The volume of each inducer dilution should be sufficient for all strains (including duplicates).
    3. While the inducer dilutions are being prepared, warm 180 µL of SPM in the incubator at 37 °C, in 96-well plates.
    4. Dilute the overnight strains from step 2.2 by a factor of 100 by serial dilutions: first dilute by a factor of 10 by mixing 100 µL of bacteria with 900 µL of SPM in 48-well plates, and then dilute again by a factor of 10 by taking 20 µL from the diluted solution into 180 µL of pre-warmed SPM, in 96-well plates suitable for fluorescent measurements.
    5. Add the diluted inducer from the inducer plate to the 96-well plates with the diluted strains according to the final concentrations.
4. Shake the 96-well plates at 37 °C for 6 h, while taking measurements of optical density at 595 nm (OD_{595}), mCherry (560 nm/612 nm) and mCerulean (460 nm/510 nm) fluorescence via a plate reader every 30 min. For normalization purposes, measure growth of SMP with no cells added.

3. Preliminary Results Analysis

1. For each day of experiment, choose a time interval of logarithmic growth according to the measured growth curves, between the linear growth phase and the stationary (T_0, T_{final}). Take approximately 6−8 time points, while discarding the first and last measurements to avoid error derived from inaccuracy of exponential growth detection (see **Figure 2A**, top panel).
   NOTE: Discard strains that show abnormal growth curves or strains where logarithmic growth phase could not be detected and repeat the experiment.
2. **Calculate the average normalized fluorescence of mCerulean and rate of production of mCherry, from the raw data of both mCerulean and mCherry fluorescence for each inducer concentration (Figure 2A).**
   1. Calculate normalized mCerulean as follows:
   $$Eq.\,1: Normalized\ mCerulean = \frac{mCerulean - blank(mCerulean)}{OD - blank(OD)}$$
   where blank(mCerulean) is the mCerulean level [a.u.] for medium only, blank(OD) is the optical density for medium only, and mCerulean and OD are the mCerulean fluorescence and optical density values, respectively.
   2. Average mCerulean over the different time points (**Figure 2B**, top two panels) as follows:
   $$Eq.\,2: Averaged\ mCerulean = \frac{\sum_{T\_0}^{T\_final} Normalized\_mCerulean}{\#\ Time\ points}$$
   where #Time points is the number of data timepoints taken into account, T_0 is the time at which the exponential growth phase begins, and T_{final} is the time at which the exponential growth phase ends.
   3. Calculate mCherry rate of production (**Figure 2B**, bottom two panels) as follows:
   $$Eq.\,3: mCherry\ production\ rate = \frac{mCherry(Tfinal) - mCherry(T0)}{\int_{T\_0}^{T\_final} ODdt}$$
   where mCherry(t) is the mCherry level [a.u.] at time t, OD is the optical density value, T_0 is the time at which the exponential growth phase begins, and T_{final} is the time at which the exponential growth phase ends.
3. Finally, plot the mCherry rate of production as a function of mCerulean, creating dose response curves as a function of RBP-mCerulean fusion fluorescence (**Figure 2C**). Such plots represent production of the reporter gene as a function of RBP presence in the cell.

4. Dose Response Function Fitting Routine and K_{RBP} Extraction

1. Under the assumption that the ribosome rate of translation with the RBP bound is constant, model the mCherry production rate as follows (see **Figure 2D**, green line):
   $$Eq.\,4: mCherry\ production\ rate = \frac{k_{unbound}}{1 + \left(\frac{[x]}{k_{RBP}}\right)^n} + C$$
   where [x] is the normalized average mCerulean fluorescence calculated according to Eq. 2, mCherry production rate is the value calculated according to Eq. 3, K_{RBP} is the relative binding affinity [a.u.], $K_{unbound}$ is the ribosome rate of translation with the RBP unbound, n is the cooperativity factor, and C is the base fluorescence [a.u.]. C, n, $K_{unbound}$, and K_{RBP} are found by fitting the mCherry production rate data to the model (Eq. 4).
2. Using data analysis software, conduct a fitting procedure on plots depicting mCherry production rate as a function of averaged mCerulean (step 3.3), and extract the fit parameters according to the formula in Eq. 4.
   NOTE: Only fitting results with $R^2 > 0.6$ are taken into account. For those fits, K_{RBP} error is mostly in the range of 0.5% to 20% of K_{RBP} values, for a 0.67 confidence interval, while those with higher K_{RBP} error can be also verified by eye.
3. Normalize K_{RBP} values by the respective maximal value of averaged mCerulean for each dose-response function.
   $$Eq.\,5: normalized\_k_{RBP} = \frac{k_{RBP}}{\max(averaged\ mCerulean)}$$
   where K_{RBP} in [a.u.] is the value extracted from the fitting procedure in Eq. 4, and max (averaged mCerulean) is the maximal averaged mCerulean signal [a.u] observed for the current strain.
   NOTE: The normalization facilitates correct comparison of the regulatory effect across strains by eliminating the dependence on the particular maximal RBP expression levels.

Representative Results

The presented method utilizes the competition between an RBP and the ribosome for binding to the mRNA molecule (**Figure 1**). This competition is reflected by decreasing mCherry levels as a function of increased production of RBP-mCerulean, due to increasing concentrations of inducer. In the case of increasing mCerulean fluorescence, with no significant changes in mCherry, a lack of RBP binding is deduced. Representative results for both a positive and a negative strain are depicted in **Figure 2**. In **Figure 2A**, the OD, mCherry, and mCerulean channels are presented as a function of time and inducer over a range of four hours, with T_0 = 1 h and T_{final} = 3.5 h. In **Figure 2B**, averaged mCerulean fluorescence (top) and mCherry rate of production (bottom) are presented as a function of inducer concentration, for the two example strains. As can be seen, the results for a positive strain display a clear down-regulatory effect in the mCherry rate of production (**Figure 2B,C**), which translates into a significant non-zero value of K_{RBP} (**Figure 2D**). For the positive strain, the fitting procedure yielded the following values: K_{RBP} = 394.6 a.u., $K_{unbound}$ = 275.6, n = 2.1, C = 11.2 a.u., and R^2 = 0.93. After normalization by the maximal mCerulean fluorescence, the K_{RBP} value was 0.24. For the negative strain, a lack of distinct response was observed (**Figure 2C**), and no K_{RBP} value was extracted (**Figure 2D**).

In **Figure 3**, we present the results of this assay for two phage coat RBPs, PP7 and MS2, on several mutated binding sites, at different locations within the initiation region of the mCherry mRNA. The results are roughly classified into three kinds of responses (**Figure 3A**): strains exhibiting a down-regulatory effect at a low mCerulean level, reflecting a low K_{RBP} value (high binding affinity); strains exhibiting down-regulatory effect at either intermediate or high mCerulean levels, reflecting a high K_{RBP} value (intermediate or low affinity); and strains exhibiting no distinct response to rising levels of mCerulean, reflecting a higher K_{RBP} value than the maximum RBP concentration in the cell (no detectible binding affinity). **Figure 3B** presents the minimal K_{RBP} value computed for every RBP−binding-site combination based on all combinations of the two RBPs and ten binding-sites at different positions. The binding sites include a negative control (no binding site), non-matching binding sites, and a positive control — the native binding site for each RBP (PP7-wt for PP7 coat protein [PCP], and MS2-wt for MS2 coat protein [MCP]). The results match the predictions, as both RBPs present a high affinity for their positive controls, and a non-detectible binding affinity for the negative controls. Additionally, previous studies using these two RBPs[27,28] have observed that they are orthogonal, which is clearly conveyed in the heatmap presented: both MCP and PCP do not bind the native site of the other RBP. Furthermore, the mutated binding sites present varying results, where some binding sites displayed a similar level of affinity as that of the native site, such as PP7-mut-1, PP7-mut-2, and MS2-mut-3, while others displayed a significantly lower affinity, such as PP7-mut-3 and MS2-mut-2. Thus, the assay presented a quantitative in vivo measurement of the binding affinity of RBPs, yielding results that are comparable to those of past experiments with these RBPs.

Since the assay is based on repression of the mCherry gene, a viable mCherry signal is required. Therefore, when designing the binding site cassette, there are two design rules to keep in mind. First, the open reading frame (ORF) of the mCherry should be kept. Since the binding-site length can vary, inserting it into the gene can cause a shift of one or two bases from the original mCherry ORF. Therefore, if needed (**Figure 4A**), insert one or two bases immediately downstream to the binding site. For example, a binding site that is 20-base long, with a δ of two bases, will yield an addition of 22 bases to the mCherry gene. To keep the ORF, we need to add two bases, for a total of 24 bases. The second design rule is to avoid insertions of stop codons into the mCherry ORF. Some binding sites, as the MS2-mut-2 (**Figure 4B**, inset), contain stop codons when positioned in one or more of the three possible ORFs. Such an example is illustrated in **Figure 4A**, where the binding site contained a stop codon that is in-frame with the mCherry ORF only when no bases are added. As can be seen in the dose-response curve for that position (**Figure 4B**), mCherry production rate was undetectable, thus the binding affinity could not be measured.

A closer look at **Figure 4B** demonstrates the effect of the spacing δ on mCherry production. For instance, for δ = 4, basal production rate was a factor of six more than those for δ = 5, ensuring a higher fold-repression effect. For δ = 14, however, the basal production levels were too low to observe a down-regulatory effect.

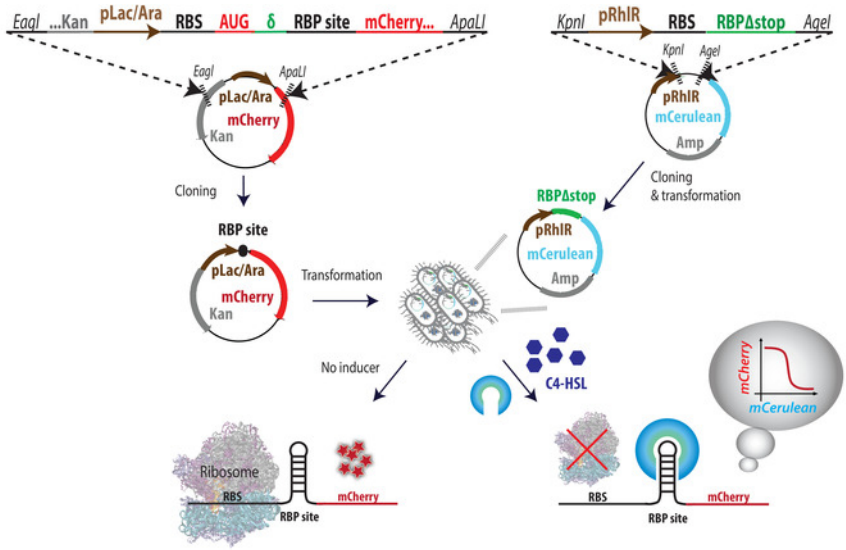


Figure 1: Overview of system design and cloning steps. Illustration of the cassette design for the binding site plasmid (left) and RBP-mCerulean plasmid (right). The next step is consecutive transformations of both plasmids into competent *E. coli* cells, with RBP plasmids first. Double-transformants are then tested for their mCherry expression levels in increasing inducer concentrations; if the RBP binds to the binding site, mCherry levels decline as a function of mCerulean (gray bubble). Please click here to view a larger version of this figure.

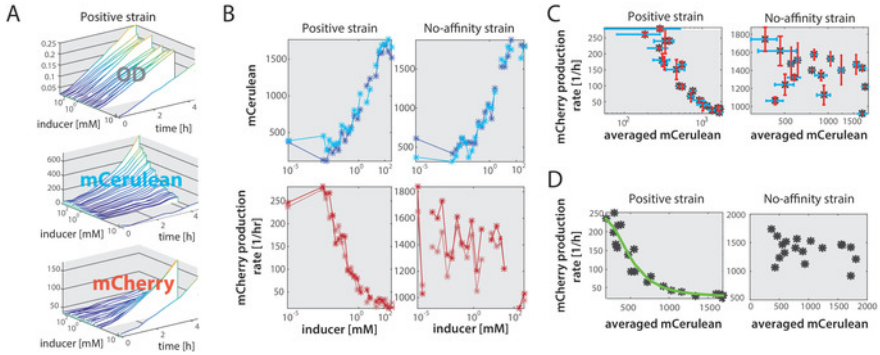


Figure 2: Analysis scheme. (**A**) Three-dimensional (3D) plots depicting raw OD levels (top), mCerulean fluorescence (middle), and mCherry fluorescence (bottom) as a function of time and inducer concentration, for a positive strain. (**B**) Top: mCerulean steady-state expression levels for each inducer concentration is computed by dividing each fluorescence level by the respective OD and averaging over all values in the 2−3 h exponential growth time window for both the positive (left) and negative (right) strains. Bottom: mCherry production rate computed according to Eq. 3 for time-points 2−3 h after induction. (**C**) mCherry production rate plotted as a function of mean mCerulean fluorescence averaged over two biological duplicates for two strains. Error bars are standard deviation of both mCherry production rate and averaged mCerulean fluorescence acquired from at least two replicates. (**D**) Fit for K_{RBP} using the fitting formula in Eq. 4 shown for the positive strain (left), exhibiting a specific binding response. For the negative strain (right), no K_{RBP} value was extracted. Data is shown in duplicate. This figure has been adapted with permission from Katz et al.[10]. Copyright 2018 American Chemical Society. Please click here to view a larger version of this figure.

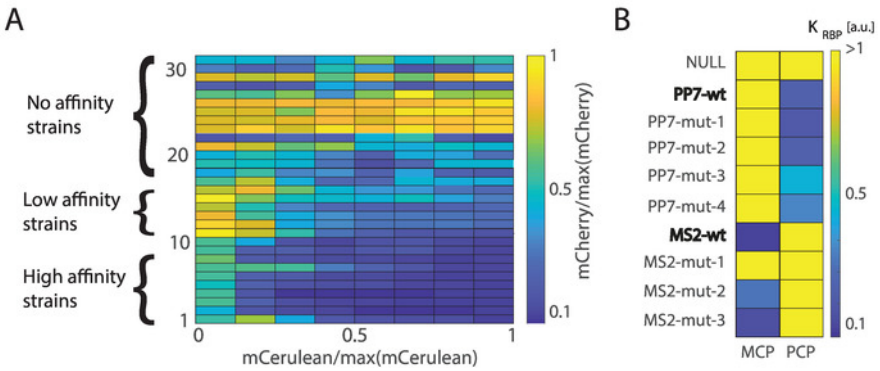


Figure 3: Representative **final results.** (**A**) Normalized dose-response curves for thirty different strains based on two RBPs and ten binding sites at different locations. Three types of responses are observed: high affinity, low affinity, and no affinity. (**B**) Quantitative K_{RBP} results for two RBPs (MCP and PCP) with five different binding site cassettes (listed). All RBP−binding-site strains were measured in duplicate. This figure has been adapted with permission from Katz et al.[10]. Copyright 2018 American Chemical Society. Please click here to view a larger version of this figure.

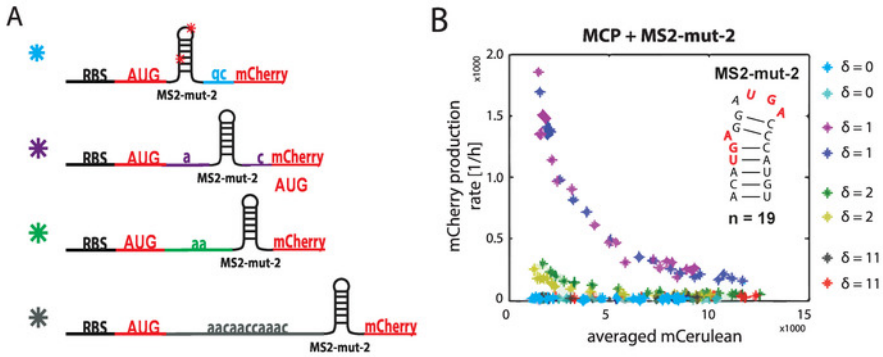


Figure 4: Example design and results for MCP with a mutant binding site. (**A**) Design illustration of the binding site cassettes in four different locations. Cassette including the ribosome binding site, start codon for the mCherry, δ spacer bases, the binding site tested, one or two bases to maintain the ORF, and the rest of the mCherry gene. Red stars indicate a stop codon. (**B**) Dose-response curves for MCP with a mutant binding site at four different locations. Inset: the sequence of the tested mutated binding site. Results presented are for duplicates of each strain. Please click here to view a larger version of this figure.

Name	Binidng site location, A in AUG = 1	Binding site sequence (RBS for controls)	Site: ATG to second mCherry codon GTG Controls: RBS to second mCherry codon GTG	Source
MS2_wt_d5	5	acatgaggattacccatgt	atgcacatgaggattacccatgtcgtg	Gen9 Inc.
MS2_wt_d6	6	acatgaggattacccatgt	atggcacatgaggattacccatgtgtg	Gen9 Inc.
MS2_wt_d8	8	acatgaggattacccatgt	atggcgcacatgaggattacccatgtcgtg	Gen9 Inc.
MS2_wt_d9	9	acatgaggattacccatgt	atggcgccacatgaggattacccatgtgtg	Gen9 Inc.
MS2_U(-5)C_d8	8	acatgaggatcacccatgt	atgcacatgaggatcacccatgtggtg	Gen9 Inc.
MS2_U(-5)C_d9	9	acatgaggatcacccatgt	atggcacatgaggatcacccatgtgtg	Gen9 Inc.
MS2_U(-5)C_d8	8	acatgaggatgacccatgt	atgcacatgaggatgacccatgtggtg	Gen9 Inc.
MS2_U(-5)G_d9	9	acatgaggatgacccatgt	atggcacatgaggatgacccatgtgtg	Gen9 Inc.
MS2_struct_d9	9	cacaagaggttcacttatg	atggccacaagaggttcacttatggtg	Gen9 Inc.
MS2_struct_d8	8	cacaagaggttcacttatg	atgccacaagaggttcacttatgggtg	Gen9 Inc.
PP7wt_d5'	5	taaggagtttatatggaaaccctta	atgctaaggagtttatatggaaaccccttacgtg	Gen9 Inc.
PP7wt_d6'	6	taaggagtttatatggaaaccctta	atgaataaggagtttatatggaaaccccttagtg	Twist Bioscience
PP7wt_d8'	8	taaggagtttatatggaaaccctta	atgaacataaggagtttatatggaaacccttacgtg	Twist Bioscience
PP7wt_d9'	9	taaggagtttatatggaaaccctta	atgaacaataaggagtttatatggaaacccttagtg	Twist Bioscience
PP7_USLSBm_d6	6	taaccgctttatatggaaagggtta	atggctaaccgctttatatggaaaggttagtg	Gen9 Inc.
PP7_USLSBm_d15	15	taaccgctttatatggaaagggtta	atgggcgccggcgctaaccgcttttatggaaagggttagtg	Gen9 Inc.
PP7_nB_d5	5	taagggtttatatggaaaccctta	atgctaagggtttatatggaaacccttagcgtg	Gen9 Inc.
PP7_nB_d6	6	taagggtttatatggaaaccctta	atggctaagggtttatatggaaacccttatgtg	Gen9 Inc.
PP7_USs_d5	5	taaggagttatatggaaccctta	atgctaaggagttatatggaaccctagtg	Gen9 Inc.
PP7_USs_d6	6	taaggagttatatggaaccctta	atggctaaggagttatatggaacccttagcgtg	Gen9 Inc.
No_BS_d1	-	-	ttaaagaggagaaaggtacccatggtg	Gen9 Inc.
No_BS_d4	-	-	ttaaagaggagaaaggtacccatggcgtg	Gen9 Inc.
No_BS_d10	-	-	ttaaagaggagaaaggtacccatggcgccggcgtg	Gen9 Inc.
Sequencing primer for binding site cassettes			gcatttttatccataagattagcgg	IDT
Sequencing primer for RBP cassettes			gcggcgctgggtctcatctaataa	IDT

Table 1: Binding sites and sequencing primers. Sequences for the binding sites and binding site cassettes used in this study, as well as the primers for the sequencing reactions detailed in the protocol (steps 1.2.5.1 and 1.3.3).

RBP name in this work	source organism name, protein	source organism gene	source organism refseq	wt aa seq	changes from wt (and references)	aa seq used in this work	nt seq used in this work
MCP	Escherichia virus MS2	cp	NC_001417.2	MASNFTQFVLV DNGGTGDVTV APSNFANGVA EWISSNSRSQ AYKVTCSVRQ SSAQNRKYTI KVEVPKVATQT VGGVELPVA AWRSYLNMEL TIPIFATNSD CELIVKAMQG LLKDGNPIPS AIAANSGIY	delF-G [1] V29I [1] taken from addgene plasmid 27121	MASNFTQFVLV DNGGTGDVTV APSNFANGIA EWISSNSRSQ AYKVTCSVRQ SSAQNRKYTI KVEVPKG AWRSYLNMEL TIPIFATNSD CELIVKAMQG LLKDGNPIPS AIAANSGIY	ATGGCTTCTA ACTTTACTCA GTTCGTTCTC GTCGACAATG GCGGAACTGG CGACGTGACT GTCGCCCCAA GCAACTTCGC TAACGGGATC GCTGAATGGA TCAGCTCTAA CTCGCGTTCA CAGGCTTACA AAGTAACCTG TAGCGTTCGT CAGAGCTCTG CGCAGAATCG CAAATACACC ATCAAAGTCG AGGTGCCTAA AGGCGCCTGG CGTTCGTACT TAAATATGGA ACTAACCATT CCAATTTTCG CCACGAATTC CGACTGCGAG CTTATTGTTA AGGCAATGCA AGGTCTCCTA AAAGATGGAA ACCCGATTCC CTCAGCAATC GCAGCAAACT CCGGCATCTAC
PCP	Pseudomonas phage PP7	cp	NC_001628.1	MSKTIVLSVGEA TRTLTEIQST ADRQIFEEKV GPLVGRLRLT ASLRQNGAKT AYRVNLKLDQ ADVVDCSTSVC GELPKVRYTQ VWSHDVTIVA NSTEASRKSL YDLTKSLVAT SQVEDLVVNL VPLGR	delF-G [2] taken from addgene plasmid 40650	MLASKTIVLSVG EATRTLTEIQ STADRQIFEE KVGPLVGRLR LTASLRQNGA KTAYRVNLKL DQADVVDSG LPKVRYTQVW SHDVTIVANS TEASRKSLYD LTKSLVATSQ VEDLVVNLVP LGR	ATGCTAGCCTC CAAAACCATC GTTCTTTCGG TCGGCGAGGC TACTCGCACT CTGACTGAGA TCCAGTCCAC CGCAGACCGT CAGATCTTCG AAGAGAAGGT CGGGCCTCTG GTGGGTCGGC TGCGCCTCAC GGCTTCGCTC CGTCAAAACG GAGCCAAGAC CGCGTATCGC GTCAACCTAA AACTGGATCA GGCGGACGTC GTTGATTCCG GACTTCCGAA AGTGCGCTAC ACTCAGGTAT GGTCGCACGA CGTGACAATC GTTGCGAATA GCACCGAGGC CTCGCGCAAA TCGTTGTACG ATTTGACCAA

							GTCCCTCGTC GCGACCTCGC AGGTCGAAGA TCTTGTCGTC AACCTTGTGC CGCTGGGCCGT
References:							
1.Peabody, D.S., Ely, K.R. Control of translational repression by protein-protein interactions. *Nucleic Acids Research*. **20** (7), 1649–1655 (1992).							
2. Chao, J.A., Patskovsky, Y., Almo, S.C., Singer, R.H. Structural basis for the coevolution of a viral RNA–protein complex. *Nature Structural & Molecular Biology*. **15** (1), 103–105, doi: 10.1038/nsmb1327 (2008)							

Table 2: RBP sequences. Amino acid and nucleotide sequences of the coat proteins used in this study.

Discussion

The method described in this article facilitates quantitative in vivo measurement of RBP-RNA binding affinity in *E. coli* cells. The protocol is relatively easy and can be conducted without the use of sophisticated machinery, and data analysis is straightforward. Moreover, the results are produced immediately, without the relatively long wait-time associated with next generation sequencing (NGS) results.

One limitation to this method is that it works only in bacterial cells. However, a previous study[12] has demonstrated a repression effect using a similar approach for the L7AE RBP in mammalian cells. An additional limitation of the method is that the insertion of the binding site in the mCherry initiation region may repress basal mCherry levels. Structural complexity or high stability of the binding site can interfere with ribosomal initiation even in the absence of RBP, resulting in decreased mCherry basal levels. If basal levels are too low, the additional repression brought on by increasing concentrations of RBP will not be observable. In such a case, it is best to design the binding site cassette with the binding site still in the initiation region, but on the verge of the transition from initiation region to elongation region (δ in the range of 12−15 bp[10,29]). We have shown that for such δ values a repression effect can still be observed. To increase the chances that the assay will work, regardless of structural complexity, we advise performing the assay on at least three different positions for a given binding site.

The main disadvantage of the method in comparison to in vitro methods, such as EMSA, is that the RBP-RNA binding affinity is not measured in absolute units of RBP concentration, but rather in terms of fusion-RBP fluorescence. This disadvantage is a direct result of the in vivo setting, which limits our ability to read out the actual concentrations of RBP. This disadvantage is offset by the benefits of measuring in the in vivo setting. For example, we have found differences in binding affinities when comparing results from our in vivo assay to previous in vitro and in situ assays. These differences may stem from discrepancies in the structure of the mRNA molecules in vivo that emerge from their presence inside cells[10,11,30,31]. Such structural differences may lead to changes in the stability of the folded states in vivo which, in turn, either stabilize or de-stabilize RBP binding.

Since the method is relatively simple and inexpensive, we advise running multiple controls alongside the actual experiment. Running a negative control, i.e., a sequence that has no affinity to the RBP yet has similar structural features, can help avoid false positives stemming from non-specific interactions with the mRNA. In the representative results shown, the two negative controls were the mCherry gene alone (no binding site), and the native binding site of the other RBP (i.e., PP7-wt for MCP and MS2-wt for PCP). Moreover, we propose incorporating a positive control (such as an RBP and its native binding site). Such a control will help in quantifying the binding affinity by presenting a reference point, and in avoiding false-negatives stemming from low fold-repression.

Finally, for those who wish to obtain a structural perspective of RBP-RNA binding, we propose carrying out a selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)[11,32,33] experiment. SHAPE-Seq is an NGS approach combined with chemical probing of RNA, which can be used to estimate secondary structure of RNA as well as RNA interactions with other molecules, such as proteins. In our previous work we conducted a SHAPE-Seq experiment on a representative strain in both in vivo conditions[34] and in vitro with purified recombinant protein[10,35]. In our case, the results revealed that RBP-binding affected a much wider segment of RNA than previously reported for these RBPs in vitro[36].

Disclosures

The authors have nothing to disclose.

Acknowledgments

This project received funding from the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation (Grant No. 152/11), Marie Curie Reintegration Grant No. PCIG11-GA- 2012-321675, and from the European Union's Horizon 2020 Research and Innovation Program under grant agreement no. 664918 - MRG-Grammar.

References

1. Cerretti, D.P., Mattheakis, L.C., Kearney, K.R., 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. *Journal of Molecular Biology.* **204** (2), 309-329 (1988).
2. Babitzke, P., Baker, C.S., Romeo, T. Regulation of translation initiation by RNA binding proteins. *Annual Review of Microbiology.* **63**, 27-44 (2009).

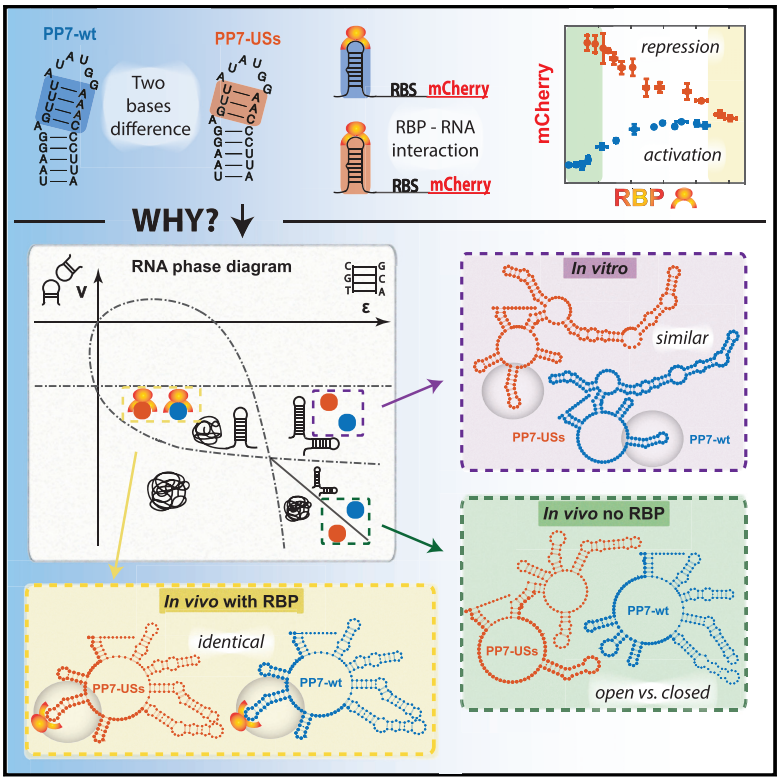
3. Van Assche, E., Van Puyvelde, S., Vanderleyden, J., Steenackers, H.P. RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Frontiers in Microbiology.* **6**, 141 (2015).
4. Chappell, J., Watters, K.E., Takahashi, M.K., Lucks, J.B. A renaissance in RNA synthetic biology: new mechanisms, applications and tools for the future. *Current Opinion in Chemical Biology.* **28**, 47-56 (2015).
5. Wagner, T.E. et al. Small-molecule-based regulation of RNA-delivered circuits in mammalian cells. *Nature Chemical Biology.* **14** (11), 1043 (2018).
6. Bendak, K. et al. A rapid method for assessing the RNA-binding potential of a protein. *Nucleic Acids Research.* **40** (14), e105 (2012).
7. Strein, C., Alleaume, A.-M., Rothbauer, U., Hentze, M.W., Castello, A. A versatile assay for RNA-binding proteins in living cells. *RNA.* **20** (5), 721-731 (2014).
8. Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., Darnell, R.B. CLIP identifies Nova-regulated RNA networks in the brain. *Science.* **302** (5648), 1212-1215 (2003).
9. Lee, F.C.Y., Ule, J. Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular Cell.* **69** (3), 354-369 (2018).
10. Katz, N. et al. An in Vivo Binding Assay for RNA-Binding Proteins Based on Repression of a Reporter Gene. *ACS Synthetic Biology.* **7** (12), 2765-2774 (2018).
11. Watters, K.E., Yu, A.M., Strobel, E.J., Settle, A.H., Lucks, J.B. Characterizing RNA structures in vitro and in vivo with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Methods.* **103**, 34-48 (2016).
12. Saito, H. et al. Synthetic translational regulation by an L7Ae-kink-turn RNP switch. *Nature Chemical Biology.* **6** (1), 71-78 (2010).
13. Gott, J.M., Wilhelm, L.J., Uhlenbeck, O.C. RNA binding properties of the coat protein from bacteriophage GA. *Nucleic Acids Research.* **19** (23), 6499-6503 (1991).
14. Peabody, D.S. The RNA binding site of bacteriophage MS2 coat protein. *The EMBO Journal.* **12** (2), 595-600 (1993).
15. Lim, F., Peabody, D.S. RNA recognition site of PP7 coat protein. *Nucleic Acids Research.* **30** (19), 4138-4144 (2002).
16. Lim, F., Spingola, M., Peabody, D.S. The RNA-binding Site of Bacteriophage Qβ Coat Protein. *Journal of Biological Chemistry.* **271** (50), 31839-31845 (1996).
17. Gibson, D.G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods.* **6** (5), 343-345 (2009).
18. *Optimizing Restriction Endonuclease Reactions.* NEB. https://international.neb.com/tools-and-resources/usage-guidelines/optimizing-restriction-endonuclease-reactions. (2018).
19. *Wizard® SV Gel and PCR Clean-Up System Protocol.* https://worldwide.promega.com/resources/protocols/technical-bulletins/101/wizard-sv-gel-and-pcr-cleanup-system-protocol/. (2018).
20. *Ligation Protocol with T4 DNA Ligase (M0202).* NEB. https://international.neb.com/protocols/0001/01/01/dna-ligation-with-t4-dna-ligase-m0202. (2018).
21. *Routine Cloning Using Top10 Competent Cells - US.* https://www.thermofisher.com/us/en/home/references/protocols/cloning/competent-cells-protocol/routine-cloning-using-top10-competent-cells.html. (2018).
22. *NucleoSpin Plasmid - plasmid Miniprep kit.* https://www.mn-net.com/ProductsBioanalysis/DNAandRNApurification/PlasmidDNApurificationeasyfastreliable/NucleoSpinPlasmidplasmidMiniprepkit/tabid/1379/language/en-US/Default.aspx. (2018).
23. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H., Roe, B.A. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *Journal of Molecular Biology.* **143** (2), 161-178 (1980).
24. Addgene. *Protocol - How to Create a Bacterial Glycerol Stock.* https://www.addgene.org/protocols/create-glycerol-stock/. (2018).
25. *Making your own chemically competent cells.* NEB. https://international.neb.com/protocols/2012/06/21/making-your-own-chemically-competent-cells. (2018).
26. *Luria-Bertani (LB) Medium Preparation · Benchling.* https://benchling.com/protocols/gdD7XI0J/luria-bertani-lb-medium-preparation. (2018).
27. Delebecque, C.J., Silver, P.A., Lindner, A.B. Designing and using RNA scaffolds to assemble proteins in vivo. *Nature Protocols.* **7** (10), 1797-1807 (2012).
28. Hocine, S., Raymond, P., Zenklusen, D., Chao, J.A., Singer, R.H. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nature Methods.* **10** (2), 119-121 (2013).
29. Espah Borujeni, A. et al. Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Research.* **45** (9), 5437-5448 (2017).
30. Ding, Y. et al. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature.* **505** (2013).
31. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., Weissman, J.S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature.* **505** (7485), 701-705 (2014).
32. Lucks, J.B. et al. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America.* **108** (27), 11063-11068 (2011).
33. Spitale, R.C. et al. Structural imprints *in vivo* decode RNA regulatory mechanisms. *Nature.* **519** (7544), 486 (2015).
34. Watters, K.E., Abbott, T.R., Lucks, J.B. Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucleic Acids Research.* **44** (2), e12 (2016).
35. Flynn, R.A. et al. Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. *Nature Protocols.* **11** (2), 273-290 (2016).
36. Bernardi, A., Spahr, P.-F. Nucleotide Sequence at the Binding Site for Coat Protein on RNA of Bacteriophage R17. *Proceedings of the National Academy of Sciences of the United States of America.* **69** (10), 3033-3037 (1972).

Article

Cell Systems

Synthetic 5′ UTRs Can Either Up- or Downregulate Expression upon RNA-Binding Protein Binding

Graphical Abstract



Highlights

- RBPs can up- or downregulate translation by direct interaction with synthetic 5′ UTR
- The type of regulation is dependent on RNA structure in the absence of the RBPs
- RNA behavior *in vivo* provides support for a tri-phasic structural model
- The tri-phasic structural model provides an explanation for the dual-regulatory role

Authors

Noa Katz, Roni Cohen, Oz Solomon, ..., Zohar Yakhini, Sarah Goldberg, Roee Amit

Correspondence

roeeamit@technion.ac.il

In Brief

Katz et al. developed RNA "parts" that are able to stimulate or repress expression of a target gene via their direct interaction with RNA-binding proteins (RBPs). The type of RBP regulation is dependent on RNA structure in their absence. This dual-regulatory role can be explained by a tri-phasic model, where each structural state of the RNA—molten-unbound, structured-unbound, and semi-structured-bound—is characterized by a state in the phase diagram. Their work provides new insight into RBP-RNA regulation and a blueprint for designing RNA-based regulatory circuits.

Katz et al., 2019, Cell Systems *8*, 93–106
July 24, 2019
https://doi.org/10.1016/j.cels.2019.04.007

CellPress

Cell Systems
Article

Synthetic 5′ UTRs Can Either Up- or Downregulate Expression upon RNA-Binding Protein Binding

Noa Katz,[1] Roni Cohen,[1] Oz Solomon,[1,3] Beate Kaufmann,[1] Orna Atar,[1] Zohar Yakhini,[2,3] Sarah Goldberg,[1] and Roee Amit[1,4,5,*]

[1]Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, 32000 Haifa, Israel
[2]Department of Computer Science, Technion - Israel Institute of Technology, 32000 Haifa, Israel
[3]School of Computer Science, Interdisciplinary Center, 46150 Herzeliya, Israel
[4]Russell Berrie Nanotechnology Institute, Technion - Israel Institute of Technology, 32000 Haifa, Israel
[5]Lead Contact
*Correspondence: roeeamit@technion.ac.il
https://doi.org/10.1016/j.cels.2019.04.007

SUMMARY

The construction of complex gene-regulatory networks requires both inhibitory and upregulatory modules. However, the vast majority of RNA-based regulatory "parts" are inhibitory. Using a synthetic biology approach combined with SHAPE-seq, we explored the regulatory effect of RNA-binding protein (RBP)-RNA interactions in bacterial 5′ UTRs. By positioning a library of RNA hairpins upstream of a reporter gene and co-expressing them with the matching RBP, we observed a set of regulatory responses, including translational stimulation, translational repression, and cooperative behavior. Our combined approach revealed three distinct states *in vivo*: in the absence of RBPs, the RNA molecules can be found in either a molten state that is amenable to translation or a structured phase that inhibits translation. In the presence of RBPs, the RNA molecules are in a semi-structured phase with partial translational capacity. Our work provides new insight into RBP-based regulation and a blueprint for designing complete gene-regulatory circuits at the post-transcriptional level.

INTRODUCTION

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 (Win and Smolke, 2008; Xie et al., 2011; Green et al., 2014; Wroblewska et al., 2015), even though RNA-based regulatory components have many advantages. Several RNA components have been shown to be functional in multiple organisms (Harvey et al., 2002; Suess et al., 2003; Desai and Gallivan, 2004; Buxbaum et al., 2015; Green et al., 2017). RNA can respond rapidly to stimuli, enabling a faster regulatory response than transcriptional regulation (Hentze et al., 1987; St Johnston, 2005; Saito et al., 2010; Lewis et al., 2017). From a structural perspective, RNA molecules can form a variety of biologically functional secondary and tertiary structures (Green et al., 2014), which enables modularity. For example, distinct sequence domains within a molecule (Khalil and Collins, 2010; Lewis et al., 2017) may target different metabolites or nucleic acid molecules (Werstuck and Green, 1998; Isaacs et al., 2006). All of these characteristics make RNA an appealing target for engineered-based applications (Hutvágner and Zamore, 2002; Rinaudo et al., 2007; Delebecque et al., 2011; Xie et al., 2011; Chen and Silver, 2012; Ausländer et al., 2014; Green et al., 2014; Sachdeva et al., 2014; Pardee et al., 2016).

Perhaps the most well-known class of RNA-based regulatory modules is riboswitches (Werstuck and Green, 1998; Winkler and Breaker, 2005; Henkin, 2008; Wittmann and Suess, 2012; Serganov and Nudler, 2013). Riboswitches are noncoding mRNA segments that regulate the expression of adjacent genes via structural change, effected by a ligand or metabolite. However, response to metabolites cannot be easily used as the basis of a regulatory network, as there is no convenient feedback or feed-forward mechanism for connection with additional network modules. Implementing network modules using RNA-binding proteins (RBPs) could enable an alternative multicomponent connectivity for gene-regulatory networks that is not based solely on transcription factors.

Regulatory networks require both inhibitory and upregulatory modules. The vast majority of known RBP regulatory mechanisms are inhibitory (Romaniuk et al., 1987; Cerretti et al., 1988; Brown et al., 1997; Schlax et al., 2001; Lim and Peabody, 2002; Sacerdot et al., 1998). A notable exception is the phage RBP Com, whose binding was demonstrated to destabilize a sequestered ribosome-binding site (RBS) of the Mu phage *mom* gene, thereby facilitating translation (Hattman et al., 1991; Wulczyn and Kahmann, 1991). Several studies have attempted to engineer activation modules utilizing RNA-RBP interactions, based on different mechanisms: recruiting the eIF4G1 eukaryotic translation initiation factor to specific RNA targets via fusion of the initiation factor to an RBP (De Gregorio et al., 1999; Boutonnet et al., 2004), adopting a riboswitch-like approach (Ausländer et al., 2014) and utilizing an RNA-binding version of the TetR protein (Goldfless et al., 2012). However, despite these notable efforts, RBP-based translational stimulation is still difficult to design in most organisms.

In this study, we employ a synthetic biology reporter assay and *in vivo* SHAPE-seq (Lucks et al., 2011; Spitale et al., 2013; Flynn et al., 2016) approach to study the regulatory effect controlled by

an RBP bound to a hairpin within the 5′ UTR of bacterial mRNA, following a design introduced by Saito et al. (2010). Our findings indicate that structure-binding RBPs (coat proteins from the bacteriophages GA [Gott et al., 1991], MS2 [Peabody, 1993], PP7 [Lim and Peabody, 2002], and Qβ [Lim et al., 1996]) can generate a range of translational responses, from previously observed downregulation (Saito et al., 2010) to upregulation. The mechanism for downregulation is most likely steric hindrance of the initiating ribosome by the RBP-mRNA complex. 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. These two experimental features indicate that the upregulatory elements constitute protein-responsive RNA regulatory elements. Our findings imply that RNA-RBP interactions can provide a platform for constructing gene-regulatory networks that are based on translational, rather than transcriptional, regulation.

RESULTS

RBP Binding Can Cause Either Upregulation or Downregulation

We studied the regulatory effect generated by four RBPs when co-expressed with a reporter construct containing native and non-native binding sites in the 5′ UTR (Figure 1A). The RBPs used (GCP, MCP, PCP, and QCP) were the coat proteins from the bacteriophages GA, MS2, PP7, and Qβ, respectively (see Table S2). In brief (Figure 1A; STAR Methods), we placed the binding site in the 5′ UTR of the mCherry gene at various positions upstream to the mCherry AUG, induced production of the RBP-mCerulean fusion by addition of N-butyryl-L-homoserine lactone (C_4-HSL) at 24 different concentrations, and measured both signals (mCherry and mCerulean) to calculate RBP response. An example signal for two duplicates of an upregulating strain using the mutated PCP-binding site PP7-wt positioned at $\delta = -31$ in the 5′ UTR is shown in Figure 1B. In the upper panel, the induction response can be seen for the PCP-mCerulean channel and in the lower panel, the mCherry rate of production for the particular 5′ UTR configuration that results from the induction is shown (see Supplemental Information for definition).

To facilitate a more efficient characterization of the dose response, we analyzed the mCherry production rate for all strains as a function of mCerulean levels. In Figure 1C (left), we present the sample dose-response results for MS2-U(-5)C, together with MCP, at all four different 5′ UTR positions assayed. A sigmoidal response can be observed for three out of the four configurations, with the fold change diminishing as the binding site is positioned closer to the RBS. For the $\delta = -23$ strain, we observed no change in response as a function of the amount of RBP in the cell. To facilitate proper comparison of the regulatory effect across strains, for each strain, we opted to normalize both the mCherry rate of production and mCerulean expression levels by their respective maximal value for each dose-response function. Such a normalization allows us to properly compare between strains fold-regulation effects, and effective dissociation constant (K_{RBP}), by in effect eliminating the dependence on basal mCherry rate of production, and the particular maximal RBP expression levels. Finally, we sorted all normalized dose responses in accordance with increasing fold upregulation effect and plotted the dose responses obtained in the experiment as a single heatmap, facilitating convenient further study and presentation of the data (Figure 1D).

We constructed our 5′ UTR variants using 11 putative binding sites for the phage coat proteins depicted in Figure 2A. These structures are based on the three native sites for the RBPs, MS2-wt, PP7-wt, and Qβ-wt (in bold). Different mutations were introduced, some structure altering, such as the PP7 upper stem short (PP7-USs) and PP7 no-bulge (PP7-nB), and some structure preserving, such as the MS2-U(-5)C and Qβ-upper stem, lower stem, and loop mutated (Qβ-USLSLm). The minimum free energy of the structure also varies, depending on the kind of mutations introduced. A few mutations in the structure of the binding site can greatly influence the stability of the structure, as is the case for PP7-nB and Qβ-USLSLm.

We positioned each of the 11 binding sites at three or four different locations upstream of the RBS, that ranged from $\delta = -21$ to $\delta = -35$ nt measured relative to the AUG of the mCherry reporter gene (see Table S1). Altogether, we constructed 44 reporter constructs (including non-hairpin controls), and co-transformed with all four RBPs, resulting in a total of 176 regulatory strains. The normalized and sorted dose-response heatmap for the 5′ UTR constructs for all strains is plotted in Figure 2B. The dose-response functions are arranged in order of increasing fold upregulation response, with the strongest-repression variants depicted at the bottom. The plot shows that there is a great diversity of responses. We found 24 upregulating strains (top of the heatmap) and 30 downregulating strains (bottom of the heatmap), and the remaining variants were not found to generate a statistically significant dose response. A closer examination indicates that the observed repression is generally weak, and at most amounts to about a factor of two reduction from basal levels (turquoise, bottom of the heat map). Notably, the top of the heatmap reveals a moderate upregulatory dose response (variant # > 140) of up to ~5-fold, which was not previously observed for these RBPs.

Next, we computed the K_{RBP} for all dose-responding strains, which is defined as the fitted dissociation constant (see STAR Methods) normalized by the maximal mCerulean expression level. The resultant $K_{RBP\ value}$s obtained for each RBP-binding-site pair are plotted as a heatmap in Figure 2C. Note that we did not find a position dependence on the values of K_{RBP} in this experiment (see Supplemental Information), and thus, the values depicted in the heatmap represents an average over multiple 5′ UTR positions. The heatmap shows similar K_{RBP} values (up to an estimated fit error of 10%) for all binding-site positions, for each of the native binding sites (MS2-wt, PP7-wt, and Qβ-wt) and for the mutated sites with a single mutation (non-structure altering) in the loop region (MS2-U(-5)C and MS2-U(-5)G). However, for mutated binding sites characterized by small structural deviations from the native structure (PP7-nB and PP7-USs), and for RBPs that bind non-native binding sites (e.g., MCP with Qβ-wt), a higher K_{RBP} was recorded. Furthermore, deviations in K_{RBP} were also observed for several of the mutated sites in comparison to a similar measurement that was reported by us recently, when the binding sites were positioned in the ribosomal initiation region (Katz et al., 2018). In particular, both Qβ-USLSLm and Qβ-LSs generated a downregulatory dose-response signal in the 5′ UTR in the presence of QCP, while no

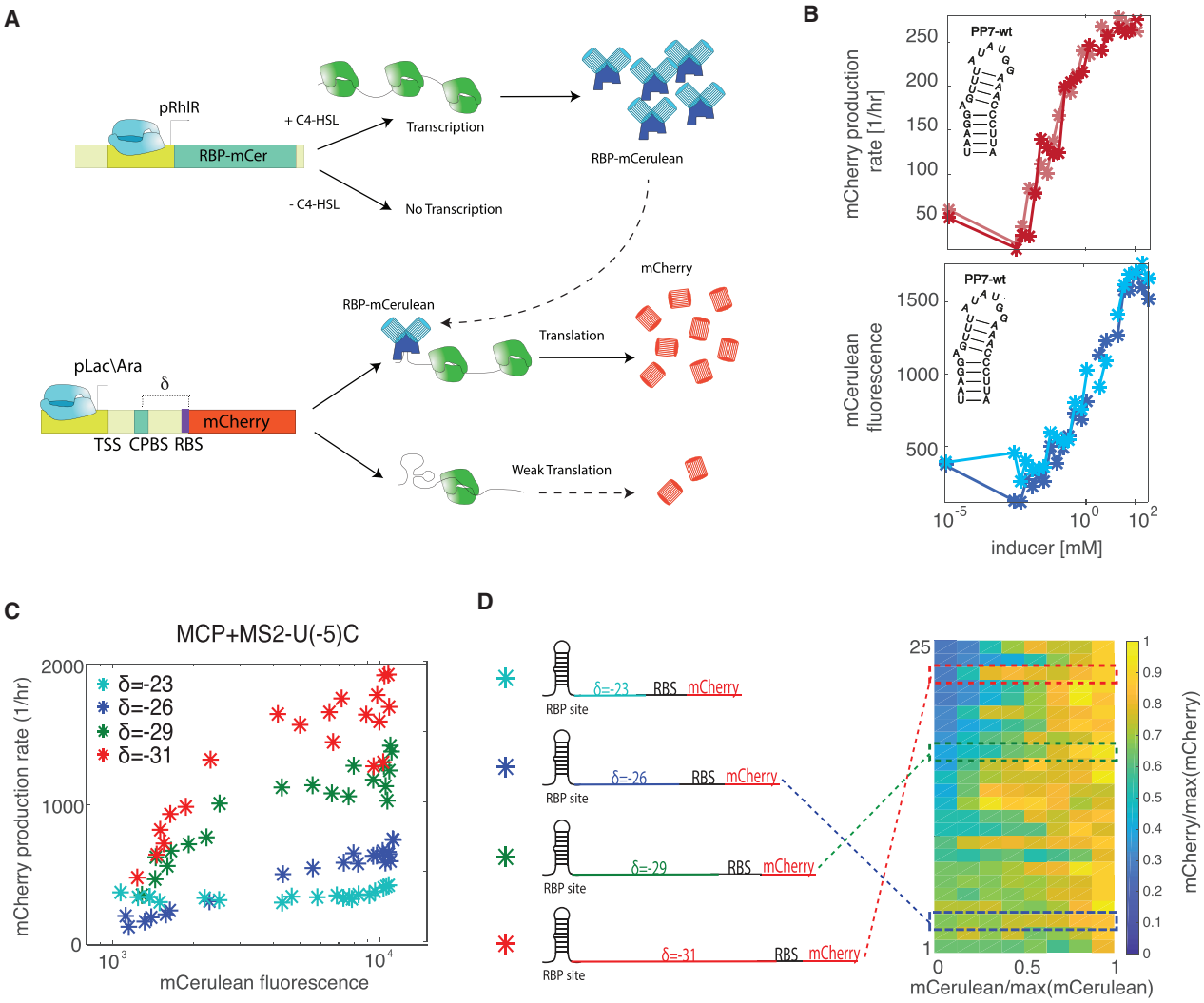


Figure 1. Experimental Schematic

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.cels.2019.04.007.

(A) Schematic of the experimental system. Top: plasmid expressing the RBP-mCerulean fusion from a pRhlR inducible promoter. Bottom: a second plasmid expressing the reporter mCherry with the RBP-binding site encoded within the 5′ end of the gene (at position $\delta < 0$). CPBS, coat-protein-binding site; TSS, transcription start site; RBS, ribosome-binding site.

(B) A sample dataset showing the two fluorescent channels separately for PP7-wt. Top: mCerulean mean production rate plotted as a function of C_4-HSL inducer concentration. Bottom: mCherry reporter expressed from a constitutive pLac/Ara promoter plotted as a function of inducer concentration showing an upregulatory response that emerges from the RBP-RNA interaction.

(C) mCherry production rate for MS2-U(-5)C at four different locations in the 5′ UTR.

(D) Illustration of the MS2-U(-5)C site at four different locations (left) and a heatmap of the dose responses for upregulating variants in the 5′ UTR (right).

response was detected in the ribosomal initiation region configurations. Conversely, QCP generated a response with the MS2-based sites, MS2-wt and MS2-U(-5)C, in the ribosomal initiation region, while no apparent response was detected when these binding sites were placed in the 5′ UTR. Finally, past *in vitro* studies have recorded a dose-response function for MS2-wt, MS2-U(-5)C, and PP7-USLSBm in the presence of QCP and PCP, while no such effect was observed here for QCP and PCP for any of these sites. Consequently, the nature of the dose response and the mere binding of a protein to a site seems to depend on additional parameters that are not localized solely to the binding site.

5′ UTR Strains Present Three Translational States

To further study the different types of dose responses (up- or downregulation), for each RBP-binding-site pair that generates a dose response, we plotted the maximal fold-change effect that was recorded over the range of 5′ UTR positions (Figure 3A). In the panel, we show both maximal down (depicted as fold values < 1) and upregulatory dose-response fold changes. The figure shows that the nature of the response does not depend on the RBP but rather on the binding sites. In particular, both MCP and GCP generate an upregulatory response for the binding sites MS2-wt, MS2-U(-5)G, and MS2-U(-5)C. Likewise, both MCP and QCP generate a downregulatory response for Qβ-wt

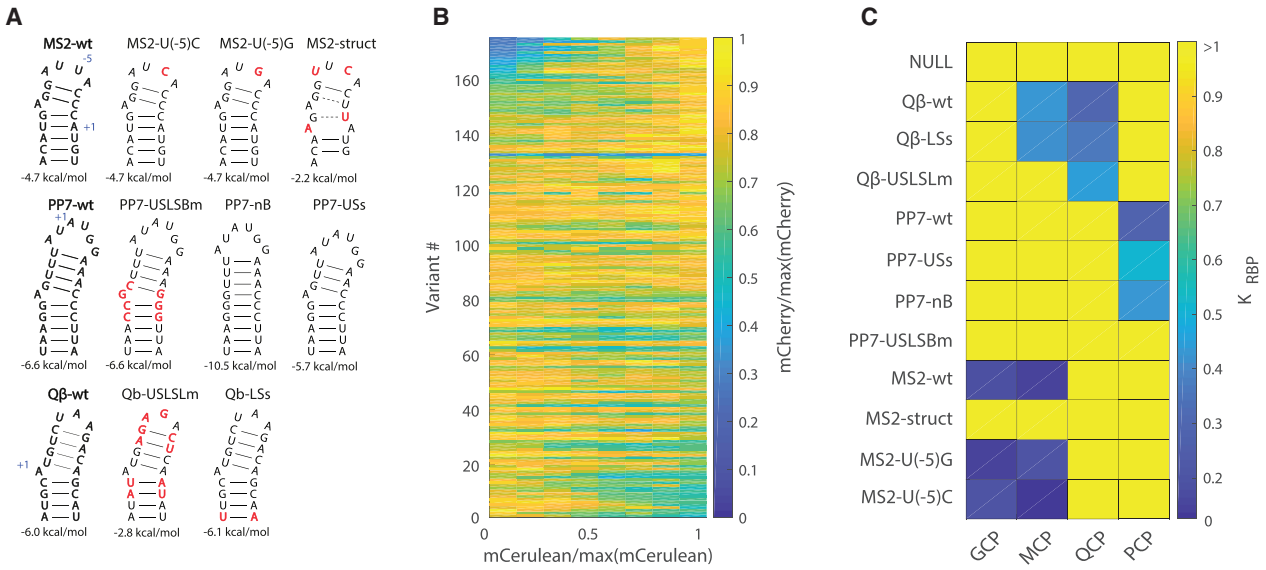


Figure 2. Translational Stimulation and Repression upon RBP Binding in the 5′ UTR

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.cels.2019.04.007.

(A) Secondary structure schematic for the 11 binding sites used in the study. Red nucleotides indicate mutations from the original wt binding sequence. US, upper stem; LS, lower stem; L, loop; B, bulge; m, mutations; s, short; struct, significant change to the binding site structure.

(B) Heatmap of the dose responses of the 5′ UTR variants. Each response is divided by its maximal mCherry/mCerulean level for easier comparison. Variants are arranged in order of increasing fold upregulation.

(C) Normalized K_{RBP} averaged over the different positions. Blue corresponds to low K_{RBP} while yellow indicates no binding. If there was no measurable interaction between the RBP and binding site, K_{RBP} was set to 1. NULL represents no binding site.

and Qβ-LSs. Conversely, structural mutations that conserve binding (PP7-USs and PP7-nB) can alter the dose response of PCP from upregulating (PP7-wt) to downregulating. Finally, for the case of MS2 with MCP, the size of the fold effect seems to depend on the exact sequence of the binding site. Here, while the native MS2-wt binding site exhibited a maximal fold-change effect of ~2, a single mutation to the loop region caused the response to increase to a factor of 5-fold activation. Taken together, our data indicate that the nature of the response is dependent on the binding-site sequence at a single-nucleotide resolution.

We next studied the relationship between the position of the binding site within the 5′ UTR and size of the fold effect. In Figure 3B (top), we plot the fold effect for all RBP-binding-site pairs as a function of 5′ UTR position. First, we note that changing the length of the sequence segment downstream to the binding sites does not alter the nature of the dose response. Second, the plots show that for both the fold repression and fold activation, the effect is mostly unaffected by changing the position of the binding site within the 5′ UTR, except when it is placed in a high proximity to the RBS (position $\delta = -23$), where the activation is diminished. Plots of the basal production rate of both types of strains show a similar picture (Figure 3B, bottom), with the fold activation diminishing as the distance from the RBS is reduced. Next, we compared the absolute rate of production levels between the upregulating and downregulating strains, for both the non-induced (Figure 3C) and fully induced (Figure 3D) states. For the non-induced states, the mean rate of production of the upregulating strains is around a factor of three less than the mean for the downregulating strains. Conversely, for the induced state, both distributions converge and present less than a factor of two difference between the two calculated mean levels. This indicates the translational level associated with the RBP-bound mRNA is similar for all 5′ UTR constructs, independent of the particular binding site or RBP present. Taken together, a picture emerges where there are three main translational states for the 5′ UTR and associated *mCherry* gene, each with its own range of resultant mCherry levels: a closed translationally inactive state occurring for the non-induced upregulating strains, where the mRNA is predominantly unavailable for translation; an open translationally active state, which occurs for the non-induced downregulating strains; and finally, a partially active translational state, which is characterized by an RBP-bound 5′ UTR.

In Vitro Structural Analysis with SHAPE-Seq Exhibits a Single Structural State

Our reporter assay analysis and past results by us and others indicate that there seem to be other factors in play that influence RBP binding and the nature of the dose response. A prime candidate is the molecular structure that forms *in vivo* in the presence and absence of the binding protein. This structure is influenced by the sequences that flank the binding site and the minimum free energy of the hairpin itself. This led us to hypothesize that each state is characterized by a structural fingerprint, which, in turn, is dependent on binding site structure and stability as well as the flanking sequences. To test our proposed scenario, we chose to focus on two 5′ UTR variants from our library, which encoded the PP7-wt and PP7-USs binding sites, both at $\delta = -29$. In this test case, the entire 5′ UTR is identical for both variants except for a deletion of two nucleotides in the upper stem of

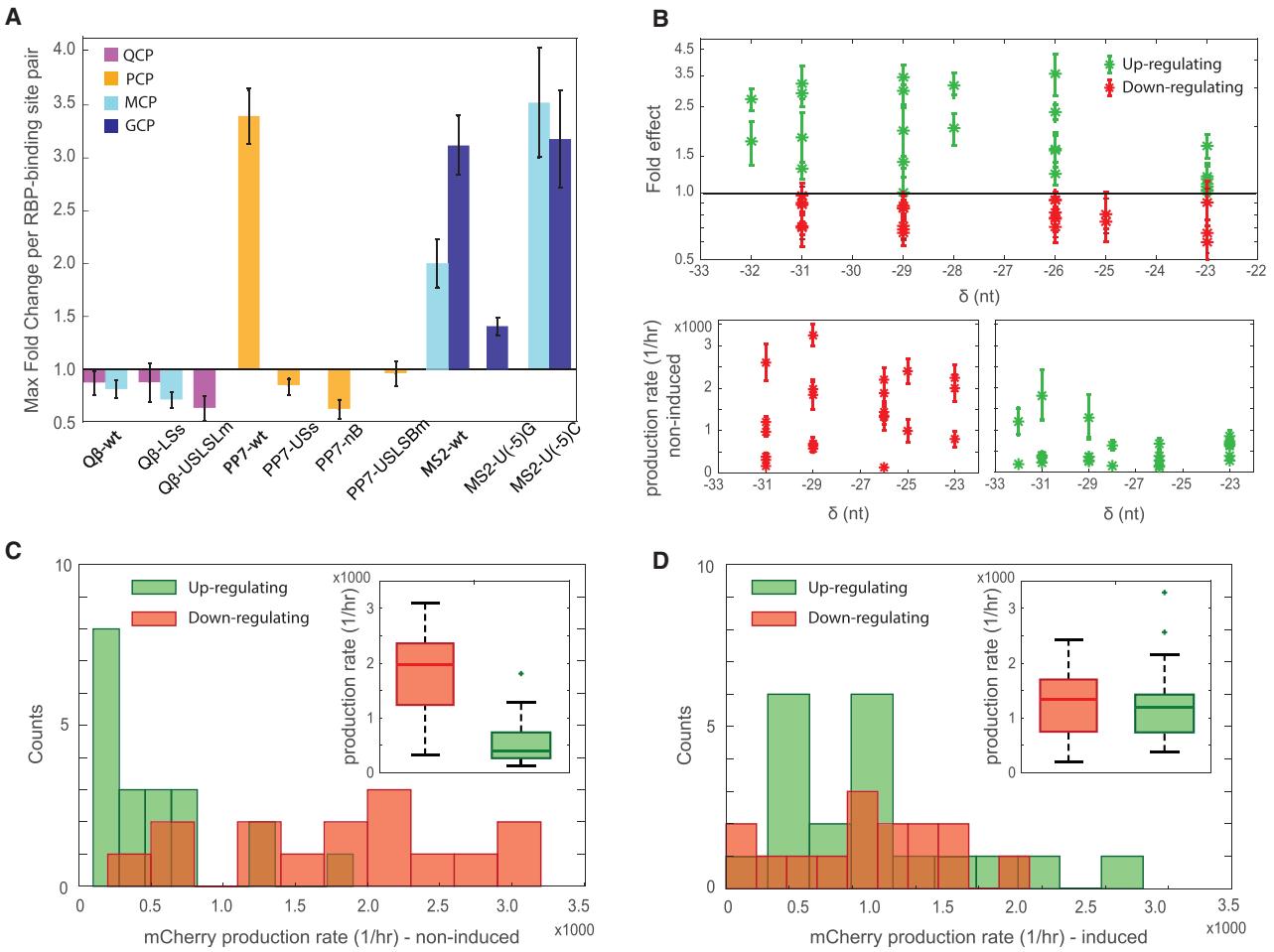


Figure 3. Reporter Assay Indicates that There May Be Three Distinct Translational States

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.cels.2019.04.007.

(A) Bar graph showing maximal fold change of each RBP-binding-site pair for all 11 binding sites as follows: QCP-mCerulean (purple), PCP-mCerulean (yellow), MCP-mCerulean (light blue), and GCP-mCerulean (dark blue). Values larger and smaller than one correspond to up- and downregulation, respectively. The MS2-struct binding site was omitted from the plot because of no observable effect with all RBPs.

(B) Top: Fold effect as a function of position for upregulating strains (green) and downregulating strains (red). Each point represents a single RBP-binding-site pair. Error bars represent standard deviation from at least two replicates. Bottom: Basal mCherry production rate as a function of position for downregulating strains (left) and upregulating strains (right).

(C and D) Histograms of mCherry production rate for both regulatory populations along with matching boxplots (inset) at the non-induced (C) and induced states (D). Mann-Whitney U test (MWW) on the two populations showed a p value of 1.5702e−04 for the non-induced state and 0.4822 for the induced state.

the PP7-wt site, which results in the PP7-USs site. This deletion reduces the stability of the PP7-USs binding site (−5.7 kcal/mol) as compared with the native PP7-wt site (−6.6 kcal/mol). First, we wanted to ensure that these variants exhibit the three translational states in their dose response (Figure 4A). Here, the PP7-wt response function exhibits a low production rate in the absence of induction (∼150 a.u./h) while rising in a sigmoidal fashion to an intermediate production rate (∼450 a.u./h) at full induction. For PP7-USs, the basal rate of production level at zero induction is nearly an order of magnitude larger at ∼1,100 a.u./h and declines gradually upon induction to an intermediate level similar to that observed for PP7-wt.

Next, we calculated the predicted structure for these two 5′ UTR variants using RNAfold (Hofacker et al., 1994). As expected, the small reduction in binding site stability did not affect the computed structures (Figure 4B), and both predicted model structures seem identical. Therefore, we chose to directly probe the mRNA structure via SHAPE-seq. We subjugated the two strains to SHAPE-seq *in vitro* using 2-methylnicotinic acid imidazole (NAI) suspended in anhydrous dimethyl sulfoxide (DMSO), with DMSO-treated cells as a non-modified control (see STAR Methods; Figure S1 for SHAPE-seq analysis of 5S-rRNA as positive control). We chose to modify a segment that includes the entire 5′ UTR and another ∼140 nt of the mCherry reporter gene. In Figure 4C, we plot the reactivity signals as a function of nucleotide position on the mRNA obtained for both the PP7-wt (blue line) and PP7-USs (red line) constructs at δ = −29 using *in vitro* SHAPE-seq, after alignment of the two signals (see STAR Methods). The reactivity of each base corresponds to the propensity of that base to be modified by NAI (for the definition of

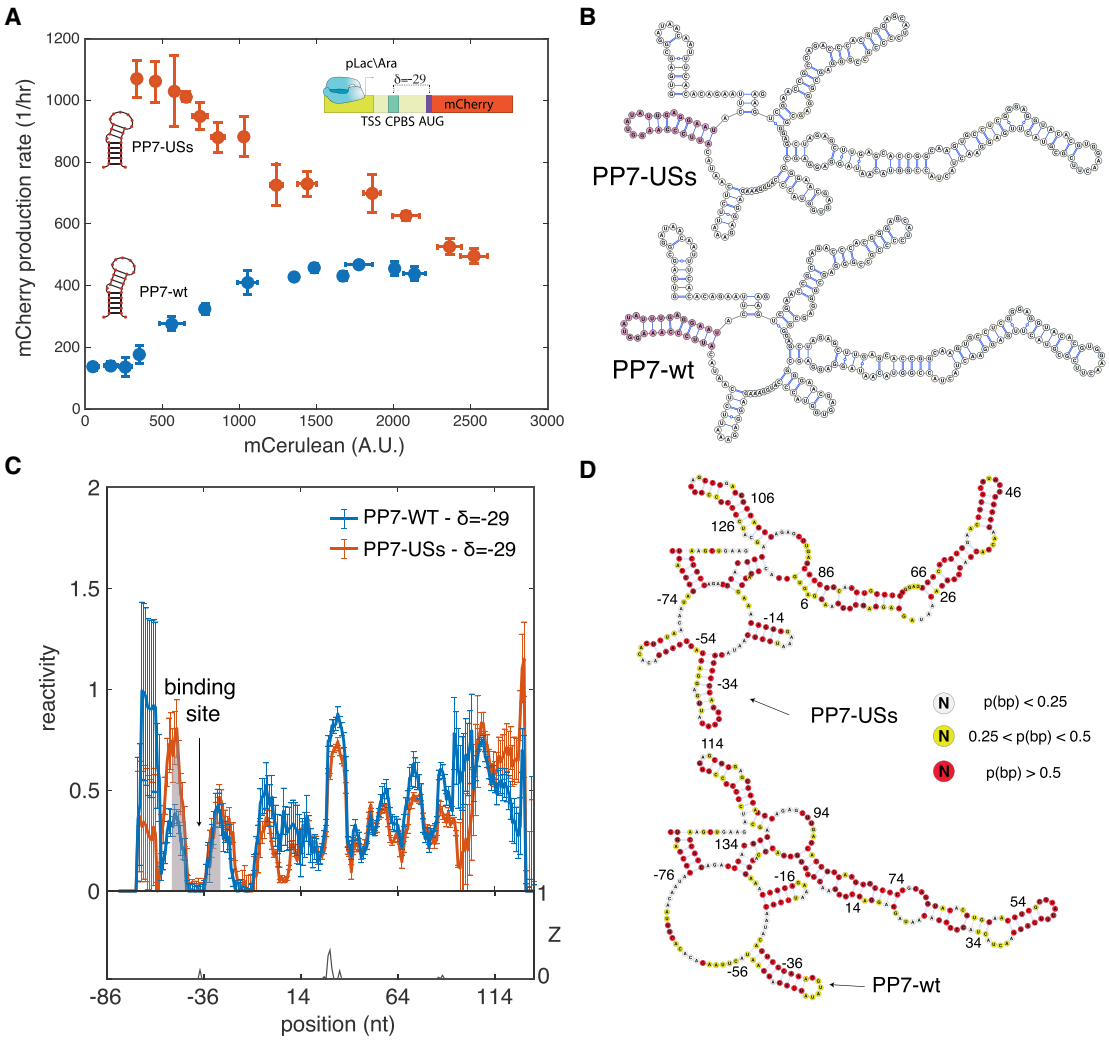


Figure 4. *In Vitro* SHAPE-Seq Analysis Does Not Reveal Two Distinct Structural States without RBP

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.cels.2019.04.007.

(A) Dose-response functions for two strains containing the PP7-wt (blue) and PP7-USs (red) binding sites at $\delta = -29$ nt from the AUG. Each data point is an average over multiple mCerulean and mCherry measurements taken at a given inducer concentration. Error bars signify the standard deviation computed from these measurements.

(B) Structure schemes predicted by RNAfold for the 5′ UTR and the first 134 nt of the PP7-wt and PP7-USs constructs (using sequence information only).

(C) *In vitro* reactivity analysis for SHAPE-seq data obtained for two constructs PP7-wt (blue) and PP7-USs (red) at $\delta = -29$. Error bars are computed using bootstrapping resampling of the original modified and non-modified libraries for each strain (see STAR Methods) and are also averaged from two biological replicates. The data from the two extra bases for PP7-wt were removed for alignment purposes.

(D) Inferred *in vitro* structures for both constructs are constrained by the reactivity scores from (B). Each base is colored by its base-pairing probability (red, high; yellow, intermediate; and white, low) calculated based on the structural ensemble via RNAsubopt (Lorenz et al., 2011). Associated with Figure S1.

reactivity, see STAR Methods). 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 yields significant distinguishability only for a narrow segment within the coding region (~+30 nt). We then used the *in vitro* reactivity data to compute the structure of the variants by guiding the computational prediction (Deigan et al., 2009; Ouyang et al., 2013; Washietl et al., 2012; Zarringhalam et al., 2012). In Figure 4D, we show that the SHAPE-derived structures for both constructs are similar to the results of the initial non-constrained RNAfold computation (Figure 4B) and are nearly indistinguishable from each other. Consequently, the *in vitro* SHAPE-derived structures and reactivity data for the two 5′ UTR variants do not reveal two distinct structural states, which are a precursor for a third RBP-bound state.

In Vivo SHAPE-Seq Reveals Three Structural States Supporting the Three-Translation-Level Hypothesis

Next, we carried out the SHAPE-seq protocol *in vivo* (see STAR Methods) on induced and non-induced samples for the two variants. In Figure 5A, we plot the non-induced (RBP−) reactivity obtained for PP7-wt (blue) and PP7-USs (red). The data show

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. In Figure 5B, we show that in the induced state (RBP+) both constructs exhibit a weak reactivity signal that is statistically indistinguishable in the 5′ UTR (i.e., Z-factor ~0 at $\delta < 0$). In particular, the region associated with the binding site is unreactive (marked in gray), indicating that the binding site and flanking regions are either protected by the bound RBP, highly structured, or both (see Figure S2 for further analysis). Consequently, contrary to the *in vitro* SHAPE analysis, for the *in vivo* case the reactivity data for the non-induced case reveal a picture consistent with two distinct translational states, for a sum total of three states when taking the induced reactivity data into account.

To generate additional structural insight, we implemented the constrained structure computation that was used for the *in vitro* samples on the PP7-wt ($\delta = -29$) and PP7-USs ($\delta = -29$) variants (Figure 5C). 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 4D). Neither displays the PP7-wt or PP7-USs binding site, and a secondary hairpin encoding a putative short anti-Shine-Dalgarno (aSD) motif (CUCUU) (Levy et al., 2017), which may partially sequester the RBS, appears only 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: an upstream flanking hairpin (−72 to −57 for PP7-wt), the binding site (−54 to −30 for PP7-wt), and downstream CUCUU aSD satellite structure (−23 to −10 for both). Taken together, the SHAPE-derived structures for the non-induced and induced strains support three distinct structural configurations for the 5′ UTR, which are consistent with the reporter assay findings and can thus be associated with their respective translational levels.

Changes to 5′ UTR Sequence Can Alter Translational State

We reasoned that we can influence the regulatory response by introducing mutations into the 5′ UTR sequence that can shift the structure from the translationally inactive state to the translationally active state. To do so, we mutated the structure of the flanking sequences in three ways (Figure 6A): first, by changing the CUCUU motif from the original strains (Figure 6A, bottom left) into an A-rich segment (Figure 6A, top right), thus potentially reducing structure formation in the 5′ UTR and potentially shifting the upregulatory response to a repression effect; second, by enhancing the aSD motif in the original strains (Figure 6A, top left), thus encouraging the formation of a structured 5′ UTR and potentially increasing the fold effect of the upregulatory strains; and finally, by extending the lower stem of MS2-wt and PP7-wt binding sites by three, six, and nine base pairs to increase binding site stability (Figure 6A, bottom right). We hypothesized that this set of new 5′ UTR variants could help us expand our understanding of the mechanism involved in translational regulation.

First, we synthesized ten additional constructs at $\delta = -29$ with PP7-nB, PP7-USs, PP7-wt, MS2-wt, or MS2-U(-5)C binding sites in which the sequence between the binding site and the RBS encoded either a strong CU-rich motif or an A-rich segment (see Table S1). We plot the basal expression level for 15 RBP-binding-site pairs containing the original spacer (green), the spacer with the CU-rich sequence (red), and the A-rich spacer lacking the aSD sequence (blue). The data (Figure 6B, left heatmap) show that the constructs with a CU-rich flanking region exhibit lower basal expression levels than the other constructs, as predicted and previously observed (Levy et al., 2017), while the different A-rich variants do not seem to affect basal expression in a consistent fashion. However, both the upregulatory and downregulatory dose responses persist independently of the flanking region content (Figure 6B, right heatmap, top and middle), compared with the response recorded for the original flanking sequences (Figure 6B, right heatmap, bottom).

To check the effect of increasing binding site stability, we designed 6 new variants for the PP7-wt binding sites by extending the length of the lower stem by three, six, and nine base pairs with complementary flanking sequences that are either GU or GC repeats (Figure S3; Table S1). When examining the dose-response functions (Figures 6C and 6D), the upregulatory responses were converted to downregulating responses for all configurations. The basal expression levels for the non-induced state was increased by 3- to 10-fold (Figure 6D, left heatmap), consistent with the levels previously observed for the non-structured, translationally active state. Upon induction, the downregulatory effect that was observed resulted in rate-of-production levels that approached the levels of the original PP7-wt construct at full induction (Figure 6C), further corroborating the three-state model. Yet, for all stem-extended constructs, the K_{RBP} increased by 2- to 3-fold (Figure S3), indicating a potentially weaker binding that may be due to the increased translational activity associated with these constructs. Finally, we checked the effect of temperature on regulation. We studied several strains (RBP-binding-site combinations) in temperatures that ranged from 22°C to 42°C and found no significant change in regulatory effect for any of the variants studied (Figure S4). Consequently, it seems that only mutations that are associated with binding site stability seem to affect the state of the non-induced state, whether it will be non-structured and translationally active or highly structured and translationally inactive.

A Tandem of Binding Sites Can Exhibit Both Cooperativity and Complete Repression

Finally, to further explore the regulatory potential of the 5′ UTR, we synthesized 28 additional 5′ UTR variants containing two binding sites from our cohort (Figure 2A), one placed in the 5′ UTR ($\delta < 0$), and the other placed in the ribosomal initiation region ($1 < \delta < 15$) of the mCherry gene (Figure 7A). In Figures 7B–7D, we plot the dose responses of the tandem variants in the presence of MCP, PCP, and QCP as heatmaps arranged in order of increasing basal mCherry rate of production. Overall, the basal mCherry production rate for all the tandem variants is lower

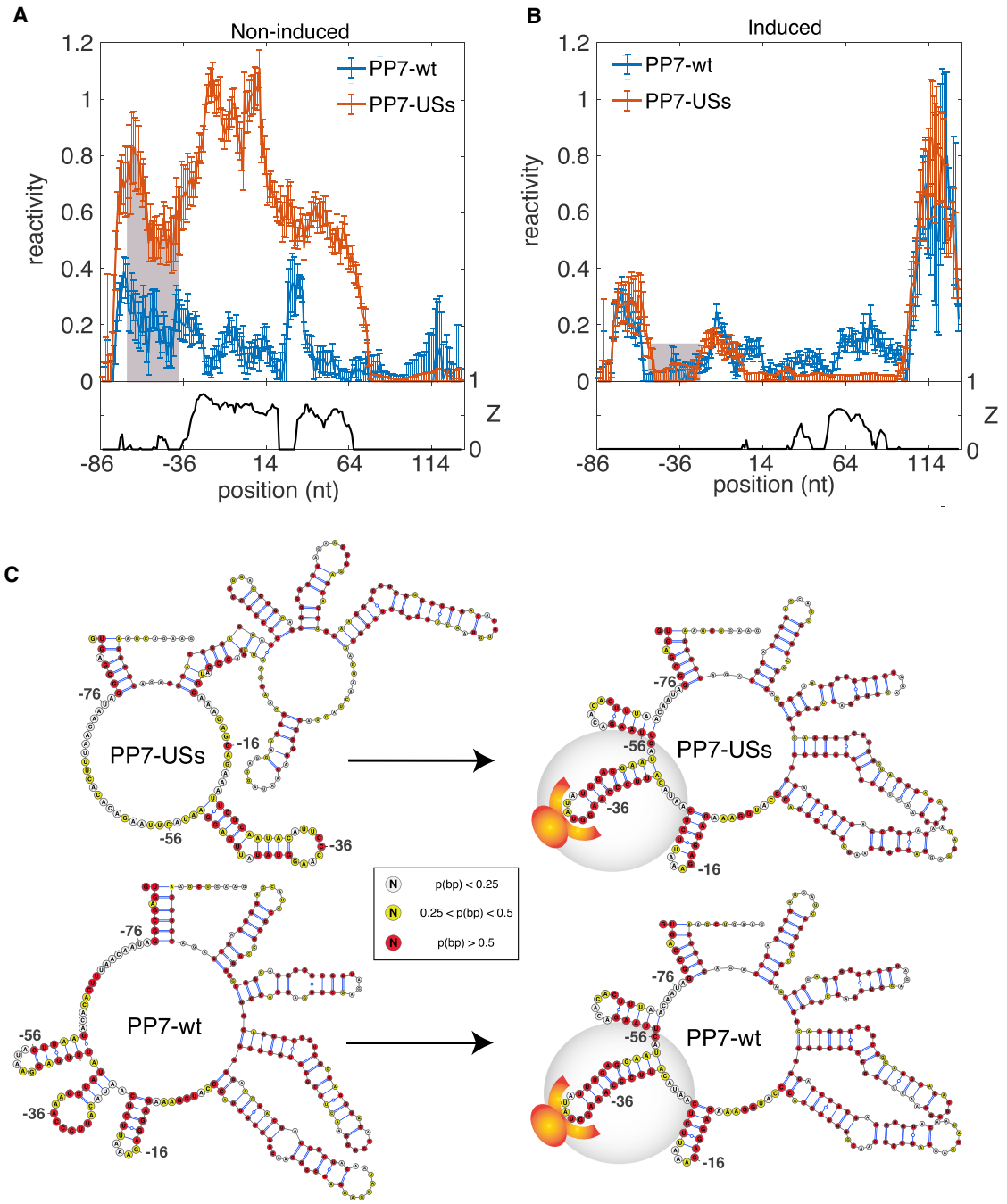


Figure 5. *In Vivo* SHAPE-Seq Analysis for PP7-wt and PP7-USs Strains Reveals Three Structural States

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.cels.2019.04.007.

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

(C) Inferred *in vivo* structures for all 4 constructs and constrained by the reactivity scores shown in (A) and (B). Each base is colored by its base-pairing probability (red, high; yellow, intermediate; and white, low) calculated based on the structural ensemble via RNAsubopt (Lorenz et al., 2011). 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.

Associated with Figure S2.

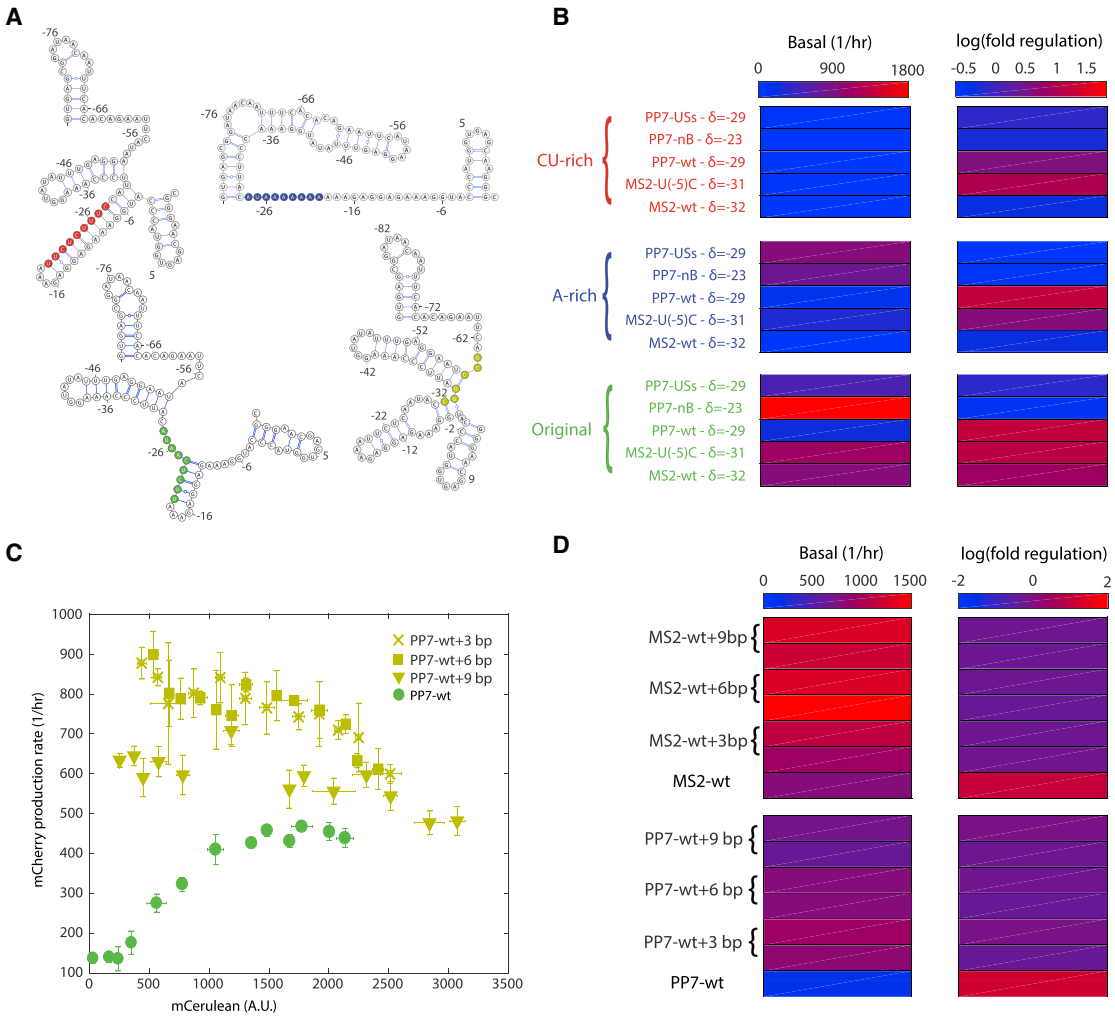


Figure 6. Nature of Fold Regulation Is Dependent on Flanking Sequences

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.cels.2019.04.007.

(A) Schematics for four sample structures computed with RNAfold (using sequence information only), where a short segment of the flanking region to the hairpin was mutated in each strain. Three structures contain the PP7-wt hairpin at $\delta = -29$. Top, left: CU-rich flanking colored in red. Top, right: A-rich flanking colored in blue. Bottom, left: original construct with "random" flanking sequence colored in green. Bottom, right: PP7-wt hairpin encoded with a longer stem colored in yellow.

(B) Variants containing 5 distinct hairpins with either CU-rich (red), A-rich (blue), or original (green) flanking sequences upstream of the RBS. While basal levels are clearly affected by the presence of a strong CU-rich flanking sequence, the nature of the regulatory effect is apparently not determined by the sequence content of the flanking region.

(C) Dose-response functions for PP7-wt binding sites with an extra 3 (x's), 6 (squares), and 9 (triangles) stem base pairs are shown relative to the dose response for PP7-wt (green). Each data point is an average over multiple mCerulean and mCherry measurements taken at a given inducer concentration. Error bars signify the standard deviation computed from these measurements.

(D) Basal levels and logarithm (base 2) of fold change for dose responses of all extended stem constructs with their corresponding RBPs (MCP or PCP). Associated with Figures S3 and S4.

than the single-binding-site variants located in the 5′ UTR. In addition, approximately half of the variants generated a significant regulatory response in the presence of the RBP, while the other half seem to be repressed at the basal level, with no RBP-related effect detected.

For MCP (Figure 7B), we observed strong repression for four of the ten variants tested, with the MS2-U(-5)G binding site positioned in the ribosomal initiation region for all four repressed variants. With different ribosomal initiation region binding sites (MS2-wt, Qβ-wt, or MS2-U(-5)C), basal mCherry rate of production was reduced to nearly zero. For PCP (Figure 7C), a similar picture emerges, with several variants exhibiting a strong dose-response repression signature, while no regulatory effect was observed for others. In terms of basal mCherry production rate, the variants in the top six all encode the PP7-nB binding site in the 5′ UTR. Moreover, all eight variants with a PP7-nB positioned in the 5′ UTR exhibit a downregulatory response. These observations are consistent with the data shown in Figures 6C and 6D, where the binding sites with longer stems resulted in larger basal mCherry rate of production, presumably

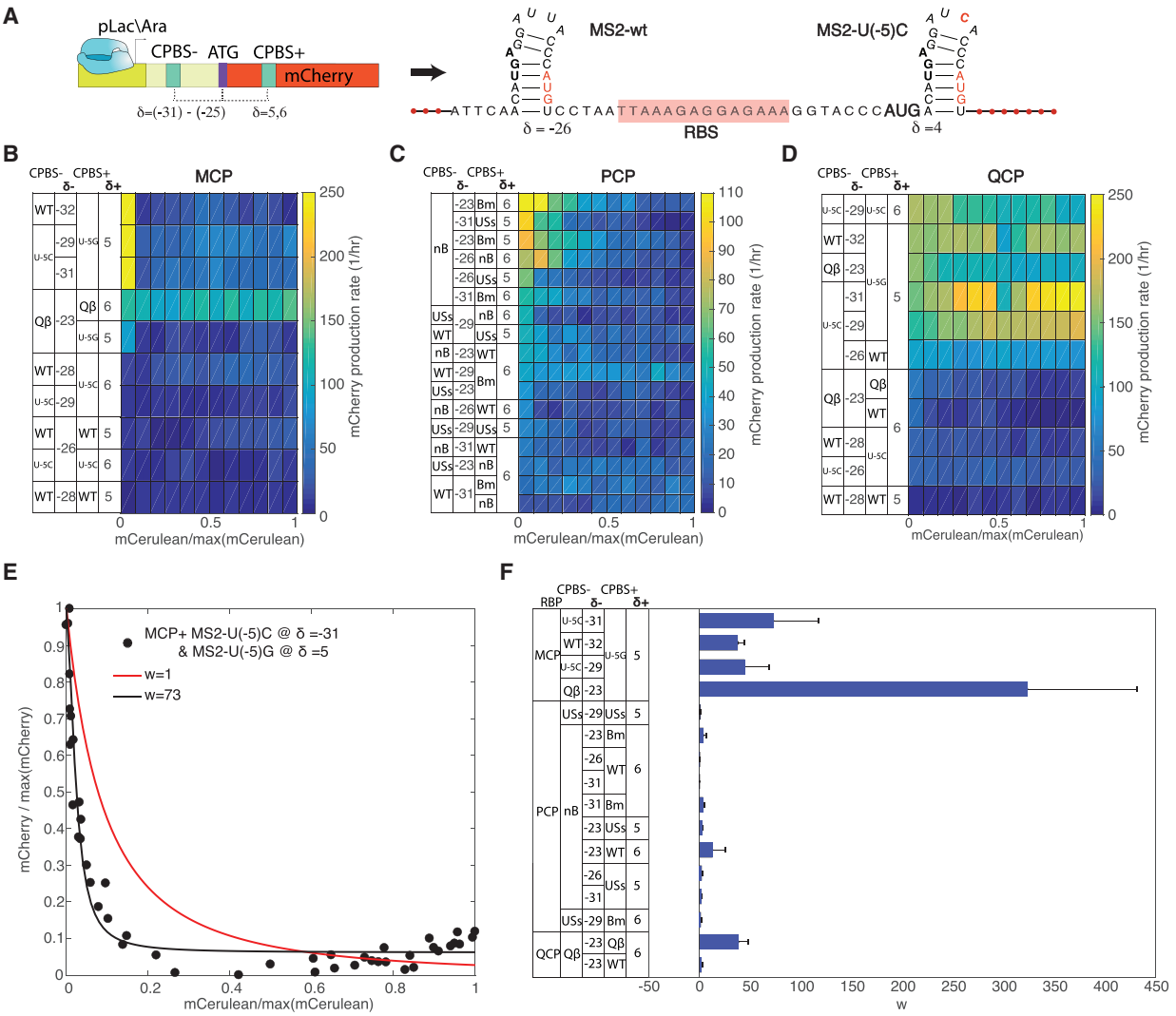


Figure 7. mRNAs with a Tandem of Hairpins

Figure360▷ For a Figure360 author presentation of this figure, see https://doi.org/10.1016/j.cels.2019.04.007.

(A) Schematic of the mRNA molecules with a single binding site at the 5′ UTR (δ < 0) and a single binding site in the gene-header region (δ > 0). Extra bases were added downstream to the binding site where necessary to retain the open reading frame.

(B–D) Heatmap corresponding the dose-response function observed for MCP (B), PCP (C), and QCP (D). In all heatmaps, the dose response is arranged in order of increasing mCherry rate of production, with the lowest-expressing variant at the bottom. The binding-site abbreviations are as follows: for MCP (B) and QCP (D), WT is MS2-wt, U(-5)G is MS2-U(-5)G, U(-5)C is MS2-U(-5)C, and Qβ is Qβ-wt. For PCP (C), WT is PP7-wt, nB is PP7-nB, Bm is PP7-LSLSBm, and USs is PP7-USs.

(E) A sample fit using the cooperativity model (see Supplemental Information).

(F) Bar plot depicting the extracted cooperativity factors *w* for all the tandems that displayed either an up- or downregulatory effect. Error bars signify the error computed in the fit for *w* using our model (see STAR Methods).

Associated with Figures S5 and S6.

because of increased hairpin stability. For other PP7-binding-site combinations, a lower basal level, and hence lower fold-repression effect, is observed.

In Figure 7D, we present the dose-response heatmaps obtained for QCP. Here, we used the same tandem variants as for MCP, due to the binding cross-talk between both proteins shown in Figure 1B. Notably, the dose responses for these tandems in the presence of QCP vary substantially as compared with that observed for MCP. While the site MS2-U(-5)G is still associated with higher basal expression when positioned in the ribosomal initiation region, only three variants (as compared with five for MCP) do not seem to respond to QCP. In particular, two variants, each containing MS2-U(-5)G in the ribosomal initiation region and MS2-U(-5)C in the 5′ UTR, exhibit a 2-fold upregulatory dose response, as compared with a strong down-regulatory effect for MCP. Given the propensity of binding sites in the ribosomal initiation region to generate a strong repression effect (Katz et al., 2018), the upregulatory effect observed here is consistent with a lack of binding of QCP to MS2-U(-5)G in the ribosomal initiation region (as was observed before), thus

facilitating the upregulation effect that was observed previously for MCP with MS2-U(-5)C in the single-binding-site strains.

Finally, we measured the effective cooperativity factor w (see Figure S5 for fitting model) for repressive tandem constructs in the presence of their corresponding cognate RBPs. In Figure 7E, we plot a sample fit for a MS2-U(-5)C/MS2-U(-5)G tandem in the presence of MCP. The data show that when taking into account the known $K_{RBP\ value}$s that were extracted for the single-binding-site variants, a fit with no cooperativity ($w = 1$) does not explain the data well (red line). However, when the cooperativity parameter is not fixed, a good description for the data is obtained for $50 < w < 80$ (best fit at $w = 73$). In Figure 7F, we plot the extracted cooperativity parameter for each of the 16 tandems displaying a regulatory response with calculated $K_{RBP\ value}$s for both sites (see Figure S5 and Table S5 for fits and parameter values, respectively). Altogether, at least 6 of the 16 tandems exhibited strong cooperative behavior. For MCP and QCP, five of the six relevant tandems displayed strong cooperativity ($w > 25$). For PCP, only two of the ten tandems displayed weak cooperativity ($1 < w < 25$). These tandems had less than 30 nt between the two PCP-binding sites.

The cooperative behavior, which reflects overall increase in affinity of the RBP to the molecule when there is more than one binding site present, may also indicate increased stability of the hairpin structures. An increased stability can explain two additional features of the tandems that were not observed for the single-binding-site constructs: the QCP upregulatory response observed for the MS2-U(-5)C/MS2-U(-5)G tandem and the decreased basal mCherry rate of production levels. Overall, the K_{RBP} of the tandem- and single-binding-site constructs together with the RBPs can be varied over a range of specificities that spans approximately an order of magnitude, depending on the chosen 5′ UTR and gene-header sequences.

DISCUSSION

In recent years, synthetic biology approaches have been increasingly used to map potential regulatory mechanisms of transcriptional and translational regulation in both eukaryotic and bacterial cells (Kinney et al., 2010; Sharon et al., 2012; Dvir et al., 2013; Weingarten-Gabbay et al., 2016; Peterman and Levine, 2016; Levy et al., 2017). Here, we built on the design introduced by Saito et al. (2010) to explore the regulatory potential of RBP-RNA interactions in bacterial 5′ UTRs, using a synthetic biology approach combined with the SHAPE-seq method. Using a library of RNA variants, we found a complex set of regulatory responses, including translational repression, translational stimulation, and cooperative behavior. The upregulation phenomenon, or translational stimulation, had been reported only once for a single natural example in bacteria yet was mimicked here by all four RBPs at multiple 5′ UTR positions.

Our expression level data on the single-binding-site constructs hint that the mechanism that drives the complexity observed can be described by a three-state system. Using both the SHAPE-seq experiment and the reporter assay, we found a translationally active and weakly structured 5′ UTR state, a translationally inactive and highly structured 5′ UTR state, and an RBP-bound state with partial translation capacity. As a result, the same RBP can either upregulate or downregulate expression, depending on 5′ UTR sequence context. This description deviates from the classic two-state regulatory model, which is often used as a theoretical basis for describing transcriptional and post-transcriptional regulation (Bintu et al., 2005). In a two-state model, a substrate can either be bound or not bound by a ligand, leading to either an active or inactive regulatory state. This implies that in the two-state scenario, a bound protein cannot be both an ''activator'' and a ''repressor'' without an additional interaction or constraint that alters the system.

The appearance of two distinct mRNA states in the non-induced case *in vivo*, as compared with only one *in vitro*, suggests that *in vivo* the mRNA molecules can fold into one of two distinct phases: a molten phase that is amenable to translation and a structured phase that inhibits translation. A previous theoretical study by Schwab and Bruinsma (SB) (Schwab and Bruinsma, 2009) showed that a first-order phase transition separating a molten and a structured phase for mRNA can occur if a strong attractive interaction between the non-base-paired segments of the molecule exists within the system (see Figure S6). Such an interaction destabilizes the base pairing of branched structures and, if sufficiently strong, leads to complete melting of the molecule into a non-structured form. It is possible that such attractive interaction between non-base-paired segments is mediated by the ribosome, which is known to destabilize base-paired structures during translation.

Furthermore, the RBP-bound states, which yielded indistinguishable *in vivo* SHAPE-seq data together with a convergence of the induced up- and downregulating expression distributions, are also consistent with the SB model. In this case, the SB phase diagram (see Figure S6) shows that a weaker attractive interaction does not yield a first-order phase transition but rather a continuous transition from a fully structured phase through a partially structured phase to the fully molten state. Since the bound RBP stabilizes the hairpin structure, counteracting the destabilizing effect of the ribosome, in the context of the SB model, this effect may lead to a reduction in the strength of the ''attractive'' interaction. Therefore, it is possible that this binding event shifts the RNA molecules into the portion of the phase diagram (see Figure S6, bottom) in which the partially folded state minimizes the free energy, leading to the observed expression level and reactivity measurements in the induced phase.

Our work presents an important step in understanding and engineering post-transcriptional regulatory networks. Throughout this paper, we attempted to increase the synthetic biology utility of our work, the highlight being the direct activation of translation via a single RBP-binding-site pair. As a result, our synthetic regulatory modules can be viewed as a new class of ''protein-sensing riboswitches,'' which, given the hypothesized phase-based characterization, may ultimately have a wide utility in gene-regulatory applications. Together with our previous work of positioning the sites in the ribosomal initiation region (Katz et al., 2018), we offer a set of modestly upregulating and a range of downregulating RBP-binding-site pairs with tunable affinities for four RBPs, three of which are orthogonal to each other (PCP, GCP, and QCP). While we emphasize that our results were obtained in *E. coli*, given the propensity of RBPs to alter the RNA structure via direct interaction, it is tempting to

speculate that such an interaction may be a generic 5′ UTR mechanism that could be extended to other RBPs and other organisms.

How difficult is it to design an upregulatory dose response for an RBP *de novo*? Unfortunately, our data do not provide a satisfactory mechanistic outcome for a quantitative prediction but a qualitative phase-based description, which is an initial step. Our experiments revealed no particular structural features that were associated with this regulatory switch, such as the release of a sequestered RBS, which has been reported before as a natural mechanism for translational stimulation (Hattman et al., 1991; Wulczyn and Kahmann, 1991). Moreover, attempting to allocate a structural state for a certain sequence *in vivo* using *in-silico*-RNA-structure-prediction tools is not a reliable approach because of mechanistic differences between the *in vivo* and *in vitro* environment, which these models understandably do not take into account. Therefore, to provide a predictive blueprint for which sequences are likely to be translationally inactive in their native RBP unbound state, a better understanding of both RNA dynamics and the interaction of RNA with the translational machinery *in vivo* needs to be established. Yet, our findings suggest that generating translational stimulation using RBPs may not be as difficult as previously thought. At present, the best approach to designing functional elements is to first characterize experimentally a small library of a variety of designs and subsequently select and optimize the variants that exhibit interesting functionality. Finally, the described constructs add to the growing toolkit of translational regulatory parts and provide a working design for further exploration of both natural and synthetic post-transcriptional gene-regulatory networks.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cels.2019.04.007.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Technion's LS&E staff (Tal Katz-Ezov and Anastasia Diviatis) for help with sequencing. This project received funding from the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation (grant no. 152/11), Marie Curie Reintegration grant no. PCIG11-GA-2012-321675, and the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 664918—MRG-Grammar.

AUTHOR CONTRIBUTIONS

N.K. designed and carried out the expression-level experiments and analysis for all constructs. N.K. also conducted several of the SHAPE-seq experiments with R.C. R.C. and B.K. designed and carried out the SHAPE-seq experiments. O.S .and Z.Y. helped analyze the SHAPE-seq data. S.G. and O.A. assisted and guided the experiments and analysis. R.A. supervised the study. N.K., R.A., S.G., and B.K. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 21, 2018
Revised: February 7, 2019
Accepted: April 26, 2019
Published: May 22, 2019

REFERENCES

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

Aviran, S., Trapnell, C., Lucks, J.B., Mortimer, S.A., Luo, S., Schroth, G.P., Doudna, J.A., Arkin, A.P., and Pachter, L. (2011). Modeling and automation of sequencing-based characterization of RNA structure. Proc. Natl. Acad. Sci. USA *108*, 11069–11074.

Bintu, L., Buchler, N.E., Garcia, H.G., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005). Transcriptional regulation by the numbers: models. Curr. Opin. Genet. Dev. *15*, 116–124.

Boutonnet, C., Boijoux, O., Bernat, S., Kharrat, A., Favre, G., Faye, J.C., and Vagner, S. (2004). Pharmacological-based translational induction of transgene expression in mammalian cells. EMBO Rep. *5*, 721–727.

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

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

Cerretti, D.P., Wignall, J., Anderson, D., Tushinski, R.J., Gallis, B.M., Stya, M., Gillis, S., Urdal, D.L., and Cosman, D. (1988). Human macrophage-colony stimulating factor: alternative RNA and protein processing from a single gene. Mol. Immunol. *25*, 761–770.

Chen, A.H., and Silver, P.A. (2012). Designing biological compartmentalization. Trends Cell Biol. *22*, 662–670.

Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: interactive drawing and editing of the RNA secondary structure. Bioinformatics *25*, 1974–1975.

De Gregorio, E., Preiss, T., and Hentze, M.W. (1999). Translation driven by an eIF4G core domain in vivo. EMBO J. *18*, 4865–4874.

Deigan, K.E., Li, T.W., Mathews, D.H., and Weeks, K.M. (2009). Accurate SHAPE-directed RNA structure determination. Proc. Natl. Acad. Sci. USA *106*, 97–102.

Delebecque, C.J., Lindner, A.B., Silver, P.A., and Aldaye, F.A. (2011). Organization of intracellular reactions with rationally designed RNA assemblies. Science *333*, 470–474.

Desai, S.K., and Gallivan, J.P. (2004). Genetic screens and selections for small molecules based on a synthetic riboswitch that activates protein translation. J. Am. Chem. Soc. *126*, 13247–13254.

Dinman, J.D. (2005). 5S rRNA: structure and function from head to toe. Int. J. Biomed. Sci. *1*, 2–7.

Dvir, S., Velten, L., Sharon, E., Zeevi, D., Carey, L.B., Weinberger, A., and Segal, E. (2013). Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. Proc. Natl. Acad. Sci. USA *110*, E2792–E2801.

Flynn, R.A., Zhang, Q.C., Spitale, R.C., Lee, B., Mumbach, M.R., and Chang, H.Y. (2016). Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. Nat. Protoc. *11*, 273–290.

Goldfless, S.J., Belmont, B.J., de Paz, A.M., Liu, J.F., and Niles, J.C. (2012). Direct and specific chemical control of eukaryotic translation with a synthetic RNA-protein interaction. Nucleic Acids Res. *40*, e64.

Gott, J.M., Wilhelm, L.J., and Uhlenbeck, O.C. (1991). RNA binding properties of the coat protein from bacteriophage GA. Nucleic Acids Res. *19*, 6499–6503.

Green, A.A., Kim, J., Ma, D., Silver, P.A., Collins, J.J., and Yin, P. (2017). Complex cellular logic computation using ribocomputing devices. Nature *548*, 117–121.

Green, A.A., Silver, P.A., Collins, J.J., and Yin, P. (2014). Toehold switches: de-novo-designed regulators of gene expression. Cell *159*, 925–939.

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

Hattman, S., Newman, L., Murthy, H.M., and Nagaraja, V. (1991). Com, the phage Mu mom translational activator, is a zinc-binding protein that binds specifically to its cognate mRNA. Proc. Natl. Acad. Sci. USA *88*, 10027–10031.

Henkin, T.M. (2008). Riboswitch RNAs: using RNA to sense cellular metabolism. Genes Dev. *22*, 3383–3390.

Hentze, M.W., Caughman, S.W., Rouault, T.A., Barriocanal, J.G., Dancis, A., Harford, J.B., and Klausner, R.D. (1987). Identification of the iron-responsive element for the translational regulation of human ferritin mRNA. Science *238*, 1570–1573.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatsh. Chem. *125*, 167–188.

Hutvágner, G., and Zamore, P.D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. Science *297*, 2056–2060.

Isaacs, F.J., Dwyer, D.J., and Collins, J.J. (2006). RNA synthetic biology. Nat. Biotechnol. *24*, 545–554.

Katz, N., Cohen, R., Solomon, O., Kaufmann, B., Atar, O., Yakhini, Z., Goldberg, S., and Amit, R. (2018). An in vivo binding assay for RNA-binding proteins based on repression of a reporter gene. ACS Synth. Biol. *7*, 2765–2774.

Keren, L., Zackay, O., Lotan-Pompan, M., Barenholz, U., Dekel, E., Sasson, V., Aidelberg, G., Bren, A., Zeevi, D., Weinberger, A., et al. (2013). Promoters maintain their relative activity levels under different growth conditions. Mol. Syst. Biol. *9*, 701.

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

Khalil, A.S., and Collins, J.J. (2010). Synthetic biology: applications come of age. Nat. Rev. Genet. *11*, 367–379.

Kinney, J.B., Murugan, A., Callan, C.G., and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. Proc. Natl. Acad. Sci. USA *107*, 9158–9163.

Levy, L., Anavy, L., Solomon, O., Cohen, R., Brunwasser-Meirom, M., Ohayon, S., Atar, O., Goldberg, S., Yakhini, Z., and Amit, R. (2017). A synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial σ54Promoters. Cell Rep. *21*, 845–858.

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

Lim, F., and Peabody, D.S. (2002). RNA recognition site of PP7 coat protein. Nucleic Acids Res. *30*, 4138–4144.

Lim, F., Spingola, M., and Peabody, D.S. (1996). The RNA-Binding Site of bacteriophage Qβ coat protein. J. Biol. Chem. *271*, 31839–31845.

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algor. Mol. Biol. *6*, 26.

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

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. Journal *17*, 10–12.

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

Ouyang, Z., Snyder, M.P., and Chang, H.Y. (2013). SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. Genome Res. *23*, 377–387.

Pardee, K., Green, A.A., Takahashi, M.K., Braff, D., Lambert, G., Lee, J.W., Ferrante, T., Ma, D., Donghia, N., Fan, M., et al. (2016). Rapid, low-cost detection of Zika virus using programmable biomolecular components. Cell *165*, 1255–1266.

Peabody, D.S. (1993). The RNA binding site of bacteriophage MS2 coat protein. EMBO J. *12*, 595–600.

Peterman, N., and Levine, E. (2016). Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. BMC Genomics *17*, 206.

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

Romaniuk, P.J., Lowary, P., Wu, H.N., Stormo, G., and Uhlenbeck, O.C. (1987). RNA binding site of R17 coat protein. Biochemistry *26*, 1563–1568.

Sacerdot, C., Caillet, J., Graffe, M., Eyermann, F., Ehresmann, B., Ehresmann, C., Springer, M., and Romby, P. (1998). 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. *29*, 1077–1090.

Sachdeva, G., Garg, A., Godding, D., Way, J.C., and Silver, P.A. (2014). In vivo co-localization of enzymes on RNA scaffolds increases metabolic production in a geometrically dependent manner. Nucleic Acids Res. *42*, 9493–9503.

Saito, H., Kobayashi, T., Hara, T., Fujita, Y., Hayashi, K., Furushima, R., and Inoue, T. (2010). Synthetic translational regulation by an L7Ae–kink-turn RNP switch. Nat. Chem. Biol. *6*, 71–78.

Schlax, P.J., Xavier, K.A., Gluick, T.C., and Draper, D.E. (2001). Translational repression of the Escherichia coli alpha operon mRNA: importance of an mRNA conformational switch and a ternary entrapment complex. J. Biol. Chem. *276*, 38494–38501.

Schwab, D., and Bruinsma, R.F. (2009). Flory theory of the folding of designed RNA molecules. J. Phys. Chem. B *113*, 3880–3893.

Serganov, A., and Nudler, E. (2013). A decade of riboswitches. Cell *152*, 17–24.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory

logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

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

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

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

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

Szymanski, M., Barciszewska, M.Z., Erdmann, V.A., and Barciszewski, J. (2002). 5S ribosomal RNA database. Nucleic Acids Res. *30*, 176–178.

Villa, E., Sengupta, J., Trabuco, L.G., LeBarron, J., Baxter, W.T., Shaikh, T.R., Grassucci, R.A., Nissen, P., Ehrenberg, M., Schulten, K., et al. (2009). Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis. Proc. Natl. Acad. Sci. USA *106*, 1063–1068.

Washietl, S., Hofacker, I.L., Stadler, P.F., and Kellis, M. (2012). RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. Nucleic Acids Res. *40*, 4261–4272.

Watters, K.E., Abbott, T.R., and Lucks, J.B. (2016). Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. Nucleic Acids Res. *44*, e12.

Weingarten-Gabbay, S., Elias-Kirma, S., Nir, R., Gritsenko, A.A., Stern-Ginossar, N., Yakhini, Z., Weinberger, A., and Segal, E. (2016). Comparative genetics. Systematic discovery of cap-independent translation sequences in human and viral genomes. Science *351*, aad4939.

Werstuck, G., and Green, M.R. (1998). Controlling gene expression in living cells through small molecule-RNA interactions. Science *282*, 296–298.

Win, M.N., and Smolke, C.D. (2008). Higher-order cellular information processing with synthetic RNA devices. Science *322*, 456–460.

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

Wittmann, A., and Suess, B. (2012). Engineered riboswitches: expanding researchers' toolbox with synthetic RNA regulators. FEBS Lett. *586*, 2076–2083.

Wroblewska, L., Kitada, T., Endo, K., Siciliano, V., Stillo, B., Saito, H., and Weiss, R. (2015). Mammalian synthetic circuits with RNA binding proteins for RNA-only delivery. Nat. Biotechnol. *33*, 839–841.

Wulczyn, F.G., and Kahmann, R. (1991). Translational stimulation: RNA sequence and structure requirements for binding of Com protein. Cell *65*, 259–269.

Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R., and Benenson, Y. (2011). Multi-input RNAi-based logic circuit for identification of specific cancer cells. Science *333*, 1307–1311.

Zarringhalam, K., Meyer, M.M., Dotu, I., Chuang, J.H., and Clote, P. (2012). Integrating chemical footprinting data into RNA secondary structure prediction. PLoS ONE *7*, e45160.

Zeevi, D., Sharon, E., Lotan-Pompan, M., Lubling, Y., Shipony, Z., Raveh-Sadka, T., Keren, L., Levo, M., Weinberger, A., and Segal, E. (2011). Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. Genome Res. *21*, 2114–2128.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
E. coli TOP10 cells	Invitrogen	C404006
Chemicals, Peptides, and Recombinant Proteins		
EagI-HF	NEB	R3505
KpnI	NEB	R0142
ApaLI	NEB	R0507
ligase	NEB	B0202S
Ampicillin sodium salt	SIGMA	A9518
Kanamycin solfate	SIGMA	K4000
Tryptone	BD	211705
glycerol	BIO LAB	071205
SODIUM CHLORIDE (NaCL)	BIO LAB	190305
MAGNESIUM SULFATE (MgSO4)	ALFA AESAR	33337
PBS buffer	Biological Industries	020235A
N-butanoyl-L-homoserine lactone (C4-HSL)	cayman	K40982552 019
2-methylnicotinic acid imidazole (NAI)	Millipore (Merck)	03-310
DMSO	Sigma Aldrich (Merck)	D8418
Max Bacterial Enhancement Reagent	Life Technologies	16122012
TRIzol	Life Technologies	466036
RiboLock RNAse inhibitor	Thermo Fisher Scientific	E00382
Superscript III reverse transcriptase	Thermo Fisher Scientific	18080044
CircLigase	Epicentre	CL4115K
glycogen	Invitrogen	R0561
Agencourt AMPure XP beads	Beackman Coulter	A63881
DynaMag-96 Side Magnet	Thermo Fisher Scientific	12331D
ExoI	NEB	M0293
Q5 HotStart Polymerase	NEB	M0493
Critical Commercial Assays		
RNeasy mini kit	QIAGEN	74104
TapeStation 2200 DNA ScreenTape assay	Agilent	N/A
Qubit fluorimeter	Thermo Fisher Scientific	N/A
HiSeq 2500 sequencing system	Illumina	N/A
Deposited Data		
SHAPE-seq sequencing data	This paper	Table S4 and GEO ID: GSE129163, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129163
Oligonucleotides		
SHAPE-seq primers and adapters	Watters et al., 2016	IDT, Table S3
Recombinant DNA sequence verification primer: acggaactcttgtgcgtaag	This study	IDT
Recombinant DNA		
Constructs with a single binding site	This study	Gen9, Table S1
RBP constructs: PP7	Wu et al	Addgene: #40650, Table S2
RBP constructs: MS2	Fusco et al	Addgene: #27121, Table S2
RBP constructs: Qbeta	NCBI #NC_001890.1	Genescript, Table S2

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
RBP constructs: GA	NCBI #NC_001426.1	IDT, Table S2
Constructs with tandem binding sites	This study	Table S5
Software and Algorithms		
Matlab analysis software	Mathworks	N/A
RNAfold WebServer	Institute for Theoretical Chemistry, University of Vienna	N/A
RNApvmin 2.4.9 WebServer	Theoretical Bioechmisty Group, Institute for Theoretical Chemistry, Univerisyt of Vienna	N/A
Other		
96-well plates	PerkinElmer	6005029
Liquid-handling robotic system	TECAN	EVO 100, MCA 96-channel
incubator	TECAN	liconic incubator
platereader	TECAN	Infinite F200 PRO

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents and resources should be directed to and will be fulfilled by the Lead Contact, Roee Amit (roeeamit@technion.ac.il).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

E. coli TOP10 cells were obtained from Invitrogen, cat number C404006 (see also Key Resources Table). Cells were grown in Laural Broth (LB) with appropriate antibiotics overnight at 37°C and 250 rpm. In the morning, they were diluted by a factor of 100 to semi-poor medium (SPM) consisting of 95% bio-assay (BA) and 5% LB with appropriate antibiotics and different inducer concentrations at 37°C and 250 rpm for 1hr to 4hrs (Method Details section for more details).

METHOD DETAILS

Design and Construction of Binding-Site Plasmids

Binding-site cassettes (see Table S1) were ordered as double-stranded DNA minigenes from either Gen9 or Twist Bioscience. Each minigene was ~500 bp long and contained the following parts: EagI restriction site, ~40 bases of the 5' end of the Kanamycin (Kan) resistance gene, pLac-Ara constitutive promoter, ribosome-binding site (RBS), and a KpnI restriction site. In addition, each cassette contained one or two wild-type or mutated RBP binding sites, either upstream or downstream to the RBS (see Table S1), at varying distances. All binding sites were derived from the wild-type binding sites of the coat proteins of one of the four bacteriophages GA, MS2, PP7, and Qβ. For insertion into the binding-site plasmid backbone, minigene cassettes were double-digested with EagI-HF and either KpnI or ApaLI (New England Biolabs [NEB]). The digested minigenes were then cloned into the binding-site backbone containing the rest of the mCherry gene, terminator, and the remainder of the Kanamycin resistance gene, by ligation and transformation into *E. coli* TOP10 cells (ThermoFisher Scientific). All the plasmids were sequence-verified by Sanger sequencing. Purified plasmids were stored in 96-well format, for transformation into *E. coli* TOP10 cells containing one of the four fusion-RBP plasmids (see below).

Design and Construction of Fusion-RBP Plasmids

RBP sequences lacking a stop codon were amplified via PCR off either Addgene or custom-ordered templates (Genescript or IDT, see Table S2). All RBPs presented (GCP, MCP, PCP, and QCP) were cloned into the RBP plasmid between restriction sites KpnI and AgeI, immediately upstream of an mCerulean gene lacking a start codon, under the so-called RhlR promoter [containing the *rhlAB* las box (Medina et al., 2003)] and induced by N-butyryl-L-homoserine lactone (C_4-HSL). The backbone contained an Ampicillin (Amp) resistance gene. The resulting fusion-RBP plasmids were transformed into *E. coli* TOP10 cells. After Sanger sequencing, positive transformants were made chemically-competent and stored at -80°C in 96-well format.

Transformation of Binding-Site Plasmids

Binding-site plasmids stored in 96-well format were simultaneously transformed into chemically-competent bacterial cells containing one of the fusion plasmids, also prepared in 96-well format. After transformation, cells were plated using an 8-channel pipettor on 8-lane plates containing LB-agar with relevant antibiotics (Kan and Amp). Double transformants were selected, grown overnight, and stored as glycerol stocks at -80°C in 96-well plates.

Single Clone Expression Level Assay

Dose-response fluorescence experiments were performed using a liquid-handling system in combination with a Liconic incubator and a TECAN Infinite F200 PRO platereader. Each measurement was carried out in duplicates. Double-transformant strains were grown at 37°C and 250 rpm shaking in 1.5 ml LB in 48-well plates with appropriate antibiotics (Kan and Amp) over a period of 16 hours (overnight). In the morning, the inducer for the rhlR promoter C_4-HSL was pipetted manually to 4 wells in an inducer plate, and then diluted by the robot into 24 concentrations ranging from 0 to 218 nM. While the inducer dilutions were being prepared, semi-poor medium consisting of 95% bioassay buffer (for 1 L: 0.5 g Tryptone [Bacto], 0.3 ml Glycerol, 5.8 g NaCl, 50 ml 1M MgSO4, 1ml 10xPBS buffer pH 7.4, 950 ml DDW) and 5% LB was heated in the incubator, in 96-well plates. The overnight strains were then diluted by the liquid-handling robot by a factor of 100 into 200 μL of pre-heated semi-poor medium, in 96-well plates suitable for fluorescent measurement. The diluted inducer was then transferred by the robot from the inducer plate to the 96-well plates containing the strains. The plates were shaken at 37°C for 6 hours. Note, that induction was only used for the rhlR promoter, which controls the expression of the RBP-mCerulean fusion. The pLac/Ara promoter controlling the mCherry reporter gene functioned as a constitutive promoter of suitable strength in our strains and did not require IPTG or Arabinose induction.

Measurement of OD, and mCherry and mCerulean fluorescence were taken via a platereader every 30 minutes. Blank measurements (growth medium only) were subtracted from all fluorescence measurements. For each day of experiment (16 different strains), a time interval of logarithmic growth was chosen (T_0 to T_{final}) according to the measured growth curves, between the linear growth phase and the stationary (T_0 is typically the third measured time point). Six to eight time points were taken into account, discarding the first and last measurements to avoid errors derived from inaccuracy of exponential growth detection. Strains that showed abnormal growth curves or strains where logarithmic growth phase could not be detected, were not taken into account and the experiment was repeated. See Figure S2 for experimental schematic and a sample data set.

SHAPE-Seq Experimental Setup

LB medium supplemented with appropriate concentrations of Amp and Kan was inoculated with glycerol stocks of bacterial strains harboring both the binding-site plasmid and the RBP-fusion plasmid and grown at 37°C for 16 hours while shaking at 250 rpm. Overnight cultures were diluted 1:100 into SPM. 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 SPM. 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) (Spitale et al., 2013), 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 (both from 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 $\delta = -29$ and PP7-USs $\delta = -29$, 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 (+C4-HSL/+NAI)
2. non-induced/modified (-C4-HSL/+NAI)
3. induced/non-modified (+C4-HSL/+DMSO)
4. non-induced/non-modified (-C4-HSL/+DMSO).

In vitro modification was carried out on DMSO-treated samples (3 and 4) and has been described elsewhere (Flynn et al., 2016). 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 (Watters et al., 2016), including reverse transcription (steps 40-51), adapter ligation and purification (steps 52-57) as well as dsDNA sequencing library preparation (steps 68-76). 1000 ng of RNA were converted to cDNA using the reverse transcription primers (for details of primer and adapter sequences used in this work see Table S3) for mCherry (#1) or 5S rRNA (#2) that are specific for either the mCherry transcripts (PP7-USs δ=-29, PP7-wt δ=-29). The RNA was mixed with 0.5 μM primer (#1) or (#2) 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 minutes, 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 (#3) (see Table S3) was ligated to the cDNA using a CircLigase 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 (#3), 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.

SHAPE-Seq Library Preparation and Sequencing

To produce the dsDNA for sequencing 10ul of purified cDNA from the SHAPE procedure (see above) were PCR amplified using 3 primers: 4nM mCherry selection (#4) or 5S rRNA selection primer (#5), 0.5μM TruSeq Universal Adapter (#6) and 0.5μM TrueSeq Illumina indexes (one of #7-26) (Table S3) with PCR reaction mix (1x Q5 HotStart reaction buffer, 0.1 mM dNTPs, 1 U Q5 HotStart Polymerase [NEB]). A 15-cycle PCR program was used: initial denaturation at 98°C for 30 s followed by a denaturation step at 98°C for 15 s, primer annealing at 65°C for 30 s and extension at 72°C for 30 s, 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. See Table S4 for read counts for all experiments presented in the manuscript.

QUANTIFICATION AND STATISTICAL ANALYSIS

Single Clone Expression Level Analysis

The average normalized fluorescence of mCerulean, and rate of production of mCherry, were calculated for each inducer concentration using the routine developed in (Keren et al., 2013), as follows:

mCerulean average normalized fluorescence: for each inducer concentration, mCerulean measurements were normalized by OD. Normalized measurements were then averaged over the N logarithmic-growth timepoints in the interval [T0, Tfinal], yielding:

$$mCerulean = \frac{1}{N}\sum_{t=T_0}^{T_{final}} \frac{mCerulean(t)}{OD(t)} \quad \text{(Equation 1)}$$

mCherry rate of production: for each inducer concentration, mCherry fluorescence at T_0 was subtracted from mCherry fluorescence at T_{final}, and the result was divided by the integral of OD during the logarithmic growth phase:

$$mCherry\ \ rate\ \ of\ \ production = \frac{mCherry(T_{final}) - mCherry(T_0)}{\int_{T_0}^{T_{final}} dtOD(t)} \quad \text{(Equation 2)}$$

Finally, we plotted mCherry rate of production [(Zeevi et al., 2011)] as a function of averaged normalized mCerulean expression, creating dose response curves as a function of RBP-mCerulean fluorescence. Our choice for computing rate of production for mCherry stems from our belief that this observable best quantifies the regulatory effect, which is a function of the absolute number of inducer protein present (i.e RBP-mCerulean) at a any given moment in time. Data points with higher than two standard deviations calculated over mCerulean and mCherry fluorescence at all the inducer concentrations of the same strain) between the two duplicates were not taken into account and plots with 25% or higher of such points were discarded and the experiment repeated.

Dose Response Fitting Routine and K_d Extraction

Final data analysis and fit were carried out on plots of rate of mCherry production as a function of averaged normalized mCerulean fluorescence at each inducer concentration. Such plots represent production of the reporter gene as a function of RBP presence in the cell. The fitting analysis and K_d extraction were based on the following two-state thermodynamic model:

$$mCherry\ \ rate\ \ of\ \ production = P_{bound}k_{bound} + P_{unbound}k_{unbound} \quad \text{(Equation 3)}$$

Here, the mCherry mRNA is either bound to the RBP or unbound, with probabilities P_{bound} and $P_{unbound}$ and ribosomal translation rates k_{bound} and $k_{unbound}$, respectively. The probabilities of the two states are given by:

$$P_{bound} = \frac{([x]/K_d)^n}{1 + ([x]/K_d)^n} \quad \text{(Equation 4)}$$

and

$$P_{unbound} = \frac{1}{1 + ([x]/K_d)^n} \quad \text{(Equation 5)}$$

where *[x]* is RBP concentration, K_d is an effective dissociation constant, and n is a constant that quantifies RBP cooperativity; it represents the number of RBPs that need to bind the binding site simultaneously for the regulatory effect to take place. Substituting the probabilities into Equation 3 gives:

$$mCherry\ \ rate\ \ of\ \ production = \frac{([x]/K_d)^n}{1 + ([x]/K_d)^n} k_{bound} + \frac{1}{1 + ([x]/K_d)^n} k_{unbound} \quad \text{(Equation 6)}$$

For the case in which we observe a down-regulatory effect, we have significantly less translation for high [*x*], which implies that $k_{bound} \ll k_{unbound}$ and that we may neglect the contribution of the bound state to translation. For the case in which we observe an up-regulatory affect for large [*x*], we have $k_{bound} \gg k_{unbound}$, and we neglect the contribution of the unbound state.

The final models used for fitting the two cases are summarized as follows:

$$mCherry\ \ rate\ \ of\ \ production \simeq \begin{cases} \frac{k_{unbound}}{1 + ([x]/K_d)^n} + C & downregulatory\ \ effect \\ \frac{([x]/K_d)^n k_{bound}}{1 + ([x]/K_d)^n} + C & upregulatory\ \ effect \end{cases} \quad \text{(Equation 7)}$$

where *C* is the fluorescence baseline. Only fit results with $R^2 > 0.6$ were taken into account. For those fits, K_d error was typically in the range of 0.5-20%, for a 0.67 confidence interval.

SHAPE-Seq Initial Reactivity Analysis

Illumina reads were first adapter-trimmed using cutadapt (Martin, 2011) and were aligned against a composite reference built from mCherry, E. coli 5S rRNA sequences, and PhiX genome (PhiX is used as a control sequence in Illumina sequencing). Alignment was performed using bowtie2 [4] in local alignment mode (bowtie2 –local).

Reverse transcriptase (RT) drop-out positions were indicated by the end position of Illumina Read 2 (the second read on the same fragment). Drop-out positions were identified using an inhouse Perl script (can be provided upon request). 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.

To facilitate proper signal comparison, all libraries (16 total - including biological duplicates) 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 Dj(x) according to:

$$\widehat{D}_j^0(x) = \frac{D_j^0(x)}{\sum_{i=1}^{L} D_j^0(x)} \quad \text{(Equation 8)}$$

where *L* is the length of the sequence under investigation after RT primer removal. The reads as a function of position from the transcription start site (TSS) are supplied in Table S4.

SHAPE-Seq Bootstrap Analysis

To compute the mean read-ratio, reactivity, and associated error bars, we employed boot-strap statistics in a classic sense. Given *M* drop out reads per library, we first constructed a vector of length M, containing the index of the read # (1...M) and an associated 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 }.

SHAPE-Seq Signal-to-Noise (Read-Ratio) Computation

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

$$R^k_s(x) = \frac{\widehat{D}^k_{s,\mathrm{mod}}(x)}{\widehat{D}^k_{s,non-\mathrm{mod}}(x)} \quad \text{(Equation 9)}$$

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:

$$\langle R_s(x)\rangle = \frac{1}{100}\sum_{k=0}^{100}\frac{\widehat{D}^k_{s,\mathrm{mod}}(x)}{\widehat{D}^k_{s,non-\mathrm{mod}}(x)}, \quad \text{(Equation 10)}$$

$$\sigma_s(x) = \langle R_s(x)\rangle - \langle R_s(x)\rangle^2. \quad \text{(Equation 11)}$$

SHAPE-Seq Reactivity Computation

The literature has several redundant definitions for reactivity, and no consensus on a precise formulation (Aviran et al., 2011; Lucks et al., 2011; Spitale et al., 2015) 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:

$$\rho^k_s(x) = \left(R^k_s(x) - 1\right)\Theta\left(R^k_s(x) - 1\right), \quad \text{(Equation 12)}$$

Where,

$$\Theta(x) = \left\{\begin{array}{ll} 0 & \text{if } x<0 \\ 1 & \text{if } x \geq 0 \end{array}\right\}. \quad \text{(Equation 13)}$$

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

$$\rho_s(x) = (\langle R_s(x)\rangle - 1)\Theta(\langle R_s(x)\rangle - 1). \quad \text{(Equation 14)}$$

For the running-average reactivity plots shown in Figure 3, we used the following procedure. First, we computed an average reactivity per position based on two boot–strapped mean reactivity scores that were obtained from the two biological replicates. We then computed a running average 10 nt window for every position δ.

SHAPE-Seq Reactivity Error Bar Computation

Error bars were computed in two steps. First, we computed the error-bar per nucleotide before running average as follows:

$$\sigma(x) = \frac{1}{N+1}\sqrt{\left(\sum_{i=1}^{N}\sigma^2_i(x) + \sigma^2_0(x)\right)}. \quad \text{(Equation 15)}$$

Where $\sigma_i(x)$ corresponds to the boot-strapped sigma computed for position x of technical repeat *i*, while $\sigma_0(x)$ is defined as the standard deviation at position *i* of the read ratio values for all *N* technical repeats. The error bar displayed for each position in the running average plot (Figures 3A and 3B) were computed as follows:

$$\tilde{\sigma}(x) = \frac{1}{10}\sqrt{\left(\sum_{i=-4}^{i=5}\sigma^2(x+i)\right)}. \quad \text{(Equation 16)}$$

SHAPE-Seq 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). To do so, we use the following formulation:

$$Z(\delta) = 1 - n\frac{\tilde{\sigma}_{-RBP}(x) + \tilde{\sigma}_{+RBP}(x)}{|\langle R_{-RBP}(x)\rangle - \langle R_{+RBP}(x)\rangle|}, \quad \text{(Equation 17)}$$

where *n* corresponds to the threshold of the number of σ′*s* that we want to use to claim a statistically significant difference between two values of the mean. For our analysis we used *n* = *3*. 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 (i.e. $\langle R_{-RBP}(\delta)\rangle - \langle R_{+RBP}(\delta)\rangle$) were considered to be protected and marked in a semi-transparent grey shading in Figures 3 and 4.

SHAPE-Seq Structural Visualization

For the structural visualization (as in Figure 3C), the mRNA SHAPE-Seq fragment of PP7-wt_d=- 29 construct was first folded *in silico* using RNAfold in default parameters. For visualization purposes, the RNAfold 2d structure prediction served as input for VARNA (Darty et al., 2009) and the SHAPE-Seq reactivity scores were used as colormap to overlay the reactivity on the predicted structure and to generate the structure image.

Using the Empirical SHAPE-Seq Data as Constraints for Structural Prediction

In order to predict more accurate structural schemes (Deigan et al., 2009; Ouyang et al., 2013; Washietl et al., 2012; Zarringhalam et al., 2012) we used the in vitro and in vivo SHAPE-Seq data as constraints to the computational structure prediction. This is done by taking the calculated reactivities of each sample, and computing a perturbation vector using RNApvmin of Vienna package (Lorenz et al., 2011) that minimizes the discrepancies between the predicted and empirically inferred pairing probabilities. Once the perturbation vector is obtained, we implement the Washietl algorithm (Washietl et al., 2012) in RNAfold to compute the inferred structure.

In order to calculate base-pairing probabilities for the structure determined by RNAfold with Washietl algorithm, the perturbation vector generated by RNApvmin is inserted as an additional input for RNAsubopt (-p 1000). A custom Perl script was used to calculate the resulted probability of pairing for each nucleotide based on the structural ensemble.

SHAPE-Seq 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 (Kertesz et al., 2010; Spitale et al., 2015; Watters et al., 2016). We analysed the SHAPE-Seq read count by computing the "reactivity" of each base corresponding to the propensity of that base to be modified by NAI. 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 S4 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 (RFAM id: RF00001, PDB id: 4V69) (Szymanski et al., 2002; Villa et al., 2009; Watters et al., 2016).

By contrast, the in vivo reactivity data (blue line) is less modified on average and especially in the 9 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 (Dinman, 2005). The reactivity scores obtained here for both the in vitro and in vivo samples (Figure S4B) are comparable to previously published 5S-rRNA reactivity analysis (Deigan et al., 2009; Szymanski et al., 2002; Watters et al., 2016).

Tandem Cooperativity Fit and Analysis

To estimate the degree of cooperativity in RBP binding to the tandem binding site, we used the following 4-state thermodynamic model:

$$Z = 1 + \frac{[RBP]}{K_{RBP1}} + \frac{[RBP]}{K_{RBP2}} + \left(\frac{[RBP]^2}{K_{RBP1}K_{RBP2}}\right)w, \qquad \text{(Equation 18)}$$

where K_{RBP1} and K_{RBP2} are the dissociation constants measured for the two single-binding-site variants, *[RBP]* is the concentration of the RNA binding proteins, and w is the cooperativity factor.

In a four state model, we assume four potential RNA occupancy states. No occupancy - receiving the relative weight 1. A state with single hairpin bound by an RBP receiving either the weight $[RBP]/k_{RBP1}$ or the weight $[RBP]/k_{RBP2}$ depending on whether the 5' UTR or gene-header states are occupied respectively.

Finally, for the state where both hairpins are occupied we have the generic weight $([RBP]^2/K_{RBP1}K_{RBP2})w$, which takes into account also a potential interaction between the two occupied states, which can be cooperative if $w > 1$ or anti-cooperative if $w < 1$. No interaction is the case where $w = 1$.

Next, we compute the relevant probabilities for translation for each weight. We know that when the ribosomal initiation region hairpin is occupied translation cannot proceed, however, some translation can result (albeit via a lower rate), when the 5' UTR hairpin is occupied. This leads to the following rate equation for protein translation:

$$\frac{d[P]}{dt} = k_{basal}[mRNA]\left(\frac{1}{1+Z=1+\frac{[RBP]}{K_{RBP1}}+\frac{[RBP]}{K_{RBP2}}+\left(\frac{[RBP]^2}{K_{RBP1}K_{RBP2}}\right)w,}\right) + k_{utr}[mRNA]\left(\frac{1}{1+Z=1+\frac{[RBP]}{K_{RBP1}}+\frac{[RBP]}{K_{RBP2}}+\left(\frac{[RBP]^2}{K_{RBP1}K_{RBP2}}\right)w,}\right) - \gamma[P], \quad \text{(Equation 19)}$$

where γ is the protein degradation rate.

When measuring rate of production and given the stability of mCherry, the degradation rate of mCherry is negligible over the 1-2 hr range of integration that was used in 2. Since we normalized the basal levels of mCherry rate of production, 20 is reduced to the following fitting formula for the data:

$$\text{Normalized mCherry rate of production} = \left(\frac{1}{1+Z=1+\frac{[RBP]}{K_{RBP1}}+\frac{[RBP]}{K_{RBP2}}+\left(\frac{[RBP]^2}{K_{RBP1}K_{RBP2}}\right)w,}\right) + \frac{k_{utr}}{k_{basal}}\left(\frac{1}{1+Z=1+\frac{[RBP]}{K_{RBP1}}+\frac{[RBP]}{K_{RBP2}}+\left(\frac{[RBP]^2}{K_{RBP1}K_{RBP2}}\right)w,}\right) - \gamma[P], \quad \text{(Equation 20)}$$

Finally, given our previous measurements for K_{RBP1} and K_{RBP2}, this formula reduces to a two parameter fit for w and k_{utr}/k_{basal}. See Figure S5 and Table S5 for the fits and associated fitting parameter details for 14 of 16 dose-response down-regulatory tandem data sets that were used in the analysis.

DATA AND SOFTWARE AVAILABILITY

The SHAPE-Seq read data is available in Table S4 and in GEO ID: GSE129163. Link: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129163.

Part 4: Unpublished Work

Overcoming the design, build, test (DBT) bottleneck for synthesis of nonrepetitive protein-RNA binding cassettes for RNA applications

Noa Katz[1], Eitamar Tripto[2], Sarah Goldberg[1], Orna Atar[1], Zohar Yakhini[3,4], Yaron Orenstein[5], and Roee Amit[1,6]

[1] Department of Biotechnology and Food Engineering, Technion - Israel Institute of Technology, Haifa 3200003, Israel.

[2] Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel.

[3] Department of Computer Science, Technion - Israel Institute of Technology, Haifa 3200003, Israel.

[4] School of Computer Science, Interdisciplinary Center, Herzliya 46150, Israel.

[5] School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 8410501, Israel.

[6] Russell Berrie Nanotechnology Institute, Technion - Israel Institute of Technology, Haifa 3200003, Israel.

In advanced review at *Nature Communications,* Link

Induction-based Sort-Seq (iSort-Seq)

We have recently shown that placing a hairpin in the ribosomal initiation region of bacteria can lead to a ~x10-100 fold repression effect when bound to an RNA-binding protein (RBP)[95,96]. The magnitude of the effect allowed us to adapt this *in vivo* binding assay to a high-throughput OL experiment. We designed 10,000 mutated versions of the single native binding sites to the phage CPs of PP7 (PCP), MS2 (MCP) and Qβ (QCP), and positioned each site at two positions within the ribosomal initiation region (Figure 4.1a). The library consists of three sub-libraries within the original library: binding sites that mostly resemble either the MS2-wt site, the PP7-wt site, or the Qβ-wt site (Figure 4.1a bottom). We introduced semi-random mutations, both structure-altering and structure-preserving, as well as deliberate mutations at positions which previous studies have shown to be crucial for binding. Additionally, we incorporated into our library several dozens of control variants. We used variants characterized in our previous study as positive and negative controls[95,97–99] as follows: positive controls are binding sites that exhibited a strong fold-repression response, and negative control variants are either random sequences or hairpins which did not exhibit a fold-repression response. For the complete library, see Table S1.

We incorporated each of the designed 10k single binding-site variants downstream to an mCherry start codon (Figure 4.1b) at each of the two positions (spacers □=C or □=GC) to ensure high basal expression and enable detection of a down-regulatory response, resulting in 20k different OL variants. Each variant was ordered with five different barcodes, resulting in a total of 100k different OL sequences.

The second component of our system included a fusion of one of the three phage CPs to green fluorescent protein (GFP) (Figure 4.1b) under the control of an inducible promoter. Thus, we created three libraries in *E. coli* cells; each with a different RBP but the same 100k binding site variants. In order to characterize the dose response of our variants, each library was first separated to six exponentially expanding cultures grown in the presence of one of six inducer concentration for RBP-GFP fusion induction. If the RBP was able to bind a particular variant, a strong fold-repression effect ensued, resulting in a reduced fluorescent expression profile (Figure 4.1c). We sorted each inducer-concentration culture into eight predefined fluorescence bins, which resulted in a 6x8 fluorescence matrix for each variant, corresponding to its dose-response behavior. We call this adaptation of Sort-Seq "induction Sort-Seq" (iSort-seq - for details see Methods). As an example, we present a high-affinity, down-regulatory dose-response for a positive variant (Figure 4.1d-bottom V1), and a no-affinity variant exhibiting no apparent regulatory effect as a function of induction (Figure 4.1d-bottom V2).

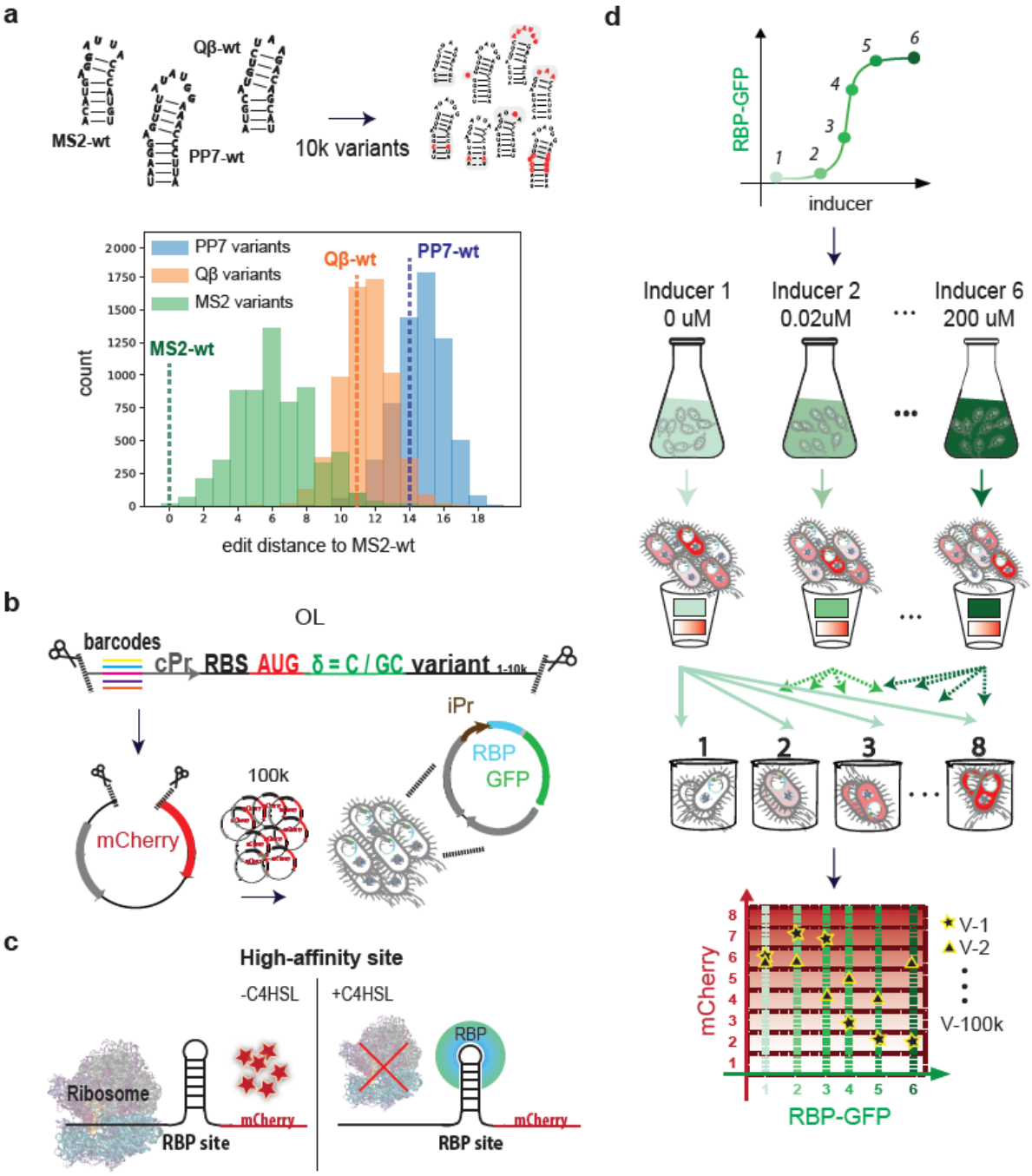


Figure 4.1: iSort-Seq overview in *E. coli*. (a) (Top) Wild-type binding sites for MS2, PP7 and Qβ phage coat proteins and illustrations of the 20k mutated variants created based on their sequences. (Bottom) Composition of the OL library. Histogram of the number of PP7-based variants (blue), Qβ-based variants (orange), and MS2-based variants (green) with different edit distances from the MS2-WT binding site. (b) Each putative binding site variant was encoded on a 210bp oligo containing the following components: restriction site, barcode, constitutive promoter (cPr), ribosome binding site (RBS), mCherry start codon, one or two bases (denoted by ⍰), the sequence of the variant tested, and the second restriction site. Each configuration was encoded with five different barcodes, resulting in a total of 100k different OL variants. The OL was then cloned into a vector and transformed into an *E. coli* strain expressing one of three RBP-GFP fusions under an inducible promoter (iPr). The transformation was repeated for all three fusion proteins. (c) The schema illustrates the behavior of a high-affinity strain: when no inducer is added, mCherry is expressed at a certain basal level that

depends on the mRNA structure and sequence. When inducer (C4-HSL) is added, the RBP binds the mRNA and blocks the ribosome from mCherry translation, resulting in a down-regulatory response as a function of inducer concentration. (d) The experimental flow for iSort-Seq. Each library is grown at 6 different inducer concentrations, and sorted into eight bins with varying mCherry levels and constant RBP-GFP levels. This yields a 6x8 matrix of mCherry levels for each variant at each induction level. (Bottom) An illustration of the experimental output of a high-affinity strain (V1) and a no-affinity strain (V2).

Calculating Binding Scores

We conducted preliminary analysis of the sequencing data to generate mCherry levels per RBP and inducer concentration for each variant (Figure 4.2). And also eliminated variants for which we acquired too little reads (see Figure 4.2 and Methods for additional details). To ascertain the validity of our assay, we first characterized the behavior of our control variants (Figure 4.3a). A linear-like down-regulatory effect as a function of RBP induction is observed for the positive control variants (green), while no response in mCherry levels is observed for the negative controls (red). Additionally, the spread in mCherry at high induction levels is significantly smaller for the positive control than that of the negative control variants.

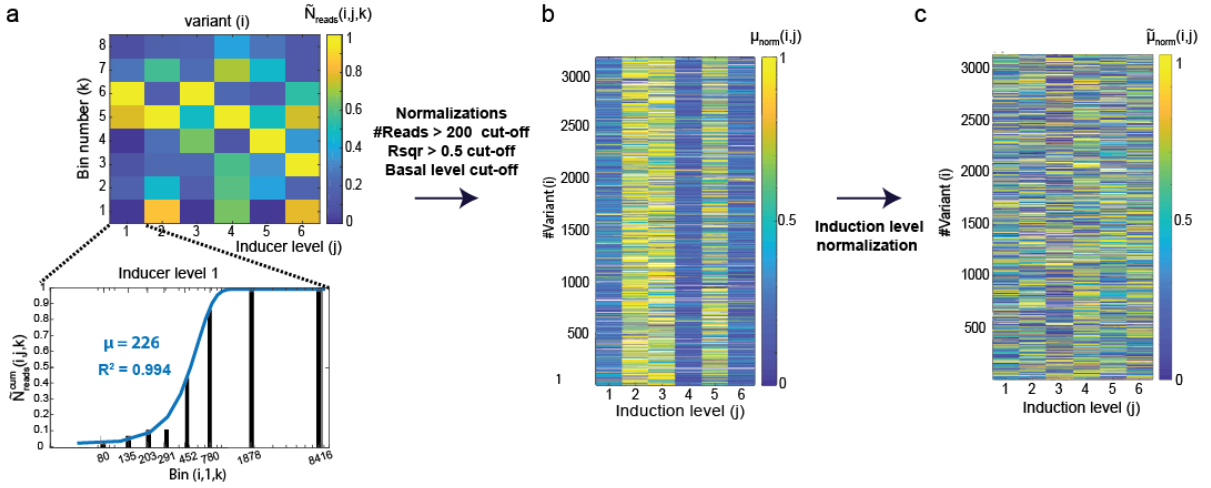


Figure 4.2: Flowchart for the preliminary analysis conducted on the reads extracted from the oligo-library experiment. (a) (Top) a sample 6x8 matrix obtained for each variant. (Bottom) Collapsing the matrix to a vector of integrated mCherry level for every inducer value. (b) Sample list for PCP of unsorted non-renormalized 6-vectors displayed as heatmap. (c) Renormalized heatmap displaying unsorted PCP responsive variants.

Next, to sort the variants in accordance with their likelihood of binding the RBP (i.e. similarity of their dose-response to the positive control's), we carried out the following computation (for details see SI). First, we characterized all variants by calculating a vector composed of three components: the slope of a linear regression, its goodness of fit (R-square), and standard deviation of the fluorescence value at the three highest induction bins (Figure 4.3b-middle). Next, we computed two multivariate Gaussian distributions using the empirical 3-component vectors that were extracted for the positive and negative controls and for the given RBP, to

yield a probability distribution function (pdf) for both the responsive and non-responsive variants, respectively (Figure 4.3b-right). The two populations are relatively well-separated from one another, presenting two only-slightly overlapping clusters. Finally, we defined the "Responsiveness score" for each variant (R_{score} – see methods for formal definition) as the logarithm of the ratio of the probabilities computed by the responsive pdf to the non-responsive pdf. This score was computed for each unique barcode, and the final result for a sequence variant was averaged over up to five vectors, one for every variant barcode that passes the read-number and basal-level thresholds.

In Figure 4.3c left, we plot the expression heatmap of the ~18k variants with PCP sorted (top to bottom) by decreasing R_{score}. The plot shows that 5470 variants exhibit an apparent down-regulatory response, defined as $\log(R_{score})>0$, corresponding to having a larger probability to belonging to the positive control distribution as compared with the negative. By comparison (Figure 4.4), MCP and QCP yielded 2604 and 7306 such variants, respectively. This indicates that while QCP may be the most promiscuous RBP in our library (i.e. tolerates a more varied set of binding sites), MCP is likely to be the most limited in terms of binding specificity. By comparison (Figure S3), MCP and QCP yielded 2604 and 7306 such variants, respectively. This indicates that while QCP may be the most promiscuous RBP in our library (i.e. tolerates a more varied set of binding sites), MCP is likely to be the most limited in terms of binding specificity. A closer observation of the top of the list (top 2000, Figure 4.3C-right) indicates that for a high R_{score}, a rapid reduction in fluorescence is detected in the second bin, which indicates that these variants also seem to exhibit the strongest binding affinity. We next plot the R_{score} obtained for all three RBPs, for each variant (Figure 4.3d). We overlay the plot with colored dots corresponding to the variants with $R_{score} > 3.5$ in each list, corresponding to the most specific variants. The plots reveal very little overlap between the subsets of variants that are highly responsive to the different RBPs, indicating that the vast majority of these highly-responsive binding sites are orthogonal (i.e. respond to only one RBP), which was expected for PCP & MCP and PCP & QCP, but not necessarily for MCP & QCP whose native sites are not mutually orthogonal[95,97,100–103].

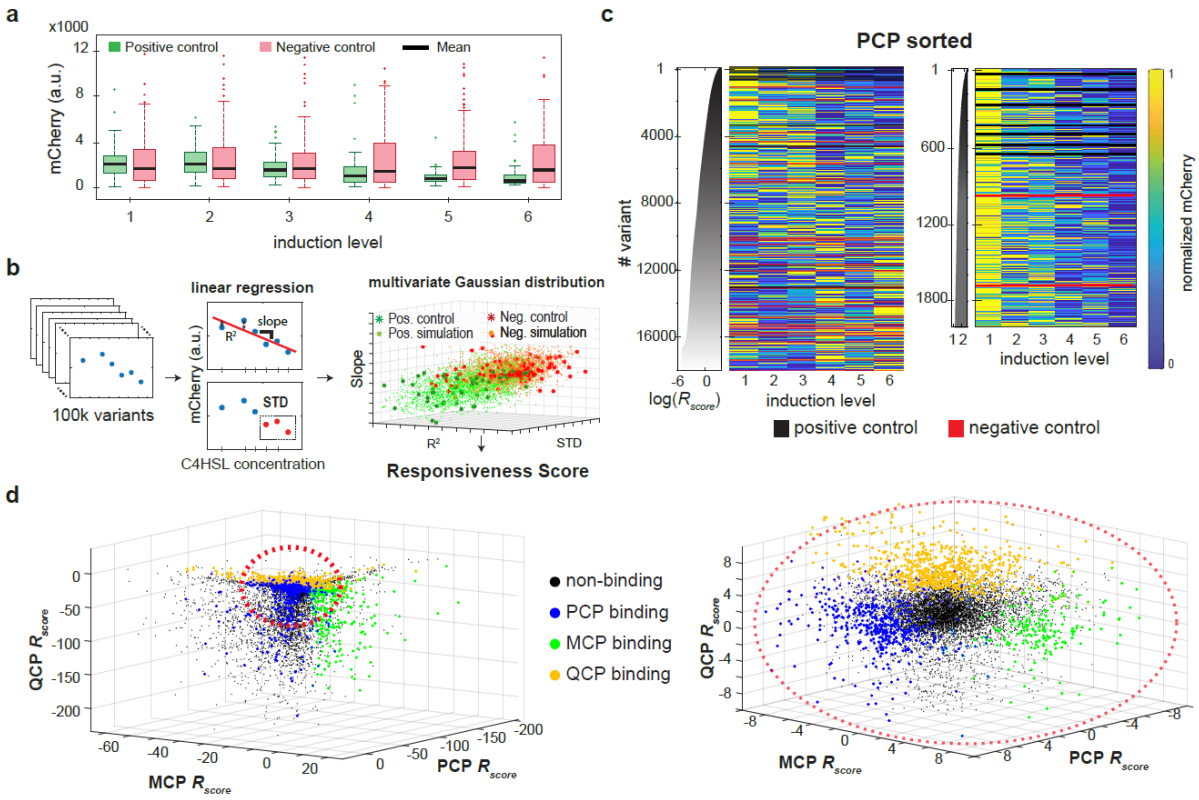


Figure 4.3. Responsiveness analysis and results. (a) Boxplots of mCherry levels for the positive and negative control variants at each of the six induction levels for PCP-GFP. (b) Schema for responsiveness score (R_{score}) analysis. (Left & middle) Linear regression was conducted for each of the 100k variants, and two parameters were extracted: slope and goodness of fit (R^2). The third parameter is the standard deviation (STD) of the fluorescence values at the three highest induction levels. (Right) Location of the positive control (dark green stars) and negative control (red stars) in the 3D-space spanned by the three parameters. Both populations (positive and negative) were fitted to 3D-Gaussians, and simulated data points were sampled from their probability density functions (pdfs) (orange for negative and green for positive). Based on these pdfs the R_{score} was calculated. (c) (Left) Heatmap of normalized mCherry expression for the ~20k variants with PCP. Variants are sorted by R_{score}. Black and red lines are positive and negative controls, respectively, and the grey graph is the R_{score} as a function of variant. (Right) "Zoom-in" on the 2,000 top-R_{score} binding sites for PCP. (d) (Left) 3D-representation of the R_{score} for every binding site in the library and all RBPs. Responsive binding sites, i.e. sites with R_{score}>3.5, are colored red for PCP, green for MCP, and orange for QCP. (Right) "Zoom-in" on the central highly concentrated region.

RBP binding sequence preferences

Using empirical R_{score} values and associated binding site sequences as training set, we developed an ML-based method that predicts the R_{score} values for every mutation in the wild type (WT) sequences. We first built a model specific to each protein and its WT binding site length to validate our OL measurements on prior knowledge of the proteins' binding specificities. To do so, we used a neural network that receives as input the sequence of a binding site the same length as the WT sequences (25nt for PP7-wt, 19nt for MS2-wt, and 20nt for Qβ-wt) and outputs a single score. We trained a specific network for each of the three RBP-

OL experiments and the two positions where the binding sites were embedded within the ribosomal initiation region (Figures 4.5a), resulting in a total of six different models. Such a model preserves the positional information for each feature, i.e. the position of each nucleotide in the WT binding site. To choose the position (δ) in which more robust scores were measured, we looked at the average AUC (area under the receiver operating characteristic curve) over 10-fold cross validation. The AUC scores for the most robust position yielded values of 0.65 for PCP with PCP-based sites and δ=C, 0.8 for MCP with MCP-based sites and δ =GC, and 0.65 for QCP with QCP-based sites and δ =GC (Figure 4.5b). Interestingly, the variant group with higher AUC score was also characterized by higher basal mCherry expression levels (Figure 4.5c), which in turn resulted in a higher fold repression effect. Thus, higher AUC, meaning more robust predictability, correlated with higher fold-repression, which provided additional validity to our analysis.

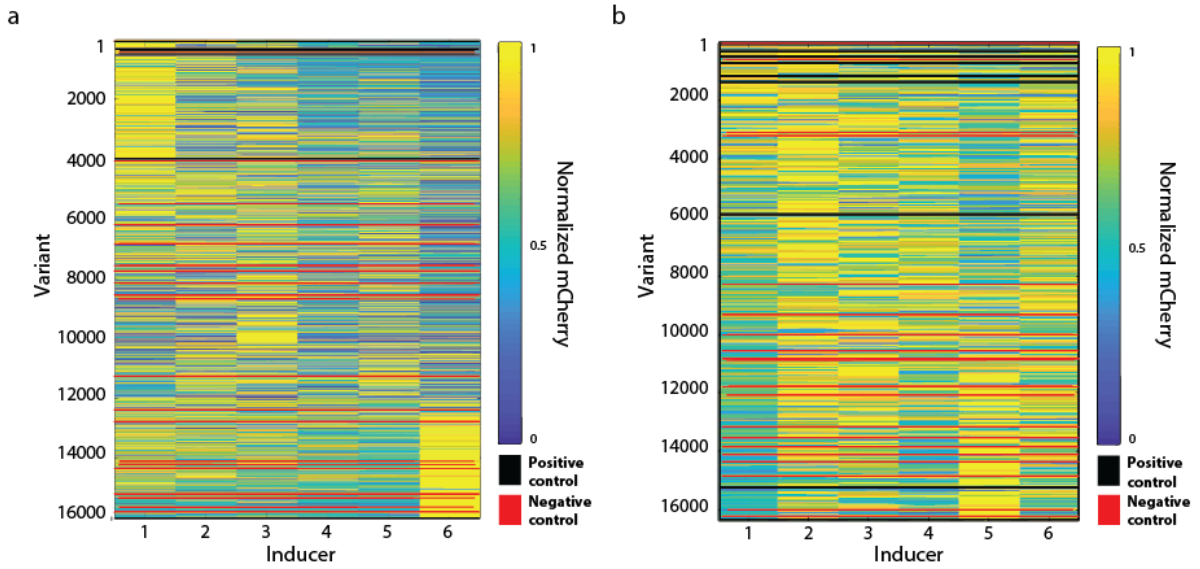


Figure 4.4: Sorted heatmaps for MCP and QCP. (a) R_{score} Sorted heat-maps of MCP, and (b) QCP with the OL. Positive and negative control are depicted in black and red, respectively.

In order to better understand the relationship between binding site sequence and binding, we used the model to analyze the effect of structure-conserving mutations in each of the WT binding-site sequences (Figure 4.5d). We present the ML model's results as "binding rules" depicted in illustrations for each of the three RBPs. The schemas represent the predicted responsiveness for every single-nucleotide mutation (SNP) in the loop or the bulge region, and every di-nucleotide mutation (DNP) preserving stem structure in the stem regions. For instance, in the schema for PCP, mutating the bulge from A to C or U sharply reduces the structure's predicted responsiveness. In addition, mutation of the second nucleotide in the loop from U to either A or G abolishes the predicted responsiveness, while mutation of the sixth nucleotide in the loop leaves binding unaffected. A clear characteristic of PCP is the

tolerance to DNPs, which is reflected by the dominance of the blue colors for most mutations in the lower stem, together with the sensitivity to SNPs in the first part of the loop. It is important to note that our results for PCP broadly correlate with past works[95,97,104], which found the loop and the bulge regions to be critical for PCP binding, while sequence variations in the stems did not alter binding significantly. For MCP, a tolerance to DNPs in the lower stem emerges from our analysis, while a strong sensitivity to SNPs in the bulge and the loop regions is revealed. Past analysis[95,99,105] also highlighted the sensitivity to mutations in the loop and

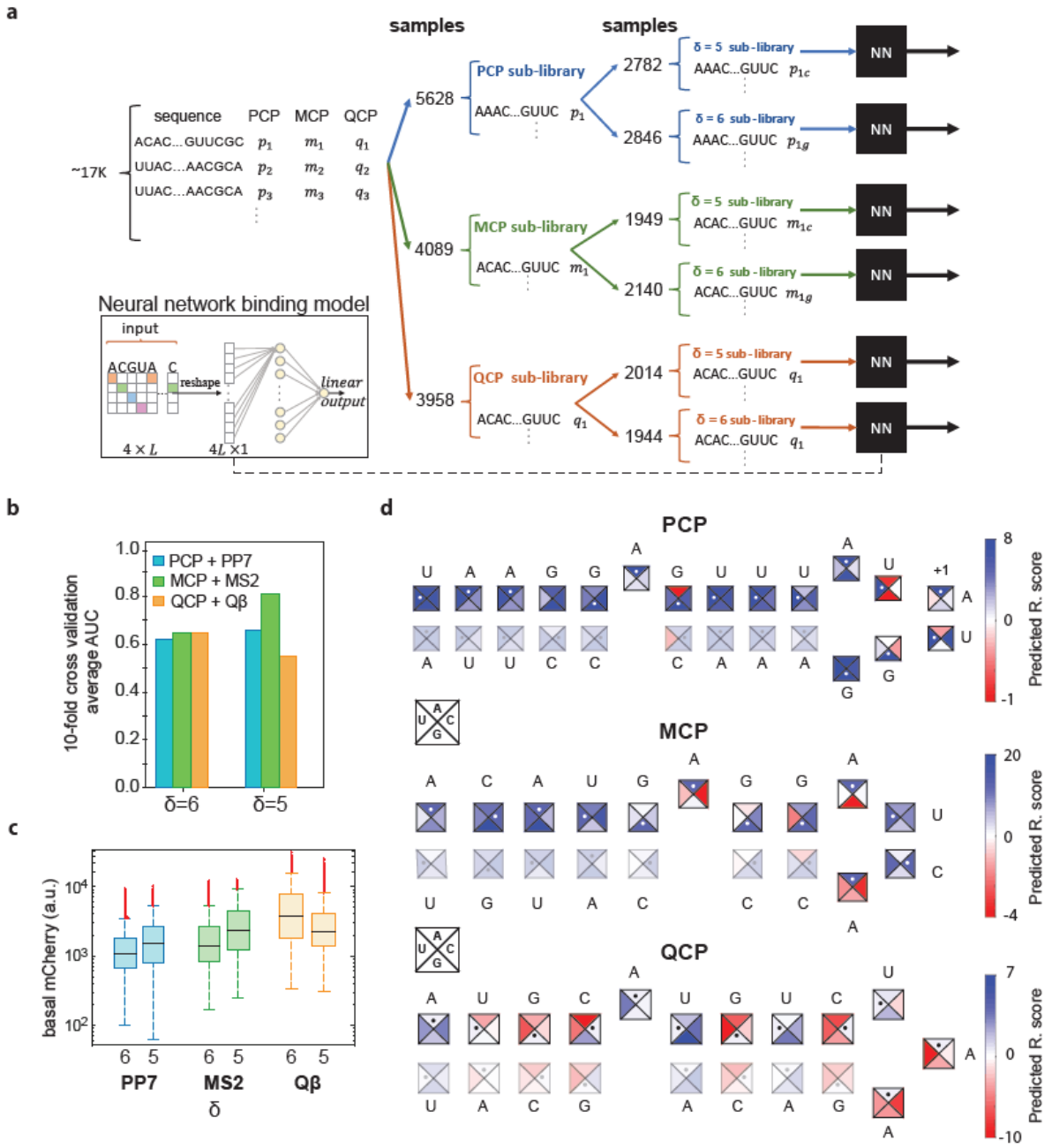


Figure 4.5. Analysis of MCP, PCP, and QCP RNA-binding sequence preferences. (a) Scheme for the data preparation and neural network (NN) architecture (inset) used. (b-c) Average AUC of the 10-fold cross validation (b) and box plots of the mCherry basal levels (c) conducted on the six sub-libraries: PCP

with PP7-based binding sites, MCP with MS2-based binding sites, and QCP with Qβ-based binding sites, all with either δ=6 or δ=5. (d) Illustrations of the NN predictions for the three sub-libraries for any single-point mutation. Each binding site is shown, with the wild-type sequence indicated as letters above and white dots inside the squares. Each square is divided to the four possible options of nucleotide identity, with the colors representing the predicted R_{score} for each option.

the bulge regions, indicating that the *in vivo* environment does not alter the overall binding characteristics of MCP.

Finally, for QCP (Figure 4.5d-bottom), a significantly different picture emerges. In some cases, it seems that the native sequence we used, as referred to in the literature[95,100,106], has a lower R_{score} than some mutated versions of it. The bulge, for instance, has a much higher R_{score} with U instead of the native A. The data seems to indicate that QCP prefers a four nucleotide K-rich (i.e. G/U) stem and a U bulge mini-motif. This motif is apparent throughout the binding site, as can be seen from the blue-colored nucleotides of both the lower and upper stems.

RBP binding structure preferences

In order to better understand the relationship between binding site structure and binding, we developed a protein-specific model based on the whole library, which we termed whole-library model (Figure 4.6a). This model, as opposed to our WT-specific NN, enables binding prediction to any site, i.e. of length different than the WT-site length. The model is based on a convolutional neural network (CNN) and receives as input both the sequence and secondary structure of the RNA binding site, as calculated by RNAfold[107].

We used this model to analyze the effect of structure-altering mutations on protein binding. To do so, we generated various binding sites with a predefined structure and used the whole-library models to predict their responsiveness score. Specifically, we looked at three types of mutations: alteration of upper-stem length, alteration of loop length, and alteration of bulge size. Overall, upper-stem length plays a big role in binding affinity for all three RBPs, though not equally (Figure 4.6b- left). PCP seems to be the most resilient to longer upper-stems, while MCP can relatively tolerate an upper-stem consisting of a single base-pair but is intolerant to stems of three base-pairs or longer. Finally, QCP exhibits tolerance to a two-base-pair stem, but a relative intolerance to any other length. Interestingly, this is consistent with QCP's known[95,98,100] weak binding affinity to the MS2-WT binding site.

Varying the loop-length suggests increased flexibility for all three RBPs (Figure 4.6b- right). PCP is the most resilient, displaying a viable binding affinity to loops that range from five to seven nucleotides in length. MCP is slightly less tolerant, displaying flexibility to structures containing loops that are three and four nucleotides in length, with some binding also observed for a small percentage of structures containing loops that are five nucleotides in

length. As for QCP's affinity to short stems, this result is also consistent with MCP's recorded low affinity to the Qβ-WT binding site. Finally, QCP is the least flexible CP, exhibiting affinity to loops that are two nucleotides in lengths, and some affinity to structures with loops of length five.

Finally, examining the importance of the bulge, a high variation in tolerance to mutations for the three RBPs is observed (Figure 4.6c). PCP can tolerate and even have higher affinity with sequences that either have no bulge, or a two-nucleotide bulge. This is depicted by a non-negligible variant density above the 3.5 threshold. MCP, on the other hand, has negligible tolerance for variants with no bulge, and very low tolerance for those with a two-nucleotide

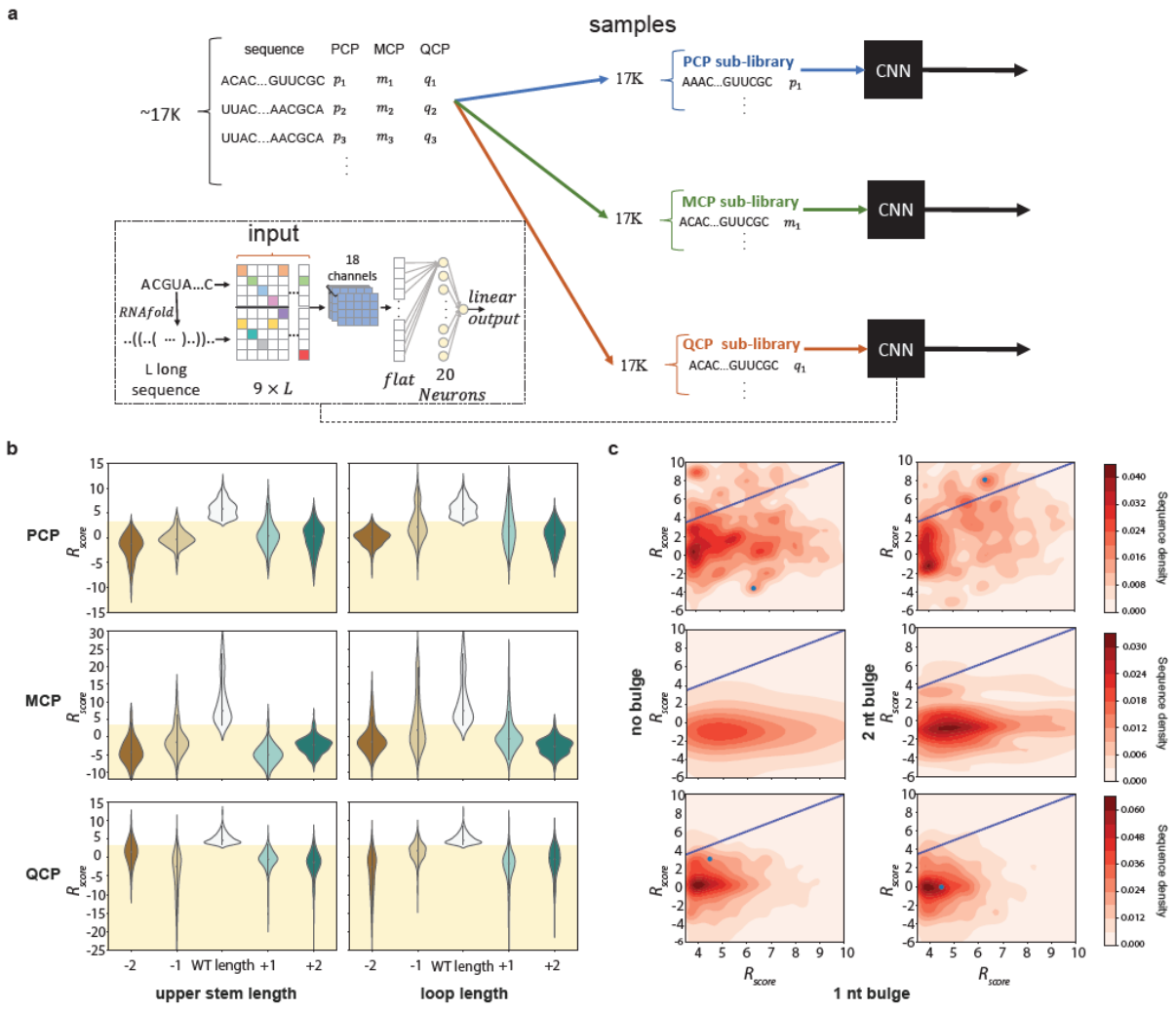


Figure 4.6. Analysis of MCP, PCP, and QCP RNA-binding structure preferences. (a) A scheme for the data preparation and neural network (NN) architecture (inset) used for the protein-specific convolutional neural network (CNN) model based on the whole library. We generated various binding sites with a predefined structure different from the wild-type and used the whole-library models to predict their responsiveness score. (b) Predicted R_{score} distributions for binding sites that differ in the length of the upper stem (left) or the loop (right) for PCP (top row), MCP (middle row), and QCP (bottom row). Stem and loop lengths were varied by ±2 base-pairs and nucleotides respectively. (c)

Density maps for predicted R_{score} for either no bulge (left-column) or a 2 nucleotide bulge (right-column) mutation of a wild-type-like structure for PCP-response (top-row), MCP-responsve (middle-row), and QCP-response (bottom-row).

bulge. This sensitivity correlates with MCP previous structure and sequence dependencies of the loop and upper stem (Figures 4.5d and Figure 4.6b). QCP displays some tolerance to both bulge mutations, though much less than PCP.

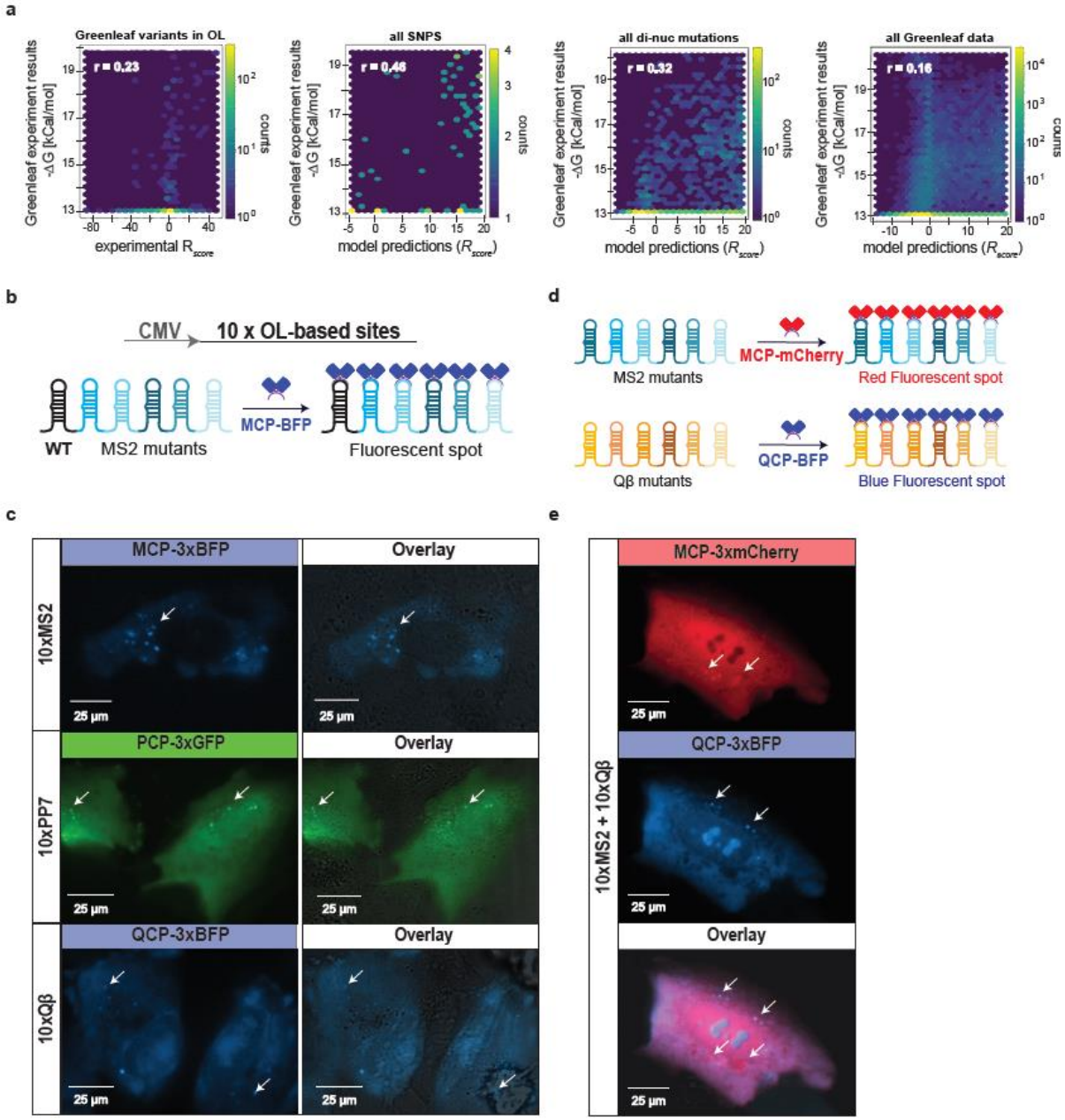


Figure 4.7. Validations: cassettes for RNA imaging in U2OS cells. (a) R_{score} comparison to ΔG results of a previous study that reported MCP binding to more than 129k sequences[25]. Each plot (from left-to-right) represents the correlation coefficient using: the experimental measurements for variants that were both in our OL and in the *in vitro* study, the R_{score} values predicted by our ML model for all single-mutation variants, for all double-mutation variants, and for the entire set of 129,248 mutated variants (b) Experiment design for the three cassettes based on the experimental binding sites. High R_{score} binding sites were incorporated into a ten-site cassette downstream to a CMV promoter. When the

matching RBP-3xFP is added (MCP-3xBFP is shown), it binds the binding-site cassette and creates a fluorescent spot. (c) The results for all three cassettes transfected with the matching RBP-3xFP plasmid into U2OS cells and imaged by fluorescence microscopy for detection of fluorescent *foci*. For each experiment, both the relevant fluorescent channel and the merged images with the differential interference contrast (DIC) channel are presented. (d) Experimental design for the orthogonality experiment: two separate cassettes with 10 predicted mutated sites for either MCP only or QCP only, respectively, were designed and transfected together with both MCP-3xmCherry and QCP-3xBFP, into U2OS cells. (e) Results for the orthogonality experiment: a cell presenting non-overlapping fluorescent *foci* from both fluorescent channels, indicating binding of MCP and QCP to different targets. Fluorescent wavelengths used in these experiments are: 400nm for BFP, 490nm for GFP, and 585nm for mCherry.

In summary, the structural analysis indicates that all three proteins prefer different structures, with some overlap that can create cross-binding (e.g. MCP to Qβ-WT). PCP seems to prefer a structure with an upper stem of length four base-pairs or longer and a variable loop size ranging from five to seven nucleotides with some sequence specificity. MCP is constrained in both structure and sequence specificity needing a bulge separating a lower and upper stem, two base-pair upper stem, and a loop length of three to five nucleotides in length with a conserved sequence signature. Finally, QCP seems to display a binding signature consistent with a repeat concatemer of 4-K-rich-stem-bulge sequence and structural motif.

Validations- new cassettes for RNA imaging

To validate both our experimental measurements and model predictions, we compared our results to a previous study that measured high-throughput *in vitro* RNA-binding of MCP[105]. In the study, the researchers employed a combined high-throughput sequencing and single molecule approach to quantitatively measure binding affinities and dissociation constants of MCP to more than 10^7 RNA sites using a flow-cell and *in vitro* transcription. The study reported ΔG values for over 120k variants, which formed a rich dataset to test correlation with our measured and predicted R_{score} values. First, we computed Pearson correlation coefficient of the purely experimental measurements for variants that were both in our library and in the *in vitro* study. The result (Figure 4.7a-left) indicates a positive and statistically significant correlation (R=0.23). We next predicted R_{score} values using the WT-specific model for all the reported variants of the *in vitro* study (Figure 4.7a left-to-right), and found a strong correlation (R=0.46) for single-mutations variants, a moderate correlation (R=0.32) for double-mutation variants and a weak correlation (R=0.16) with the entire set of 129,248 mutated variants. Given the large difference between the experiments and the different sets of variants used (e.g. *in vitro* vs. *in vivo*, microscopy-based vs. flow cytometry-based), the positive correlation coefficients (p-values<0.0002 for all reported coefficients) indicate a good agreement for both sets of experimental data, and a wide applicability for the learned binding models for MCP.

To further validate the results of our experiment and test the wider applicability of the findings, we generated new cassettes containing multiple non-repetitive RBP binding sites identified by our experimental data set, and tested them in mammalian cells. Once labelled with a fusion of the RBP to a fluorescent protein, functional cassettes appear as trackable bright fluorescent *foci*. We designed three binding site cassettes based on library variants that were identified as highly responsive for each RBP (Figure 4.7b). Each cassette was designed with ten different binding sites, all characterized by a large edit distance (i.e. at least 5) from the respective WT site, thus creating a sufficiently non-repeating cassette that IDT was able to synthesize in three working days. In addition, all selected binding sites exhibited non-responsive behavior to the two other RBPs in our experiment. We cloned the cassettes into a vector downstream to a CMV promoter for mammalian expression and transfected them into U2OS cells together with one of the RBP-3xFP plasmid encoding either PCP-3xGFP, MCP-3xBFP, or QCP-3xBFP. In a typical cell (Figure 4.7c), all three cassettes generated more than five fluorescent puncta, dispersed throughout the cytoplasm. The puncta were characterized by rapid mobility within the cytoplasm, and a lack of overlap with static granules or distinct features which also appear in the DIC channel (see Supplementary movies).

To expand our claim to orthogonal and simultaneous imaging of multiple promoters, we ordered two additional cassettes with MS2 and Qβ variants, respectively, and co-transfected them with a plasmid encoding for both of the matching fusion proteins: MCP-3xmCherry and QCP-3xBFP (Figure 4.7d). For each cassette, the sites were chosen with two constraints: to minimize repeat sequences and to maximize orthogonality to the other RBP (e.g. both MS2-WT and Qβ-WT binding sites were not included as they exhibit cross-responsiveness and are thus not orthogonal). In Figure 4.5e we plot sample cell images depicting single and double channel views. The images show that both cassettes produce a spatially distinct set of puncta (Figure 4.7e-top and middle), which can be definitively associated with one of the two proteins (Figure 4.7e-bottom). This indicates that our binding sites are sufficiently orthogonal to allow tracking of more than one cassette simultaneously. Moreover, there is little difference between the number of puncta of the two sequences and the fluorescent intensity for all puncta seem to fluctuate unimpeded in all three directions (x, y and z) inside the cell. Taken together, the microscopy experiments conducted in mammalian cells demonstrate the universal applicability of the results obtained from the high-responsiveness binding sites identified in the OL experiment to the advancement of RNA imaging in a variety of cell types.

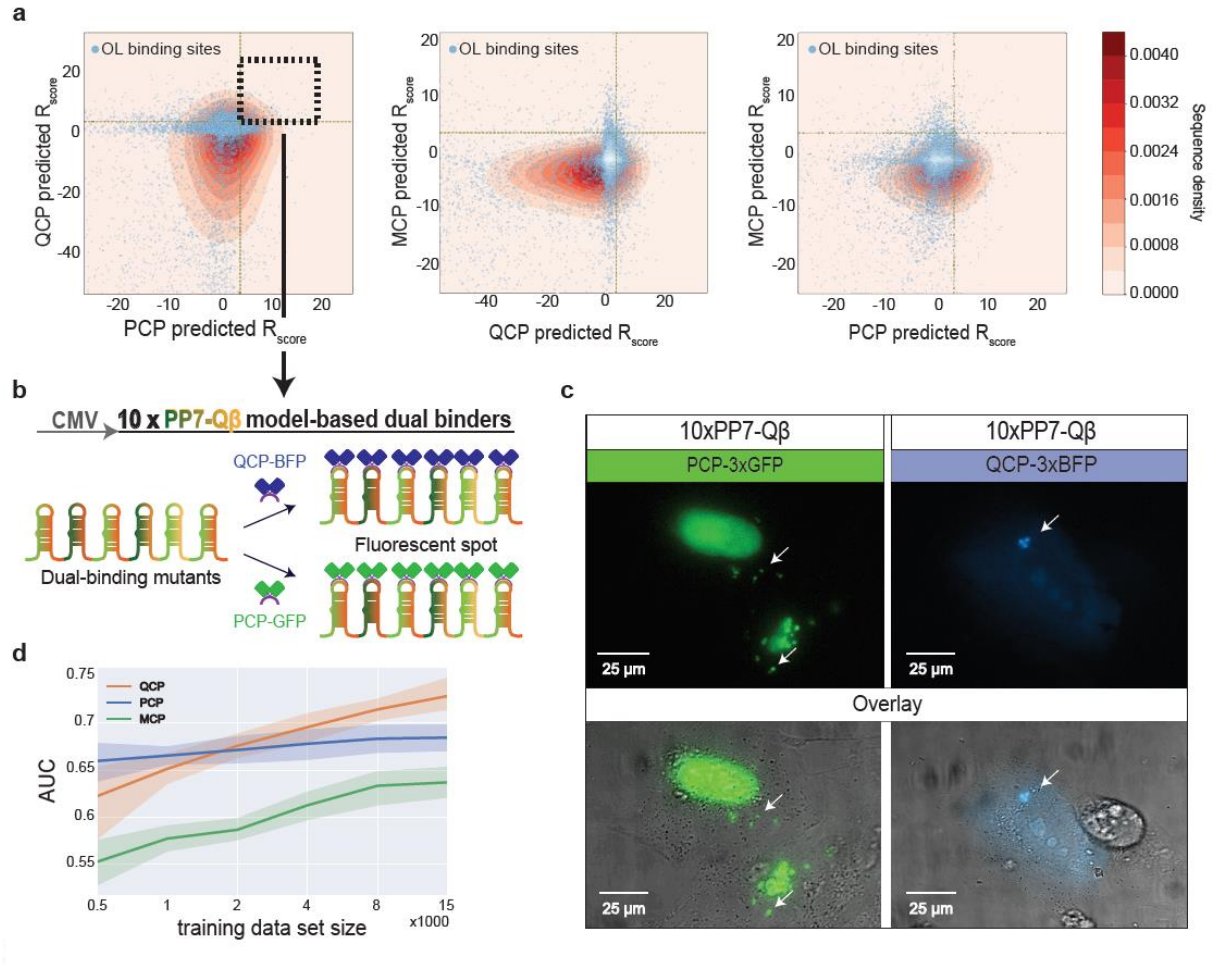


Figure 4.8. *De novo* design of dual-binding site cassettes in U2OS cells. (a) 2D density plots (pink-red scale) depicting the predicted R_{score} values for one million ML variants binding to (left-to-right): PCP and QCP, MCP and QCP, and MCP and PCP. QCP-PCP dual-binding variants are located in the black dashed square. Blue-white dots represent the experimental OL variants. (b) Based on the dual-binding mutants for QCP and PCP from our model predictions, we designed an additional cassette. (c) Results for the dual-binding experiment. Fluorescent *foci* can be observed for the cassette expressed with either PCP-3xGFP or QCP-3xBFP. For both experiments, both the relevant fluorescent channel and the merged images with the DIC channel are presented. Fluorescent wavelengths used in these experiments are: 400nm for BFP and 490nm for GFP. (d) Evaluation of prediction accuracy based on size of the training set. For each training set size, a random set of more than 1,000 training-set variants was withheld for computational testing post-training. Performance is reported as average AUC over 10 random training and test sets (and standard deviation in shade).

De novo design of dual-binding site cassettes

Finally, we wished to further validate our predictive power by creating cassettes with binding preferences that do not currently exist. We used the whole-library models to predict *de novo* functional binding site sequences, which could bind multiple RBPs. To do so, we generated all possible variants with Hamming distance 3-7 to one of the three WTs. From this set of sequences, we randomly selected one million sequences and used the models to predict the responsiveness score for each of the three RBPs. In Figure 4.8a, we plot the variant density

distribution based on a predicted R_{score} values. The plots show that the highest density of sequences appears at R_{score} values that hover around 0 for all three proteins. The plots further show that there is a bias towards negative responsiveness values for all three proteins in the computed sequences. This is consistent with having a small region of sequence space which facilitates specific binding, which in turn is easy to abolish with a small number of mutations. In contrast, high responsiveness scores are only computed for a small number of the sequences, as can be seen by the sharp gradient in the density plot for positive responsiveness values. Finally, each plot shows a non-negligible region where the same sequence exhibits a high responsiveness score for both RBPs. These sequences are predicted to be double binders. By overlaying the empirical responsiveness score for all the variants in our library (white and blue dots), we observe that the dual-binder region is inhabited by a handful of experimental variants for each possible RBP pair.

To test the predictions of the whole-library models experimentally, we designed another 10x binding site cassette (Figure 4.8b), where each binding site was selected from the set of predicted sequences whose responsiveness scores for QCP and PCP were both above 3.5 (see dashed square in Figure 4.8a-left panel). Therefore, we expected the cassette to generate fluorescent *foci* when bound by either QCP or PCP. As before, we cloned the cassette into a vector downstream of a CMV promoter for mammalian expression and transfected it into U2OS cells together with a plasmid encoding for either PCP-3xGFP or QCP-3xBFP. In Figure 4.8c, we plot fluorescent and DIC images for PCP (left) and QCP (right), depicting bright fluorescent *foci* that are located outside of the nucleus and which do not overlap with a DIC feature. The plots show distinct puncta observed with both relevant RBPs confirming the dual binding nature of the cassette. Consequently, these images support our model's ability to accurately predict MCP, PCP, and QCP binding sequences with known function with respect to all three RBPs.

Part 5: Discussion

During my PhD I studies protein-RNA interactions, focusing on their Synthetic Biology applications. The first goal was to improve a universal method for live RNA imaging using Synthetic Biology tools, a liquid-handling robot, and machine learning. The first step to achieving this goal was to create an assay for quantifying protein affinity in a cellular environment, based on a combined synthetic biology and SHAPE-seq approach.

Using our library of RNA regulatory variants comprising of binding site downstream to the AUG of an mCherry gene, we identified and characterized a position-dependent repression of translation when the hairpin was bound by an RBP. The extent of the repression effect was strongly dependent on position, and diminished for δ>15. The localization of a strong inhibition effect to region nearby the AUG for at least two different RBP-hairpin pairs suggests that this region may be particularly susceptible for repression effects. Previous works[108,109] have provided evidence that the ribosomal initiation region extends from the RBS to about 9-11 nucleotides downstream of the AUG (δ=12 to δ=14 as in our coordinate system). Furthermore, these authors also showed that structured stems of 6 bp or longer in the N-terminus can silence expression up to +11-13 from the AUG, but show negligible silencing when positioned further downstream. Thus, the region where the strong regulatory effects were detected in our experiments likely overlaps with the presumed ribosomal initiation region. This suggests that translation initiation may be susceptible to regulation, which can be an important guideline for RNA-based synthetic biology circuit design. The strong fold repression effect generated by the RBP within the initiation region allowed us to characterize the specific *in vivo* interaction of each RBP-binding-site pair by an effective K_{RBP}, which we found to be independent of binding site location. Interestingly, the *in vivo* K_{RBP} measured for some of the binding sites relative to their native site, differ from past *in vitro* and *in situ* measurements. Such discrepancies may be due to structural constraints, as our *in vivo* RNA constructs were significantly longer than what was used previously *in vitro* and included 700 nt reporter gene. Another reason for these differences may stem from variations in structure of RNA molecules that emerges from their presence inside cells.

This work establishes a blueprint for an *in vivo* assay for measuring the dissociation constant of RBPs with respect to their candidate binding sites in a more natural *in vivo* setting. This assay can be used to discover additional binding sites for known RBPs, which could be utilized in synthetic biology applications where multiple non-identical or orthogonal binding sites are needed. Therefore, we proceeded to write a methods paper based on our binding assay, to make it easier for researchers who wish to implement this technique.

In the paper, we highlighted the advantages of our method- a relatively easy protocol that can be conducted without the use of sophisticated machinery, data analysis is straightforward, and the results are produced immediately, without the relatively long wait-time associated with New Generation Sequencing results.

One limitation to this method is that we have demonstrated that it only works in bacterial cells. However, a previous study[12] has demonstrated a repression effect using a similar approach for the L7AE RBP in mammalian cells. An additional limitation of the method is that the insertion of the binding site in the mCherry initiation region may repress basal mCherry levels. Structural complexity or high stability of the binding site can interfere with ribosomal initiation even in the absence of RBP, resulting in decreased mCherry basal levels. If basal levels are too low, the additional repression brought on by increasing concentrations of RBP will not be observable.

The main disadvantage of the method in comparison to *in vitro* methods, such as EMSA, is that the RBP-RNA binding affinity is not measured in absolute units of RBP concentration, but rather in terms of fusion-RBP fluorescence. This disadvantage is a direct result of the *in vivo* setting, which limits our ability to read out the actual concentrations of RBP. This disadvantage is offset by the benefits of measuring in the *in vivo* setting. As mentioned previously in this section, we have found differences in binding affinities when comparing results from our in vivo assay to previous in vitro and *in situ* assays.

After the successful implementation of the binding assay as single-clone experiments, we proceeded to develop this technique to a high-throughput OL platform in bacteria[95,96]. The goal was to generate a sufficiently large dataset for training an ML-based model to reliably predict functional non-repeating binding elements for the RBPs PP7, MS2, and Qβ. This way, we acquired a computational tool that allows us to bypass the DBT-cycle when designing new molecules encoding multiple repeats of these binding sites. This tool substantially shortens the time from design to functional applications and removes many of the previous restrictions associated with these systems, such as the need for repetitive cloning cycles, repeat-based structure formation, and limitation on the number of functional binding sites. We also demonstrated that our MCP and QCP sites are orthogonal to one another, allowing for an additional orthogonal channel. These achievements provide the community with a reliable design tool for new phage coat protein binding cassettes in a variety of organisms.

In addition to solving an important technological bottleneck, we were also inspired by the need to develop new approaches for understanding RNA-related problems. It is generally believed that the combinatorial nature of RNA sequence and its intramolecular interactions

lead to high complexity, making simulations based on biophysical models a difficult task with limited degree of success[107,110–112], even when cellular environment is not taken into account. As a result, little is known about the evolutionary constraints on RNA structures, making bioinformatic identification of functional RNAs difficult[105]. In this work, using the OL-ML approach, we were able to quantitatively model *in vivo* binding of three phage coat-protein to RNA, at single-base-pair resolution, and for every possible single-nucleotide mutation. Based on the model, we found that each RBP prefers a different set of structural and sequence specificities. In addition, we concluded that the wild-type binding site for QCP is sub-optimal, and we could design *de novo* dual-binding binding sites (PCP *and* QCP) as well as orthogonal binding sites (MCP-*only* or *Q*CP-*only*) that did not exist naturally. For such an endeavor to work, we had to achieve a level of understanding beyond that accomplished by typical single-clone approaches. Furthermore, our demonstration that modeling of single RNA binding sites in bacteria is sufficient for generating a reliably predictive model for multi-binding site cassettes in mammalian cells is evidence that, at least for this set of proteins, the RNA-RBP module can accommodate multiple cellular environments, thus constraining the complexity of the overall system. Consequently, our work paves the way for characterizing and predicting binding of additional RBPs in any cellular environment, in addition to providing a proof-of-concept for the OL-ML approach.

Finally, our work not only provides a blueprint for studying RNA-related systems, but also partially answers the question of how much data is needed to train a reliably predictive model that will allow one to bypass the DBT-cycle. In our case, several thousand variants were sufficient (Figure 4.6d). At the present time, it is impossible to tell whether this number is typical or “surprisingly” small, as there are very few experiments to compare with. However, given our previous (albeit partial) mechanistic and structural understanding regarding PCP, MCP, and QCP binding to RNA that informed the OL design process, a reduction in the number of OL variants needed for the learning process was expected. It is reasonable to assume that partial knowledge of a system could reduce the size of the useful training set, and is likely to be an important ingredient in generating a complete computational understanding of a system. Future work on other complex RNA-based molecular interaction systems will determine whether the OL-ML approach is indeed a useful tool for providing new mechanistic and structural insights into these systems.

In a follow-up work that is currently in review in *Nature Physics*[113], a fellow PhD student from our lab has used OL-based cassettes to show that protein-RNA complexes formed inside the cell trigger liquid-liquid phase separation in the cytosol. Using PP7-coat protein and Qβ-coat protein together with multi-binding-site RNA scaffolds and real-time tracking of RNA-protein

complexes into and out of the biocondensates reveals that the cytosol is divided into a dense liquid phase in the nucleoid-dominated region, and a dilute liquid phase in the polar regions. We provide evidence for this assertion using stationary phase cells, where emergence of non-polar biocondensate formation is consistent with a reduction in size of the dense-nucleoid phase. The bi-phasic hypothesis for the *E.coli* cytosol has implications for various transcriptional and translational processes, and could provide an alternative explanation for the Super-Poisson dynamics attributed to transcriptional bursts.

The second goal of my PhD research was to increase our understanding, and in turn, our ability for engineering post-transcriptional regulatory networks. We placed RBP binding sites in the 5' UTR region of an mCherry gene and expressed the matching RBPs at rising concentrations, looking for the effect different structures and sequences in the 5'UTR have on protein function. Additionally, we implemented a method for probing RNA structure in live cells called SHAPE-Seq in order to study the effect of structure on RNA function. Using a library of RNA variants, we found a complex set of regulatory responses, including translational repression, translational stimulation, and cooperative behavior. The up-regulation phenomenon, or translational stimulation, had been reported only once for a single natural example in bacteria, yet was mimicked by all four RBPs at multiple 5' UTR positions.

The interesting story of the two binding sites PP7-USS and PP7-wt, which differ only in two bases yet present opposing regulatory responses- down-regulation and upregulation respectively, has let us to the conclusion that the mechanism which drives the complexity observed can be described by a three-state system. We found a translationally-active and weakly-structured 5' UTR state (PP7-USs without protein), a translationally-inactive and highly-structured 5' UTR state (PP7-wt without protein), and an RBP-bound state with partial translation capacity (both constructs with protein). As a result, the same RBP can either up-regulate or down-regulate expression, depending on 5' UTR sequence context. This description deviates from the classic two-state regulatory model, which is often used as a theoretical basis for describing transcriptional and post-transcriptional regulation [114]. In a two-state model, a substrate can either be bound or not bound by a ligand, leading to either an active or inactive regulatory state. This implies that in the two-state scenario, a bound protein cannot both be an "activator" and a "repressor" without an additional interaction or constraint which alters the system. The appearance of two distinct mRNA states in the non-induced case *in vivo*, as compared with only one *in vitro,* suggests that *in vivo* the mRNA molecules can fold into one of two distinct phases: a molten phase that is amenable to translation, and a structured phase that inhibits translation.

On the more "applicative" side, these synthetic regulatory modules can be viewed as a new class of “protein-sensing-riboswitches”, which may ultimately have a wide utility in gene regulatory applications. Together with the previous work of positioning the sites in the ribosomal initiation region[115], we offer a set of modestly up-regulating and a range of down-regulating RBP-binding site pairs with tuneable affinities for four RBPs, three of which are orthogonal to each other (PCP, GCP, and QCP). While we emphasize that our results were obtained in *E. coli*, given the propensity of RBPs to alter the RNA structure via direct interaction, it is tempting to speculate that such an interaction may be a generic 5’ UTR mechanism that could be extended to other RBPs and other organisms.

For any follow-up work to take place, we must first ask how difficult is it to design an up-regulatory dose-response for an RBP *de novo*? Unfortunately, our data does not provide a satisfactory mechanistic outcome for a quantitative prediction, but a qualitative phase-based description, which is an initial step. Our experiments revealed no particular structural features that were associated with this regulatory switch, such as the release of a sequestered RBS, which has been reported before as a natural mechanism for translational stimulation[57,58]. Moreover, attempting to allocate a structural state for a certain sequence *in vivo* using *in-silico* RNA structure prediction tools is not a reliable approach, due to mechanistic differences between the *in vivo* and *in vitro* environment, which these models understandably do not take into account. Therefore, to provide a predictive blueprint for which sequences are likely to be translationally inactive in their native RBP unbound state, a better understanding of both RNA dynamics and the interaction of RNA with the translational machinery *in-vivo* needs to be established. Yet, our findings suggest that generating translational stimulation using RBPs may not be as difficult as previously thought. Finally, the described constructs add to the growing toolkit of translational regulatory parts and provide a working design for further exploration of both natural and synthetic post-transcriptional gene regulatory networks.

Taken together, my work has advanced the field of protein-RNA interactions in both an applicative sense and in our understanding of the underline mechanisms that drive RNA function. It has demonstrated the large impact a small change in sequence of an RNA molecule in the 5' UTR region can have on its structure, and in turn, on its function. It has also significantly advanced synthetic biology techniques that are based on protein-RNA interactions, such as the simple assay we developed to quantify the affinity between protein and RNA, and the development of repeat-free cassettes for RNA imaging and RNA-based genetic manipulation.

Part 6: Bibliography

1. Stülke, J. Control of transcription termination in bacteria by RNA-binding proteins that modulate RNA structures. *Arch Microbiol* **177**, 433–440 (2002).

2. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986 (2008).

3. Lee, Schedl, M., T. The Online Review of C. elegans Biology. in (WormBook, 2006).

4. Stefl, R., Skrisovska, L. & Allain, F. H.-T. RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Rep* **6**, 33–38 (2005).

5. Olsthoorn, R. C., Garde, G., Dayhuff, T., Atkins, J. F. & Van Duin, J. Nucleotide sequence of a single-stranded RNA phage from Pseudomonas aeruginosa: kinship to coliphages and conservation of regulatory RNA structures. *Virology* **206**, 611–625 (1995).

6. Forrest, K. M. & Gavis, E. R. Live Imaging of Endogenous RNA Reveals a Diffusion and Entrapment Mechanism for nanos mRNA Localization in Drosophila. *Current Biology* **13**, 1159–1168 (2003).

7. Spingola, M., Lim, F. & Peabody, D. S. Recognition of diverse RNAs by a single protein structural framework. *Archives of Biochemistry and Biophysics* **405**, 122–129 (2002).

8. Gott, J. M., Wilhelm, L. J. & Uhlenbeck, O. C. RNA binding properties of the coat protein from bacteriophage GA. *Nucleic Acids Res.* **19**, 6499–6503 (1991).

9. Chao, J. A., Patskovsky, Y., Almo, S. C. & Singer, R. H. Structural basis for the coevolution of a viral RNA–protein complex. *Nat Struct Mol Biol* **15**, 103–105 (2008).

10. Spingola, M. & Peabody, D. S. MS2 coat protein mutants which bind Qβ RNA. *Nucl. Acids Res.* **25**, 2808–2815 (1997).

11. Peabody, D. S. Translational repression by bacteriophage MS2 coat protein expressed from a plasmid. A system for genetic analysis of a protein-RNA interaction. *J. Biol. Chem.* **265**, 5684–5689 (1990).

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

13. Buenrostro, J. D. *et al.* Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat Biotech* **32**, 562–568 (2014).

14. Mitra, K. *et al.* Structure of the E. coli protein-conducting channel bound to a translating ribosome. *Nature* **438**, 318–324 (2005).

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

16. Olsthoorn, R. C., Zoog, S. & van Duin, J. Coevolution of RNA helix stability and Shine-Dalgarno complementarity in a translational start region. *Mol. Microbiol.* **15**, 333–339 (1995).

17. Munson, L. M., Stormo, G. D., Niece, R. L. & Reznikoff, W. S. lacZ translation initiation mutations. *J. Mol. Biol.* **177**, 663–683 (1984).

18. Smit, M. H. de & Duin, J. van. Control of Translational Initiation by mRNA Secondary Structure: A Quantitative Analysis. in *Post-Transcriptional Control of Gene Expression* (eds. McCarthy, J. E. G. & Tuite, M. F.) 169–184 (Springer Berlin Heidelberg, 1990).

19. Schauder, B. & McCarthy, J. E. The role of bases upstream of the Shine-Dalgarno region and in the coding sequence in the control of gene expression in Escherichia coli: translation and stability of mRNAs in vivo. *Gene* **78**, 59–72 (1989).

20. Butkus, M. E., Prundeanu, L. B. & Oliver, D. B. Translocon 'pulling' of nascent SecM controls the duration of its translational pause and secretion-responsive secA regulation. *J. Bacteriol.* **185**, 6719–6722 (2003).

21. Chen, G. & Yanofsky, C. Features of a leader peptide coding region that regulate translation initiation for the anti-TRAP protein of B. subtilis. *Mol. Cell* **13**, 703–711 (2004).

22. Gu, Z., Harrod, R., Rogers, E. J. & Lovett, P. S. Anti-peptidyl transferase leader peptides of attenuation-regulated chloramphenicol-resistance genes. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5612–5616 (1994).

23. Mayford, M. & Weisblum, B. Conformational alterations in the ermC transcript in vivo during induction. *EMBO J.* **8**, 4307–4314 (1989).

24. Nudler, E. & Mironov, A. S. The riboswitch control of bacterial metabolism. *Trends in Biochemical Sciences* **29**, 11–17 (2004).

25. Win, M. N. & Smolke, C. D. Higher-Order Cellular Information Processing with Synthetic RNA Devices. *Science* **322**, 456–460 (2008).

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

27. Green, A. A., Silver, P. A., Collins, J. J. & Yin, P. Toehold Switches: De-Novo-Designed Regulators of Gene Expression. *Cell* **159**, 925–939 (2014).

28. Wroblewska, L. *et al.* Mammalian synthetic circuits with RNA binding proteins for RNA-only delivery. *Nat Biotech* **33**, 839–841 (2015).

29. Harvey, I., Garneau, P. & Pelletier, J. Inhibition of translation by RNA-small molecule interactions. *RNA* **8**, 452–463 (2002).

30. Suess, B. *et al.* Conditional gene expression by controlling translation with tetracycline-binding aptamers. *Nucleic Acids Res* **31**, 1853–1858 (2003).

31. Desai, S. K. & Gallivan, J. P. Genetic Screens and Selections for Small Molecules Based on a Synthetic Riboswitch That Activates Protein Translation. *J. Am. Chem. Soc.* **126**, 13247–13254 (2004).

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

33. Green, A. A. *et al.* Complex cellular logic computation using ribocomputing devices. *Nature* **548**, 117–121 (2017).

34. Hentze, M. W. *et al.* Identification of the iron-responsive element for the translational regulation of human ferritin mRNA. *Science* **238**, 1570–1573 (1987).

35. St Johnston, D. Moving messages: the intracellular localization of mRNAs. *Nat. Rev. Mol. Cell Biol.* **6**, 363–375 (2005).

36. Lewis, C. J. T., Pan, T. & Kalsotra, A. RNA modifications and structures cooperate to guide RNA-protein interactions. *Nat Rev Mol Cell Biol* **18**, 202–210 (2017).

37. Khalil, A. S. & Collins, J. J. Synthetic biology: applications come of age. *Nature Reviews Genetics* **11**, 367–379 (2010).

38. Werstuck, G. & Green, M. R. Controlling gene expression in living cells through small molecule-RNA interactions. *Science* **282**, 296–298 (1998).

39. Isaacs, F. J., Dwyer, D. J. & Collins, J. J. RNA synthetic biology. *Nat. Biotechnol.* **24**, 545–554 (2006).

40. Hutvágner, G. & Zamore, P. D. A microRNA in a Multiple-Turnover RNAi Enzyme Complex. *Science* **297**, 2056–2060 (2002).

41. Rinaudo, K. *et al.* A universal RNAi-based logic evaluator that operates in mammalian cells. *Nat Biotech* **25**, 795–801 (2007).

42. Delebecque, C. J., Lindner, A. B., Silver, P. A. & Aldaye, F. A. Organization of Intracellular Reactions with Rationally Designed RNA Assemblies. *Science* **333**, 470–474 (2011).

43. Chen, A. H. & Silver, P. A. Designing biological compartmentalization. *Trends in Cell Biology* **22**, 662–670 (2012).

44. Ausländer, S. *et al.* A general design strategy for protein-responsive riboswitches in mammalian cells. *Nat Meth* **11**, 1154–1160 (2014).

45. Sachdeva, G., Garg, A., Godding, D., Way, J. C. & Silver, P. A. In vivo co-localization of enzymes on RNA scaffolds increases metabolic production in a geometrically dependent manner. *Nucleic Acids Res* **42**, 9493–9503 (2014).

46. Pardee, K. *et al.* Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell* **165**, 1255–1266 (2016).

47. Winkler, W. C. & Breaker, R. R. Regulation of Bacterial Gene Expression by Riboswitches. *Annual Review of Microbiology* **59**, 487–517 (2005).

48. Henkin, T. M. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev* **22**, 3383–3390 (2008).

49. Wittmann, A. & Suess, B. Engineered riboswitches: Expanding researchers' toolbox with synthetic RNA regulators. *FEBS Letters* **586**, 2076–2083 (2012).

50. Serganov, A. & Nudler, E. A Decade of Riboswitches. *Cell* **152**, 17–24 (2013).

51. Romaniuk, P. J., Lowary, P., Wu, H. N., Stormo, G. & Uhlenbeck, O. C. RNA binding site of R17 coat protein. *Biochemistry* **26**, 1563–1568 (1987).

52. Cerretti, D. P. *et al.* Human macrophage-colony stimulating factor: Alternative RNA and protein processing from a single gene. *Molecular Immunology* **25**, 761–770 (1988).

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

54. Schlax, P. J., Xavier, K. A., Gluick, T. C. & Draper, D. E. Translational repression of the Escherichia coli alpha operon mRNA: importance of an mRNA conformational switch and a ternary entrapment complex. *J. Biol. Chem.* **276**, 38494–38501 (2001).

55. Lim, F. & Peabody, D. S. RNA recognition site of PP7 coat protein. *Nucl. Acids Res.* **30**, 4138–4144 (2002).

56. Sacerdot Christine *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. *Molecular Microbiology* **29**, 1077–1090 (2002).

57. Hattman, S., Newman, L., Murthy, H. M. & Nagaraja, V. Com, the phage Mu mom translational activator, is a zinc-binding protein that binds specifically to its cognate mRNA. *PNAS* **88**, 10027–10031 (1991).

58. Wulczyn, F. G. & Kahmann, R. Translational stimulation: RNA sequence and structure requirements for binding of Com protein. *Cell* **65**, 259–269 (1991).

59. Gregorio, E. D., Preiss, T. & Hentze, M. W. Translation driven by an eIF4G core domain in vivo. *The EMBO Journal* **18**, 4865–4874 (1999).

60. Boutonnet, C. *et al.* Pharmacological-based translational induction of transgene expression in mammalian cells. *EMBO Rep* **5**, 721–727 (2004).

61. Goldfless, S. J., Belmont, B. J., de Paz, A. M., Liu, J. F. & Niles, J. C. Direct and specific chemical control of eukaryotic translation with a synthetic RNA-protein interaction. *Nucleic Acids Res.* **40**, e64 (2012).

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

63. Stapleton, J. A. *et al.* Feedback control of protein expression in mammalian cells by tunable synthetic translational inhibition. *ACS Synth Biol* **1**, 83–88 (2012).

64. Endo, K., Hayashi, K., Inoue, T. & Saito, H. A versatile cis-acting inverter module for synthetic translational switches. *Nature Communications* **4**, 2393 (2013).

65. Endo, K., Stapleton, J. A., Hayashi, K., Saito, H. & Inoue, T. Quantitative and simultaneous translational control of distinct mammalian mRNAs. *Nucleic Acids Res.* **41**, e135 (2013).

66. Levy, L. *et al.* A Synthetic Oligo Library and Sequencing Approach Reveals an Insulation Mechanism Encoded within Bacterial σ54 Promoters. *Cell Reports* **21**, 845–858 (2017).

67. Langan, R. A. *et al.* De novo design of bioactive protein switches. *Nature* **572**, 205–210 (2019).

68. Koodli, R. V. *et al.* EternaBrain: Automated RNA design through move sets and strategies from an Internet-scale RNA videogame. *PLOS Computational Biology* **15**, e1007059 (2019).

69. Vainberg Slutskin, I., Weingarten-Gabbay, S., Nir, R., Weinberger, A. & Segal, E. Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. *Nat Commun* **9**, 529 (2018).

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

71. Nielsen, A. A. K. *et al.* Genetic circuit design automation. *Science* **352**, aac7341 (2016).

72. Reis, A. C. *et al.* Simultaneous repression of multiple bacterial genes using nonrepetitive extra-long sgRNA arrays. *Nat Biotechnol* 1–8 (2019) doi:10.1038/s41587-019-0286-9.

73. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W. & Prasher, D. C. Green fluorescent protein as a marker for gene expression. *Science* **263**, 802–805 (1994).

74. Bertrand, E. *et al.* Localization of ASH1 mRNA Particles in Living Yeast. *Molecular Cell* **2**, 437–445 (1998).

75. Bertrand, E. *et al.* Localization of ASH1 mRNA Particles in Living Yeast. *Molecular Cell* **2**, 437–445 (1998).

76. Haim, L., Zipor, G., Aronov, S. & Gerst, J. E. A genomic integration method to visualize localization of endogenous mRNAs in living yeast. *Nat Meth* **4**, 409–412 (2007).

77. Campbell, P. D., Chao, J. A., Singer, R. H. & Marlow, F. L. Dynamic visualization of transcription and RNA subcellular localization in zebrafish. *Development* **142**, 1368–1374 (2015).

78. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* **123**, 1025–1036 (2005).

79. Singer, R. H. Reminiscences on my life with RNA: a self-indulgent perspective. *RNA* **21**, 508–509 (2015).

80. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).

81. Chubb, J. R., Trcek, T., Shenoy, S. M. & Singer, R. H. Transcriptional Pulsing of a Developmental Gene. *Current Biology* **16**, 1018–1025 (2006).

82. Yunger, S., Rosenfeld, L., Garini, Y. & Shav-Tal, Y. Single-allele analysis of transcription kinetics in living mammalian cells. *Nat Meth* **7**, 631–633 (2010).

83. Forrest, K. M. & Gavis, E. R. Live Imaging of Endogenous RNA Reveals a Diffusion and Entrapment Mechanism for nanos mRNA Localization in Drosophila. *Current Biology* **13**, 1159–1168 (2003).

84. Shechner, D. M., Hacisüleyman, E., Younger, S. T. & Rinn, J. L. CRISPR Display: A modular method for locus-specific targeting of long noncoding RNAs and synthetic RNA devices in vivo. *Nat Methods* **12**, 664–670 (2015).

85. Biswas, J., Rahman, R., Gupta, V., Rosbash, M. & Singer, R. H. MS2-TRIBE evaluates protein-RNA interactions and nuclear organization of transcription by RNA editing. *bioRxiv* 829606 (2019) doi:10.1101/829606.

86. Haimovich, G. *et al.* Use of the MS2 aptamer and coat protein for RNA localization in yeast: A response to "MS2 coat proteins bound to yeast mRNAs block 5′ to 3′ degradation and trap mRNA decay products: implications for the localization of mRNAs by MS2-MCP system". *RNA* **22**, 660–666 (2016).

87. KOBAYASHI, T. Ribosomal RNA gene repeats, their stability and cellular senescence. *Proc Jpn Acad Ser B Phys Biol Sci* **90**, 119–129 (2014).

88. Medina, G., Juárez, K., Valderrama, B. & Soberón-Chávez, G. Mechanism of Pseudomonas aeruginosa RhlR Transcriptional Regulation of the rhlAB Promoter. *Journal of Bacteriology* **185**, 5976–5983 (2003).

89. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

90. Keren, L. *et al.* Promoters maintain their relative activity levels under different growth conditions. *Mol. Syst. Biol.* **9**, 701 (2013).

91. Zeevi, D. *et al.* Compensation for differences in gene copy number among yeast ribosomal proteins is encoded within their promoters. *Genome Res.* **21**, 2114–2128 (2011).

92. Spitale, R. C. *et al.* RNA SHAPE analysis in living cells. *Nat Chem Biol* **9**, 18–20 (2013).

93. Flynn, R. A. *et al.* Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. *Nat. Protocols* **11**, 273–290 (2016).

94. Watters, K. E., Abbott, T. R. & Lucks, J. B. Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucl Acids Res* **44**, e12–e12 (2016).

95. Katz, N. *et al.* An in Vivo Binding Assay for RNA-Binding Proteins Based on Repression of a Reporter Gene. *ACS Synth Biol* **7**, 2765–2774 (2018).

96. Katz, N., Cohen, R., Atar, O., Goldberg, S. & Amit, R. An Assay for Quantifying Protein-RNA Binding in Bacteria | Protocol. (2019).

97. Lim, F. & Peabody, D. S. RNA recognition site of PP7 coat protein. *Nucleic Acids Res* **30**, 4138–4144 (2002).

98. Lim, F., Spingola, M. & Peabody, D. S. The RNA-binding Site of Bacteriophage Qβ Coat Protein. *J. Biol. Chem.* **271**, 31839–31845 (1996).

99. Johansson, H. E. *et al.* A thermodynamic analysis of the sequence-specific binding of RNA by bacteriophage MS2 coat protein. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 9244–9249 (1998).

100. Witherell, G. W. & Uhlenbeck, O. C. Specific RNA binding by Q.beta. coat protein. *Biochemistry* **28**, 71–76 (1989).

101. Peabody, D. S. Translational repression by bacteriophage MS2 coat protein expressed from a plasmid. A system for genetic analysis of a protein-RNA interaction. *J. Biol. Chem.* **265**, 5684–5689 (1990).

102. Spingola, M. & Peabody, D. S. MS2 coat protein mutants which bind Qbeta RNA. *Nucleic Acids Res* **25**, 2808–2815 (1997).

103. Lim, F., Spingola, M. & Peabody, D. S. The RNA-binding Site of Bacteriophage Qβ Coat Protein. *J. Biol. Chem.* **271**, 31839–31845 (1996).

104. Lim, F., Downey, T. P. & Peabody, D. S. Translational repression and specific RNA binding by the coat protein of the Pseudomonas phage PP7. *J. Biol. Chem.* **276**, 22507–22513 (2001).

105. Buenrostro, J. D. *et al.* Quantitative analysis of RNA-protein interactions on a massively parallel array for mapping biophysical and evolutionary landscapes. *Nat Biotechnol* **32**, 562–568 (2014).

106. Lim, F., Spingola, M. & Peabody, D. S. The RNA-binding Site of Bacteriophage Qβ Coat Protein. *J. Biol. Chem.* **271**, 31839–31845 (1996).

107. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26 (2011).

108. Paulus, M., Haslbeck, M. & Watzele, M. RNA stem–loop enhanced expression of previously non-expressible genes. *Nucleic Acids Research* **32**, e78–e78 (2004).

109. Espah Borujeni, A. *et al.* Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Research* **45**, 5437–5448 (2017).

110. Bundschuh, R. & Bruinsma, R. Melting of branched RNA molecules. *Phys. Rev. Lett.* **100**, 148101 (2008).

111. Miao, Z. *et al.* RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* **21**, 1066–1084 (2015).

112. Leamy, K. A., Assmann, S. M., Mathews, D. H. & Bevilacqua, P. C. Bridging the gap between in vitro and in vivo RNA folding. *Quarterly Reviews of Biophysics* **49**, (2016).

113. Granik, N., Katz, N., Goldberg, S. & Amit, R. Synthetic liquid-liquid phase separated RNA-protein biocondensates reveal a bi-phasic cytosol in E.coli. *bioRxiv* 682518 (2020) doi:10.1101/682518.

114. Bintu, L. *et al.* Transcriptional regulation by the numbers: models. *Curr Opin Genet Dev* **15**, 116–124 (2005).

115. Katz, N. *et al.* An in Vivo Binding Assay for RNA-Binding Proteins Based on Repression of a Reporter Gene. *ACS Synth. Biol.* (2018) doi:10.1021/acssynbio.8b00378.

האינטראקציה בין חלבונים ל-RNA: יישומים בביולוגיה סינטטית

חיבור על מחקר לשם מילוי חלקי של הדרישות לקבלת התואר דוקטור לפילוסופיה

נועה בץ

הוגש לסנט הטכניון – מכון טכנולוגי לישראל

אב תש"פ, חיפה, יולי 2020

המחקר נעשה בהנחיית פרופסור רועי עמית בפקולטה להנדסת ביוטכנולוגיה ומזון.

אני מודה לטכניון – מכון טכנולוגי לישראל על התמיכה הכספית הנדיבה בהשתלמות.

תקציר

בעבודה זו חקרנו את העולם המורכב של האינטרקציות בין חלבונים ורנ"א, בדגש על אפליקציות לעולם הביולוגיה הסינטטית. העבודה כללה שני חלקים עקריים, הראשון שעוסק בשיפור ושדרוג של כלי בעולם הדימות והמניפולציה של רנ"א, והשני בחקר הקשר בין הרצף של הרנ"א למבנה שלו ולתפקודו בתא.

תחילה, שמנו לנו למטרה לשפר מערכת קיימת למעקב אחר מולקולות רנ"א בתאים חיים. המערכבת מבוססת על חיבור בין חלבונים פלורסנטיים ואתרי קישור מרנ"א, כאשר למולקולות הרנ"א שאחריה מעוניינים לעקוב מוסיפים קסטה המורכבת ממספר חזרות של אתר קישור לחלבון בעל מבנה של סיכה. החזרתיות של אותו הרצף בעל מבנה מספר פעמים מובילה לבעיות ביכולות שלנו לייצר את הקסטה, לשמר אותה באורגניזם, ולקבל תוצאות כמותיות ומהימנות. על כן, המטרה הייתה לפתח כלי ממוחשב שיוכל de-novo לייצר אתרי קישור לחלבונים עם אפיניות גבוהה לחלבון אך רצף שונה.

למטרה זו, התחלנו בפיתוח פרוטוקול המאפשר כימות של אפיניות חלבון לאתר הקישור שלו בתוך תאי חיידק, בהתבסס על תחרות בין הריבוזום לחלבון לקישור הרנ"א. מיקמנו אתרי קישור באזור האינציאציה של הריבוזום ב-ORF של גן פלורסנטי, וביטאנו בריבוזים שונים את החלבון. התוצאות העידו על תלות חזקה בין הירידה בביטוי הגן בעקבות חיבור החלבון לבין מיקום האתר בקרבת ה-AUG. העיכוב בביטוי היה גבוה כל עוד האתר היה ממוקם באזור האיציאציה, ברגע שהמרחק בין האתר ל-AUG היה מספיק גדול כדי להיות מחוץ לאזור האיניציאציה, הביטוי נשאר ללא שינוי.

בעקבות ההצלחה של הפרוטוקול על ארבעה חלבונים, החלטנו לפרסם את הפרוטוקול בכתב עת מיוחד, המאפשר למדענים ברחבי העולם לבצע אותו ביתר קלות. כחלק מעבודה זו, תיכננתנו רובוט על מנת שיבצע את הניסויים ביעילות רבה.

לאחר מכן, ברצונינו היה לבדוק מספר רב של אתרי קישור על מנת לייצר רצפים de-novo, בשביל שמעבדות בכל העולם יוכלו להשתמש בכלי זה ולהזמין את האתרים שרצויים להם ללא צורך בשלב ניסוי ותעיה. למטרה זו, פיתחנו את טכניקת ה- induction-based Sort-Seq המבוססת על הפרוטוקול שהוזכר קודם לכן. בעזרת טכניקה זו, מדדנו אפיניות של 20,000 אתרים מוטנטים בו זמנית לשלושה חלבונים. בעזרת שיטות של מערכות לומדות ורשת נוירונים הגדלנו את מרחב האתרים למיליונים, וע"י כך זיהינו את המאפיינים החשובים לקישור החלבון מבחינת רצף ומבנה הרנ"א. לבסוף, תכננו קסטות רנ"א חדשות המרוכבות מאתרי קישור ללא חזרות לשלושת החלבונים, ווידאנו שהן עובדות בתאים אנימאליים. אחת הקסטות שנבדקה הייתה מבוססת על אתרי קישור עם אפיניות גבוהה לשניים משלושת החלבונים, דבר שאינו קיים בטבע. לסיכום חלק זה, פיתחנו כלי לקהילה המדעית המאפשר תכננון של קסטות ללא חזרות, למטרות מעקב אחר רנ"א בתאים חיים. באופן זה, קיצרנו משמעותית את הזמן בין הקמת המערכת ולדימות ומעקב אחר גנים, במקביל לשדרוג הטכניקה ולפיתוח היכולת לקבל מידע אמין וכמותי. עד כה, מעבדות מאוניברסטאות הרווארד, ברקלי, וקאלטק, התעניינו בתוצאות שלנו.

החלק השני של העבודה התמקד בקשר בין מבנה לתפקוד הרנ"א. ע"י שימוש במערכת דומה לזו שתוארה קודם לכן, הדגמנו שמחיקת שני בסיסים בלבד מתוך 25 בסיסים באתר הקישור וכ-2000 במולקולת הרנ"א השלמה, משנה את המבנה של כל המולקולה. כתוצאה מכך, החלבון שקודם לכן גרם לירידה בביטוי הרנ"א עכשיו גורם לאקטיבציה ולביטוי ביתר. אקטיבציה "ישירה" הנובעת משינוי קונפורמציה של הרנ"א ע"י החלבון בלבד, מהווה protein-based riboswitches, תופעה אשר נצפתה בעבר בטבע פעם אחת בלבד. בנוסף, השינוי הדרסטי בהשפעת קישור החלבון על ביטוי הרנ"א בעקבות שינוי קטן מאוד ברצף לו, הינו מפתיע ומחזק את העובדה שהיחס בין מבנה ותפקוד של רנ"א עודו נושא מלהיב ופתוח.