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.
* Lead contact

Abstract

The design-build-test (DBT) cycle in synthetic biology is considered to be a major bottleneck for progress in the field. The emergence of high-throughput experimental techniques, such as oligo libraries (OLs), combined with machine learning (ML) algorithms, provide the ingredients for a potential “big-data” solution that can generate a sufficient predictive capability to overcome the DBT bottleneck. In this work, we apply the OL-ML approach to the design of RNA cassettes used in gene editing and RNA tracking systems. RNA cassettes are typically made of repetitive hairpins, therefore hindering their retention, synthesis, and functionality. Here, we used a strong repression effect that occurs when an RBP binds its binding site within the ribosomal initiation region of *E. coli* to generate using an OL-experiment thousands of new binding sites for the phage coat proteins of bacteriophages MS2 (MCP), PP7 (PCP), and Qβ (QCP). We then applied a connected neural network ML approach to vastly expand this space of binding sites to millions of additional predicted sites, which allowed us to identify the structural and sequence features that are critical for the binding of each RBP. To verify our approach, we designed new non-repetitive binding site cassettes of ten sites each containing either experimentally verified or predicted sites, and tested their functionality in U2OS mammalian cells. We found that all of our cassettes exhibited multiple trackable puncta. Additionally, we designed and verified a cassette containing predicted binding sites, where each site can bind both PCP and QCP. Consequently, we provide the scientific community with a novel resource for rapidly creating functional non-repetitive binding site cassettes using one or more of three phage coat proteins with a variety of binding affinities for any application spanning bacteria to mammalian cells.

Introduction

The design, build, test (DBT) cycle in synthetic biology is considered by many to be a major bottleneck for progress in the field. Specifically, the field is still lacking computational methods that will enable users to reliably design their system of choice without going through multiple time-consuming DBT cycles before arriving at the desired solution. Recent studies have begun to tackle this challenge and to devise methods for diverse RNA-, DNA- and protein-based applications, with varying degrees of success (Koodli et al., 2019; Langan et al., 2019; Vainberg Slutskin et al., 2018). Perhaps the best known methods are the Cello algorithm and RBS calculator, which are limited to bacterial chassis (Salis et al., 2009; Nielsen et al., 2016).

The challenge of formulating such algorithms is rooted in the different molecular mechanisms, each governed by diverse dynamics, which in turn require different assumptions and parameter values to model properly using a kinetic or thermodynamic modelling approach. For example, computing the efficiency of ribosome binding sites (i.e. the RBS calculator) requires both a quantitative understanding of ribosome-RNA interactions and RNA structure formation at the translation initiation site. Alternatively, the evaluation of gene regulatory circuits (i.e. the Cello algorithm) requires an understanding of the kinetics of transcription factors binding to DNA, dynamics of mRNA transport, and kinetics of translational processes. Given the particular molecular characteristics needed for each model, it is therefore improbable to generalize these approaches to even related molecular systems (e.g. predicting ribosome binding in mammalian cells). Consequently, since knowledge of the necessary biological parameters is often lacking due to experimental restrictions, kinetic and thermodynamic models are largely qualitative for the vast majority of molecular systems with limited predictive power.

Reliable algorithms are especially needed for the design of RNA-centric functional modules for various applications. A notable study has been recently published, where the authors were successful in designing non-repetitive sgRNA cassettes for regulating multiple metabolic genes in bacteria (Reis et al., 2019) demonstrating the utility of such an approach to studying whole systems of genes. 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. In particular, the latter technology has been utilized in many studies and proven to be an important tool in gene editing and RNA-tracking related applications (Biswas et al., 2019; Chen et al., 2018; Lari et al., 2019; Shechner et al., 2015; Tutucci et al., 2018; Urbanek et al., 2014). However, current cassettes are characterized by repeated sequence elements, making them nearly impossible to synthesize, clone, and maintain, which limits their utility for robust quantitative measurements (KOBAYASHI, 2014) as well as expansion to more complex multi-genic applications.

Previous findings have determined that specificity in phage coat protein binding to RNA is determined by the structural elements formed by specific sequence motifs (Buenrostro et al., 2014; Katz et al., 2018, 2019; Lim and Peabody, 2002; Lim et al., 1996a, 2001; Peabody, 1990; Witherell and Uhlenbeck, 1989). This implies that for a given phage coat protein, many different sequences may fold into a common functional structure. The DBT problem for phage coat protein-binding cassette design can thus be

solved by generating a database of functional binding sites that are divergent from a sequence perspective, and then utilizing different sequences with the same functional structure in place of multiple repeats of the same RNA sequence. The emergence in recent years of high-throughput oligo library (OL) based-experiments (Kinney et al., 2010; Sharon et al., 2012; Gasperini et al., 2016; Levy et al., 2017; Vainberg Slutskin et al., 2018; Weingarten-Gabbay et al., 2016; Wetmore et al., 2015) provides a platform for testing hundreds of thousands of potential binding-site variants. While extremely useful for identifying functional variants, the OL scale is much smaller than the available sequence space for ~20nt-long binding sites, and thus many functional variants are not sampled. Recently-developed machine-learning (ML) algorithms (Park and Kellis, 2015; Wainberg et al., 2018; Eraslan et al., 2019) provide the necessary tool for computationally expanding the variant database to millions of potentially functional sequences, using the OL as an empirical training dataset. The result is an ML algorithm which can computationally score any sequence for the desired functionality.

In this work, we apply the OL-ML approach to the design of phage coat protein binding cassettes made of repeated hairpin structures. We generate an OL of many candidate RNA binding sites for the phage-coat proteins MS2, PP7, and Qβ, that are designed to maintain functional RNA structure, despite changes in sequence. We evaluate the function of the resulting RNA hairpins in a massively-parallel *in vivo* expression assay, in bacteria. We then utilize ML tools trained on the OL sequences and their experimental function scores to computationally predict and experimentally verify novel sequences that can bind the phage coat proteins with high affinity. Consequently, we demonstrate that sequences with non-repeating elements can be reliably designed, synthesized, cloned; and, once transcribed, exhibit the functionality expected from the original repeated hairpins, in both bacterial and mammalian cells. This achievement provides researchers with a new resource to rapidly design functional customized cassettes for RNA-based applications for multiple genes in any organism, effectively eliminating the DBT bottleneck for this technology.

Results

Induction-based Sort-Seq (iSort-Seq)

We previously showed 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)(Katz et al., 2018, 2019). The size 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 coat proteins of PP7, MS2 and Qβ, and positioned each site at two positions within the ribosomal initiation region (Figure 1A top). 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 1A bottom and Figure S1). 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(Johansson et al., 1998; Katz et al., 2018, 2019; Lim and Peabody, 2002; Lim et al., 1996b) 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 1B) at each of the two positions (spacers δ=C or δ=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 three phage coat proteins to GFP (Figure 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 (Fig. 1C). We sorted each inducer-concentration culture into eight predefined fluorescence bins, which resulted in an 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 show a high-affinity, down-regulatory dose-response for a positive variant (Figure 1D-bottom V1), and a no-affinity variant exhibiting no apparent regulatory effect as a function of induction (Figure 1D-bottom V2).

Calculating Binding Scores

We conducted preliminary analysis of the sequencing data to acquire mCherry levels per RBP and inducer concentration, for each variant (Figure S2 and SI). We proceeded only with variants that were covered by a significant amount of reads (above 200 reads), and high mCherry basal levels (see SI for further details), for which there is better potential for observing a repression effect. Next, to ascertain the validity of our assay, we characterized the behavior of our control variants (Figure 2A). 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. We characterized each variant of the positive and negative controls using a vector composed of three components: the goodness of fit (R-square), the slope of a linear regression, and standard deviation of the fluorescence value at the three highest induction bins (Figure 2B-middle).

Next, we identified subsets of the non-control RNA sequences that were either highly responsive to each RBP (i.e. binding sites with high effective binding affinity), or non-responsive sequences (i.e. sites with no effective binding affinity) based on the positive and negative controls. We computed two multivariate Gaussian distributions using the empirical 3-component vectors that were extracted for the negative and positive controls and for the given RBP, to yield a probability distribution function (pdf) for both the responsive and non-responsive variants (Fig. 2B-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 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 2C left, we plot the expression heatmap of the ~18k variants with PCP (i.e variants that exhibited more than 200 read) , sorted (top to bottom) by decreasing R_{score} (see also Figure S3 for MCP and QCP). The plot shows that 5470 variants (counted from the top) 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 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 2C-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 (Fig. 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(Katz et al., 2018; Lim and Peabody, 2002; Lim et al., 1996c; Peabody, 1990; Spingola and Peabody, 1997; Witherell and Uhlenbeck, 1989)**.**

Computational analysis of RNA-binding 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 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 and outputs a single score. We train a specific network for each of the three RBP-OL experiments (Figure 3A and SI). The network only considers variants with the same length as the three WT sequences (25nt for PP7-wt, 19nt for MS2-wt, and 20nt for Qβ-wt), and all variants were split into two categories according to their position within the ribosomal initiation region (Figures 3A and S4), 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 (spacer 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 with δ=C, 0.8 for MCP with MCP-based sites with δ=C, and 0.65 for QCP with QCP-based sites with δ=GC (Figure 3B). Interestingly, the variant group with higher AUC score was also characterized by higher basal mCherry expression levels (Figure 3C), 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.

In order to better understand the relationship between binding site sequence and binding, we used the model to analyze the effect of single-point mutations in each of the WT binding-site sequences, for each of the OL-RBPs (Figure 3D). 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 to the WT sequences. 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 quickly 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 single nucleotide mutations, which is reflected by the dominance of the blue colors for most mutations in the lower stem. The PCP binding site also appears to be tolerant to mutations in the upper stem, while the loop region (particularly the first three nucleotides) appear to be highly sensitive to mutations. It is important to note that our results for PCP broadly correlate with past works (Katz et al., 2018; Lim and Peabody, 2002; Lim et al., 2001), which found that loop region and bulge to be critical for PCP binding.

For MCP, a tolerance to mutations in the lower stem emerges from our analysis, while a strong sensitivity to mutations in the bulge and the loop regions is revealed. Past analysis (Buenrostro et al., 2014; Johansson et al., 1998; Katz et al., 2018) also highlighted the sensitivity to mutations in the loop and bulge regions, indicating that the *in vivo* environment does not alter the overall binding characteristics of MCP. Finally, for QCP (Fig. 3D-bottom), a significantly different picture emerges, as it seems that the native sequence we used (as referred to in the literature (Katz et al., 2018; Lim et al., 1996a; Witherell and Uhlenbeck, 1989)) is far from the highest-affinity sequence for binding. In most cases, it has a lower R_{score} than the mutated versions. The bulge, for instance, has a much higher R_{score} with U instead of the native A. Consequently, our data seems to indicate that QCP prefers a GC-rich stem and a U bulge, which is structurally different than the bulge-stem-loop and stem-loop preferences of MCP and PCP, respectively. Altogether, the three proteins preferably bind different motifs and structural elements, which explains the orthogonality observed in our data (Figure 2D).

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(Buenrostro et al., 2014). 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 using the purely experimental measurements for variants that were both in our library and in the *in vitro* study. The result (Figure 3E-left) indicates a positive and statistically significant correlation (R=0.23). We next predicted R_{score} values using our model for all the reported variants of the *in vitro* study (Figure 3E 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) indicate a good agreement for both sets of experimental data, and a wide applicability for the learned binding models for MCP.

Universal new cassettes for RNA imaging

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 4A). 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 euther PCP-3xGFP, MCP-3xBFP, or QCP-3xBFP. In a typical cell (Figure 4B), 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 4C). 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 Fig. 4D we plot sample cell images depicting single and double channel views. The images show that both cassettes produce a spatially distinct set of puncta (Fig. 4D-top and middle), which can be definitively associated with one of the two proteins (Fig. 4D-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.

De novo design of dual-binding site cassettes

Finally, to further verify the validity of the ML-approach, we opted to test the predictive power of the model. To do so, we altered our initial ML approach by developing a protein-specific model based on the whole library, which we termed whole-library model (Figure 5A, Fig. S5 for AUC curves, and further details in the SI). This model, as opposed to our 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). It receives as input both the sequence and structure of the RNA binding site, as calculated by RNAfold(Lorenz et al., 2011). It is composed of one convolution layer, a pooling layer, a fully connected layer and a single output node. All three CNNs showed improved predictive capabilities when the structural data was added into the network (Figure S6).

We used the whole-library model 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 5B, 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 *de novo* cassette sequences that consisted only of predicted dual-binder variants, with orthogonal WT sites. We designed another 10x binding site cassette (Fig. 5C), 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 Fig. 5B-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 5D, 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.

Discussion

In this study, we adapted our previously-developed binding assay for quantifying RBP-RNA binding site affinity *in vivo* to a high-throughput OL platform in bacteria (Katz et al., 2018, 2019). Consequently, we were able to generate a sufficiently large dataset for training an ML-based model to reliably predict functional non-repeating binding elements for the phage coat proteins of PP7, MS2, and Qβ. As a result, we now possess 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 (Bundschuh and Bruinsma, 2008; Leamy et al., 2016; Lorenz et al., 2011; Miao et al., 2015), 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 (Buenrostro et al., 2014). In this work, using the OL-ML approach, we were able to quantitatively model *in vivo* binding of three phage coat proteins to RNA, at single-base-pair resolution, and for every possible single-nucleotide mutation. Based on the model, we concluded that the binding site for QCP is not 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 (Fig. 5E). 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.

Supplementary Data

5 Supplementary figures

6 Supplementary tables

5 Supplementary movies

ACKNOWLEDGEMENT

The authors would like to acknowledge the Technion's LS&E staff (Tal Katz-Ezov and Anastasia Diviatis) for help with sequencing.

AUTHOR CONTRIBUTIONS

NK designed and carried out the oligo library and microscopy experiments and analysis for all constructs. ET wrote the code for the machine learning models and carried out the ML-based analysis. SG and OA assisted and guided the experiments and analysis. ZY assisted with the

OL design and statistical analysis of results. RA and YO supervised the study. NK, RA, and YO, wrote the manuscript.

FUNDING

This project received funding the European Union's Horizon 2020 Research And Innovation Programme under grant agreement no. 664918 - MRG-Grammar.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

Figure Legends

Figure 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 δ), 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 3 fusions. (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 (C4HSL) 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). Associated with Fig. S1

Figure 2. 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 (see SI) 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. Associated with Fig. S2, S3, and S4.

Figure 3. Analysis of MCP, PCP, and QCP RNA-binding 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=GC or δ=C. (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. (E) R_{score} comparison to ΔG results of a previous study that reported MCP binding to more than 129k sequences(Buenrostro et al., 2014). 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. Associated with Figure S5

Figure 4. Universal cassettes for RNA imaging in U2OS cells.

(A) 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. (B) 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. See also Movie S1 and S2 for MCP and PCP respectively. (C) Experimental design for the orthogonality experiment: two separate cassettes with 10 predicted dual-binding, mutated (no-WT) sites for only MCP or only QCP, respectively were designed and transfected together with both MCP-3xmCherry and QCP-3xBFP, into U2OS cells. (D) 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. See also Movie S3.

Figure 5. *De novo* design of dual-binding site cassettes in U2OS cells.

(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. (B) 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. (C) Based on the dual-binding mutants for QCP and PCP, we designed an additional cassette based on our models' predictions alone (no binding sites from the original OL). (D) 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. See also Movie S4 and S5 with PCP-GFP and QCP-BFP respectively. (E) 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). Associated with Fig. S6.

STAR * METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to and will be fulfilled by the corresponding author Roee Amit (roeeamit@technion.ac.il).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The Human Bone Osteosarcoma Epithelial Cell line (U2OS) female was incubated and maintained in 100x20mm cell culture under standard cell culture conditions at 37°C in humidified atmosphere containing 5% CO2 and were passaged at 80-85% confluence. Cells were washed once with 1x PBS and subsequently treated with 1mL trypsin/EDTA 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.

E. coli TOP10 cells were obtained from Invitrogen, cat number C404006 (see also Key resource 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 (Methods Details section for more details).

METHOD DETAILS

Bacterial OL work

Construction of the bacterial library

We designed 10,000 mutated versions of the WT binding sites to the phage coat proteins of PP7 (Figure 1 and S1), MS2 and Qβ, and positioned them at two positions within the ribosomal initiation region. 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 tested 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.

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.

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 Amplicilin – 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.

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 solution were 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, 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 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 sec. 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 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.

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.

Mammalian Cassette microscopy experiments

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 micrsocpy 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.

Mammalian Microscopy assay

Cell culture:
The Human Bone Osteosarcoma Epithelial Cell line (U2OS, see Key Resource table) 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 (XX) transfection reagent according to the manufacture's instructions. Typical DNA for transfection was 150ng from RBP-3xFP and 850ng from the cassette plasmid. After inculabtion of 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.

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 an 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.

QUANTIFICATION AND STATISTICAL ANALYSIS

Responsiveness score

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

Read normalization and filtration

Read number was normalization 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.

$$\text{Eq. 1:}\quad \mathrm{Nreads}(i,j,k) = \mathrm{Rreads}(i,j,k) \times \%\mathrm{Cells}(j,k), \quad \begin{matrix} i = 1:100k \\ j = 1:6 \\ k = 1:8 \end{matrix},$$

where Nreads is the number of normalized reads, Rreads is the number of raw reads, %Cells is the percentages of the cells in each bin 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.

mCherry μ estimation from reads

For each concentration, we have an 8-bin histogram for which we need to calculate the mCherry averaged fluorescence (μ). First, we normalize each strain number of reads per bin by the maximum number of reads in that inducer level for that variant (each column in the read matrix, Figure S2 top left).

$$\text{Eq. 2:}\quad \mathrm{Nreads_perLevel}(i,j) = \frac{\mathrm{Nreads}(i,j,k)}{\max\{\mathrm{Nreads}(i,j,k);\ k = 1:8\}}, \quad \begin{matrix} i = 1:100k \\ j = 1:6 \\ k = 1:8 \end{matrix},$$

where Nreads_perLevel is the normalized number of reads per level of inducer and N_reads is the number of normalized reads from Eq. 1.

We do this by fitting to a cumulative Gaussian (Figure S2 bottom left):

$$\text{Eq. 3:}\quad \mathrm{Nreads_perLevel_cumu}(i,j) = 0.5 + 0.5 \text{ x } \mathrm{erf}\left(\frac{\mathrm{maxBin}(j,k) - \mu}{\sigma\sqrt{2}}\right), \quad \begin{matrix} i = 1:100k \\ j = 1:6 \\ k = 1:8 \end{matrix},$$

where μ is the averaged mCherry fluorescence extracted from the fitting procedure, σ is the standard deviation for mCherry fluorescence extracted from the fitting procedure, maxBin is the maximum mCherry fluorescence for each bin as supplied by the sorter, Nreads_perLevel_cumu is the cumulative number of normalized reads over the eight bins per variant and per induction level.

Only induction levels that had a goodness-of-fit higher than 0.5 were taken into account.

μ normalization and filtration

Since each inducer concentration was done in different conditions (e.g. duration of incubation on ice, O/N shaking, binning time) and at a different time (different days), mCherry levels varied greatly as a function of induction. For example, very low levels are present at induction level 1 roughly for all variants displayed, while induction level 2 presents very high levels for all (Figure S2 middle). Therefore, we scaled each mCherry value based on the averaged μ values for the negative control variants only for each induction level. First, we calculated the mCherry scaled value for every variant:

$$\text{Eq. 4: } NmCh(i,j) = \frac{mCh(i,j)}{\max\{mCh(i,j);\ j=1:6\}},\quad \begin{matrix} i=1:100k \\ j=1:6 \end{matrix},$$

where NmCh is the normalized mCherry level per induction level for every variant and mCh is the mCherry level per induction level for every variant.

Then, we calculated the averaged normalized mCherry level of the negative controls only per induction level and used it to normalize all variants:

$$\text{Eq. 5: } SNmCh(i,j) = \frac{NmCh(i,j)}{\text{mean}\{NmCh_{NEG}(i_{NEG},j);i_{NEG}=1:neg_controls\}},\quad \begin{matrix} i=1:100k \\ j=1:6 \end{matrix},$$

where SNmCh is the scaled mCherry value of each variant and induction level, NmCh is the normalized mCherry level from Eq. 4, $NmCh_{NEG}$ is the normalized mCherry level for only the negative variants, and neg_controls are the total number of negative controls for each RBP experiment (220 for PCP, 160 for MCP and QCP).

After scaling and normalization, mCherry levels now present no particular trend as a function of induction level (Figure S2 right).

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.

Calculating the three principle components

On each variant's 6-value vector we next conducted a linear regression, to extract the slope and goodness of fit from the following equation:

$$\text{Eq. 6: } SNmCh(i) = slope \times ind_conc + const,\ i=1:100k,$$

where SNmCh is the value calculated in Eq. 5, ind_conc is the inducer concentration [uM], slope is the slope of the calculated linear fit, and const is the constant in the calculated linear fit.

The third principle component is the standard deviation of the fluorescence value at the three highest induction bins:

$$\text{Eq. 7: } mCh_std(i) = \text{std}\{SNmCh(i,j);j=4:6\},\ i=1:100k$$

where mCh_std is the mCherry standard deviation at the three highest induction level, four to six, and SNmCh is the same value calculated in Eq. 5.

Once acquired, we conducted a multi-variant Gaussian regression based on the three principle components (the goodness of fit (R-square) and the slope from a linear regression, and standard deviation of the fluorescence value at the three highest

induction bins) on two populations – the positive control and negative control. Based on the parameters extracted - μ and Σ - we then calculated the probability of each variant to belong to each of the two populations.

$$\text{Eq. 8: } \mathrm{pdf}\left(\mathrm{slope}, R^2, \mathrm{std}\right) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}{\sqrt{(2\pi)^3 |\Sigma|}},,$$

where pdf is the probability density function (pdf), X is a 3-by-1 vector with the three principle components, μ is the averaged values (3-by-1), and Σ is the co-variance matrix (3-by-3).

Finally, we calculated the R_{score}, which represents the similarity of each variant to either the positive or negative controls.

$$\text{Eq. 9: } \mathrm{R.score}(i,n) = \frac{\sum_{j=1}^{n} \log\left(\frac{p\left(X_{ij} \mid pos\right)}{p\left(X_{ij} \mid neg\right)}\right)}{\sqrt{n}}, \quad \begin{matrix} n = 1/2/\ldots/5 \\ i = 1:20k \end{matrix},$$

where R.score is the responsiveness score for each variant, p is the pdf result for a variant to belong to either in the positive distribution (pos) or the negative distribution (neg), and n is the number of barcode variants taken into account.

Machine-learning methods

We developed two types of models to predict the binding preferences of the three RNA binding proteins (RBPs): WT-specific and whole-library. Here, we will describe in detail the models, the choice of hyper-parameters and their training on experimental data. First, we will cover the features common to the two models. Then, we will provide details relevant to each of the two model types separately.

Dataset

The dataset contains RNA-binding intensity of three proteins (MCP, PCP and QCP) to approximately 17,000 sequences (PCP 17,177, MCP 17,213, QCP 16,041, and 12,245 in the intersection of the three). All sequences were either a variant of a known wild type (WT) binding site of one of the three proteins or a non-similar sequence that was used as control (PCP 42, MCP 40, QCP 38). The edit distance of the derived sequences from their WT mostly span 4 to 8 mutations or indels (Figure S1). The binding intensity score, also denoted as the responsiveness score, empirically spanned the range of -281 to 47. Each sequence has a positional feature, which defines its prefix and suffix, i.e. upstream and downstream flanking sequences, respectively. The prefix is either C or GC and the corresponding suffix is one out of three options: T, CT or no suffix. The choice of suffix is done in a way that guarantees no shift in the reading frame.

Data encoding

To provide the sequence data as input to the computational framework we used, it first needs to be transformed to numerical values. Each sequence was encoded using a traditional one-hot encoding of the sequence. Each nucleotide is converted to a four-bit vector with one bit set in the position corresponding to that nucleotide and all other positions set to zero. This way an L-long sequence is transformed into a 4 x L binary

matrix. L is either the WT length in the WT-specific model or 50 for the whole-library model.

Model evaluation

We performed 10-fold cross-validation to evaluate the binding models. We partitioned the dataset randomly into 10 equal-sized folds. Then, we trained and tested the model 10 times, each time using a different fold as the test set and the other nine folds combined as the training set. We used two measurements to gauge model performance: Pearson correlation and area under the ROC curve. Pearson correlation measures the linear agreement between two vectors, and is a common measure to evaluate intensity prediction. Area under the receiver operating curve (AUC) is a common measure to evaluate classification of positive and negative data points. We defined positive sequences as those having a binding intensity above 3.5 and negative sequences as those having a binding intensity of zero or less.

Parameters search

We used a hyper parameter search to optimize model performance. The first step of this search process was to randomly select a set of parameters from the parameter space defined for each of the models (Tables S1 and S2 below), build, train and test the model. That step was repeated 10 times. Each random parameter combination was evaluated using 10-fold cross-validation as described above. From the 10 random parameters sets, the best performing set was selected based on average Pearson correlation over the 10 folds. The second step of the search was “fine tuning” of the chosen parameter set. In this step, we checked sets of parameters in the “surrounding” of the set that was selected during the first step. This process of parameters selection was done for each protein and for each of the model types separately.

WT-specific binding model

Dataset division

We first tackled the challenge of developing a model based on a WT and its variants of the same length. For this aim, we used a different subset of the data for each protein. The protein-specific subset contained only the sequences that have the same length as its WT binding site (MS2 – 19nt, Qβ – 20nt, PP7 – 25nt). Then, we split the subset again by the prefix of the sequence (C or GC). The rationale for the second split is the low correlation in binding intensities observed between C and GC positions (Figure S4). This process is summarized in Figure 3A, main text.

Model description and optimization

Each WT-specific model is composed of 1-2 hidden layers with 10-40 nodes and one output layer with a single node (Figure 3A). Each protein and its sub-library have different parameters that were chosen specifically for it.

This optimization process was done as described under the Parameters search section above. The details of the parameters we examined are described in Table S5.

Parameter space			Final parameters (protein, prefix)					
Parameter	Initial space	Surrounding space	MCP-C	MCP-GC	QCP-C	QCP-GC	PCP-C	PCP-GC
Nodes	5-50	±5	38	32	39	39	10,10	10,10
Layers	1-3	-	1	1	1	1	2	2
Activation function	linear, tanh, relu	-	tanh	tanh	tanh	tanh	tanh	tanh

Epochs	20,30..100	±15 (strides of 5)	70	70	70	70	70	70

Table S5. *Parameters search space for WT-specific model, The parameter space for each of the two steps of the hyper parameters search and the final models parameters. Unless noted otherwise, the range specified is of stride 1.*

In addition to the parameters in Table , which are unique to each model, there are additional parameters that are common to all of them: learning rate 0.001 (default), batch size 8, optimizer ADAM, loss function MSE (mean squared error) and dropout of 0.2 for each hidden layer. The output layer consisted of one node with a linear activation function.

Results

Overall, our WT-specific models achieved good prediction performance, i.e. average AUC of more or almost 0.6 and average Pearson correlation of more or almost 0.5, in 10-fold CV (Figure S5). As explained before, the sub-library of each RBP was divided in to two sub-libraries based on its prefix. We trained a model specific for each of the two sub-libraries and tested it in 10-fold CV. We then choose the better performing model out of these two according to its average Pearson correlation in 10-fold CV, and used it in our downstream analysis. This resulted in using the C library for MCP and PCP, and the GC library for QCP.

Whole-library binding model

Padding sequences for whole-library models

We next developed a protein-specific binding model based on the whole library of RNA sequences and their responsiveness scores. Since the binding sites have different lengths, they need to be converted to have equal lengths for the learning process. All sequences were padded to the same length of 50nt. The binding sites were part of an RNA transcript. Hence, we upstream-padded them up to the transcription start site (TSS), with C or GC prefix according to their position. Downstream-padding of the sequences was done by their transcriptomic context up to a complete length of 50nt.

The upstream nucleotides are:

AATTGTGAGCGCTCACAATTATGATAGATTCAATTGGATTAATTAAAGAGGAGAAA
GGTACCCATG

The downstream nucleotides are:

GTGAGCAAGGGCGAGGAGGATAACATGGCCATCATCAAGGAGTTCATGCGCTTC
AAGGTGCACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGA
GGGCGAGGGCCGCCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGTGACCA
AGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTACG
GCTCCAAGGCCTACGTGAAGCACC

RNA secondary structure information

For the whole-library binding model, we augmented the one-hot encoded sequence information by RNA secondary structure information. We used RNAfold algorithm (Vienna package) to predict the structure of each sequence. The input to RNAfold is the binding site, and it outputs the predicted secondary structure in parenthesis notation, i.e. opening and closing parenthesis for base-pairs and a dot for unpaired nucleotide.

We converted this notation into an encoding of RNA structural contexts. This was done by a MATLAB script that encodes the RNA structure as a one-hot matrix with one bit set in each column for the corresponding structural context. For a binding site of length

n, the n-long parenthesis annotation is transformed to a 5 x n binary matrix. The structural contexts we used were lower stem (LS), bulge (B), upper stem (US), loop (L), and no-hairpin (N). The one-hot encoded structure matrix not on the binding site was set to zero. The RNA structure matrix was concatenated to the sequence matrix, (Figure 4A, main text). In total, for a sequence of length L, this results in a binary matrix of size (4+5) x L.

Model description and optimization

The model is composed of one convolution layer, one hidden layer and an output layer (Figure 5A, main text).

The optimization of the model was done in the same manner as described above. Briefly, 10 random parameters sets were tested, and the best preforming one was chosen followed by fine tuning.

Parameter space			Final parameters		
Parameter	Initial space	Surrounding space	MCP	QCP	PCP
Nodes	5-40	±5	20	20	20
Layers	1-3	-	1	1	1
Kernel length	4-10	±3	6	5	5
Kernel number	4-35	±5	18	18	11
Epochs	10,20..100	±15 (strides of 5)	40	40	40

Table S6. *Parameters search for whole-library models. The parameter space for each of the two steps of the hyper parameters search and the final models parameters. Unless noted otherwise, the range specified is of stride 1.*

In addition to the parameters in Table S, which are unique to each model, there are additional parameters that are common to all of them: learning rate 0.001, batch size 16, optimizer ADAM, loss function MSE (mean squared error), activation function for the convolution layer ‘relu’ and for the hidden layer ‘tanh’. The output layer consists of one node with a linear activation function.

Results

The prediction performance achieved by the whole-library models are similar to the WT-specific ones, i.e. average AUC of more than 0.6 and average Pearson correlation of around 0.5 for each of the three proteins (Figure S5). Although the WT-specific model achieved better results for two out of the three proteins, this whole-library model has the advantage that it can be applied to a binding site of any length, and not just that of the WT. This enables the prediction of binding of all three proteins to the same sequence set. To show case the contribution of RNA structure to our whole-library models, we trained and tested a model without the additional RNA structure information, and prediction performance degraded with the removal of RNA structural information (Figure S6).

Generation of sequences for experimental validation

To test the predicted binding cassette generated according to our models’ predictions, we created one million synthetic binding sites. We generated one million random sequences that are in hamming distance of 3-7 from one of the WT binding sites. Overall, we randomly selected the one million out of 1.5 billion options. Because the number of possible variants rises as the length of the sequence, uniform selection of sequences will result in more variants of the long WT (PCP, 25-nt long) and less variants of the short WT (MCP, 19-nt). To overcome this bias, we divided the random selection into three parts; in each part we randomly selected 333,333 sequences from the variants of one WT. We computed the binding intensity of each of the proteins to

the set of one million sequences using the whole-library models. Then, to experimentally validate model accuracy, we chose a sample out of the one million. We selected ten sequences that are single binders (i.e. bound by a single protein and not by the two others), and ten that are double binders (i.e. bound by two proteins and not by the third). All are in edit distance of at least 4 from one another and all were not included in the original experimental library.

DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTARY TABLES TITLES:

1. Table S1: Oligo library sequences
2. Table S2: DNA oligos
3. Table S3: RNA Binding Proteins sequences
4. Table S4: RNA binding site cassettes sequences
5. Table S5: Parameters search space for WT-specific model (in text)
6. Table S6: Parameters search for whole-library models (in text)

REFERENCES

Biswas, J., Rahman, R., Gupta, V., Rosbash, M., and Singer, R.H. (2019). MS2-TRIBE evaluates protein-RNA interactions and nuclear organization of transcription by RNA editing. BioRxiv 829606.

Buenrostro, J.D., Chircus, L.M., Araya, C.L., 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 for mapping biophysical and evolutionary landscapes. Nat Biotechnol *32*, 562–568.

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

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., and Gregor, T. (2018). Dynamic interplay between enhancer-promoter topology and gene activity. Nat Genet *50*, 1296–1303.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat Rev Genet *20*, 389–403.

Gasperini, M., Starita, L., and Shendure, J. (2016). The power of multiplexed functional analysis of genetic variants. Nature Protocols *11*, 1782–1787.

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.

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.

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 *9*, 93-106.e8.

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

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.

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

Koodli, R.V., Keep, B., Coppess, K.R., Portela, F., Participants, E., and Das, R. (2019). EternaBrain: Automated RNA design through move sets and strategies from an Internet-scale RNA videogame. PLOS Computational Biology *15*, e1007059.

Langan, R.A., Boyken, S.E., Ng, A.H., Samson, J.A., Dods, G., Westbrook, A.M., Nguyen, T.H., Lajoie, M.J., Chen, Z., Berger, S., et al. (2019). De novo design of bioactive protein switches. Nature *572*, 205–210.

Lari, A., Farzam, F., Bensidoun, P., Oeffinger, M., Zenklusen, D., Grunwald, D., and Montpetit, B. (2019). Live-Cell Imaging of mRNP-NPC Interactions in Budding Yeast. Methods Mol. Biol. *2038*, 131–150.

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

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 σ54 Promoters. Cell Reports *21*, 845–858.

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. (1996a). The RNA-binding Site of Bacteriophage Qβ Coat Protein. J. Biol. Chem. *271*, 31839–31845.

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

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

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

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. Algorithms for Molecular Biology *6*, 26.

Miao, Z., Adamiak, R.W., Blanchet, M.-F., Boniecki, M., Bujnicki, J.M., Chen, S.-J., Cheng, C., Chojnowski, G., Chou, F.-C., Cordero, P., et al. (2015). RNA-Puzzles

Round II: assessment of RNA structure prediction programs applied to three large RNA structures. RNA *21*, 1066–1084.

Nielsen, A.A.K., Der, B.S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E.A., Ross, D., Densmore, D., and Voigt, C.A. (2016). Genetic circuit design automation. Science *352*, aac7341.

Park, Y., and Kellis, M. (2015). Deep learning for regulatory genomics. Nature Biotechnology *33*, 825–826.

Peabody, D.S. (1990). 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.

Reis, A.C., Halper, S.M., Vezeau, G.E., Cetnar, D.P., Hossain, A., Clauer, P.R., and Salis, H.M. (2019). Simultaneous repression of multiple bacterial genes using nonrepetitive extra-long sgRNA arrays. Nat Biotechnol 1–8.

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

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.

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

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

Tutucci, E., Vera, M., Biswas, J., Garcia, J., Parker, R., and Singer, R.H. (2018). An improved MS2 system for accurate reporting of the mRNA life cycle. Nature Methods *15*, 81–89.

Urbanek, M.O., Galka-Marciniak, P., Olejniczak, M., and Krzyzosiak, W.J. (2014). RNA imaging in living cells – methods and applications. RNA Biol *11*, 1083–1095.

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

Wainberg, M., Merico, D., Delong, A., and Frey, B.J. (2018). Deep learning in biomedicine. Nature Biotechnology *36*, 829–838.

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

Wetmore, K.M., Price, M.N., Waters, R.J., Lamson, J.S., He, J., Hoover, C.A., Blow, M.J., Bristow, J., Butland, G., Arkin, A.P., et al. (2015). Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. MBio *6*, e00306-15.

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

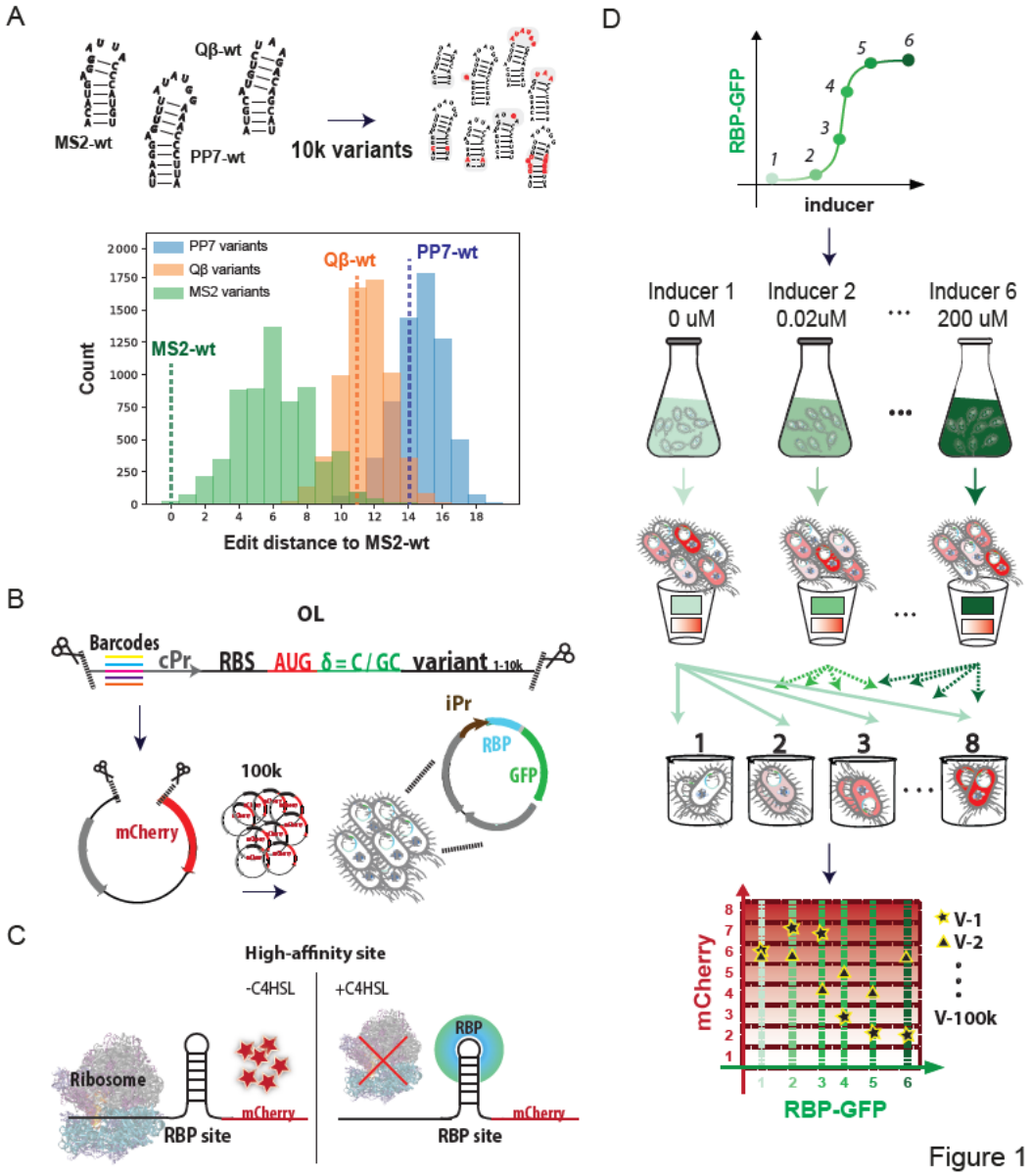
A
MS2-wt
Qβ-wt
PP7-wt
10k variants
PP7 variants
Qβ variants
MS2 variants
Qβ-wt
PP7-wt
MS2-wt
Count
Edit distance to MS2-wt
B
OL
Barcodes
cPr
RBS
AUG
δ = C / GC
variant 1-10k
iPr
RBP
GFP
mCherry
100k
C
High-affinity site
-C4HSL
+C4HSL
Ribosome
RBP
mCherry
RBP site
mCherry
RBP site
D
RBP-GFP
inducer
Inducer 1
0 uM
Inducer 2
0.02uM
Inducer 6
200 uM
1
2
3
8
mCherry
RBP-GFP
V-1
V-2
V-100k

Figure 1

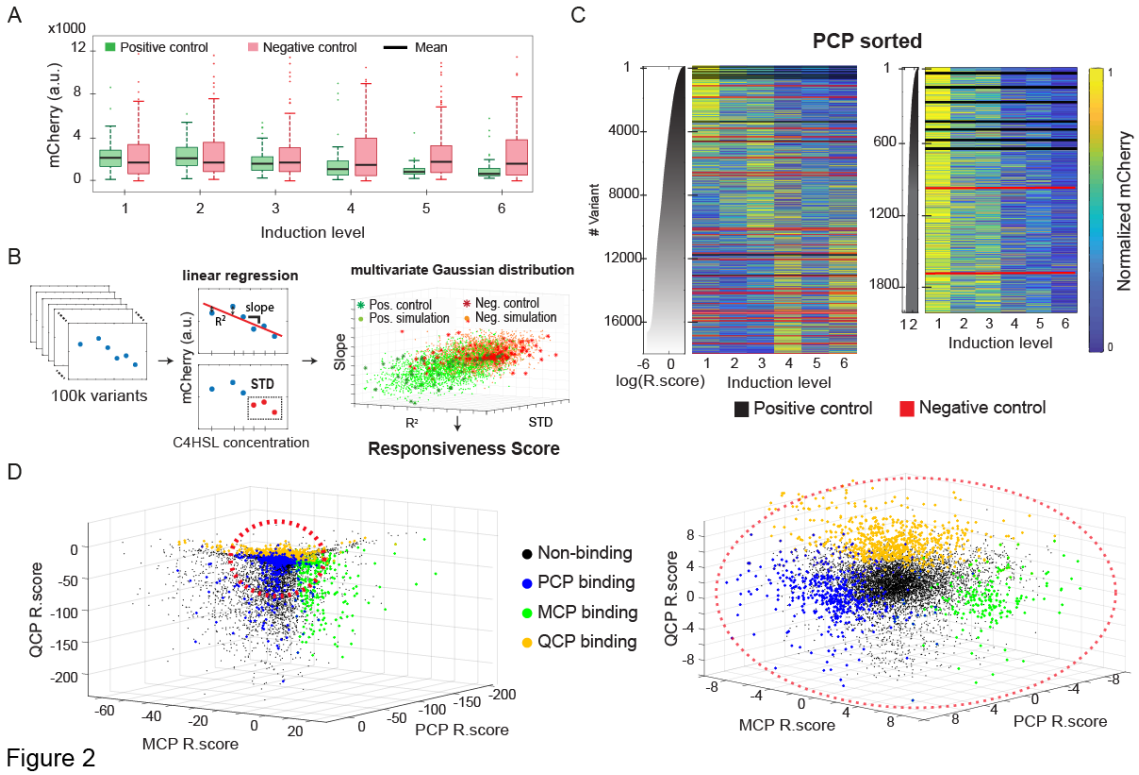
A
x1000
Positive control
Negative control
Mean
mCherry (a.u.)
Induction level
B
100k variants
linear regression
slope
R²
STD
mCherry (a.u.)
C4HSL concentration
multivariate Gaussian distribution
Pos. control
Pos. simulation
Neg. control
Neg. simulation
Slope
R²
STD
Responsiveness Score
C
PCP sorted
# Variant
log(R.score)
Induction level
Normalized mCherry
Positive control
Negative control
D
QCP R.score
MCP R.score
PCP R.score
Non-binding
PCP binding
MCP binding
QCP binding

Figure 2

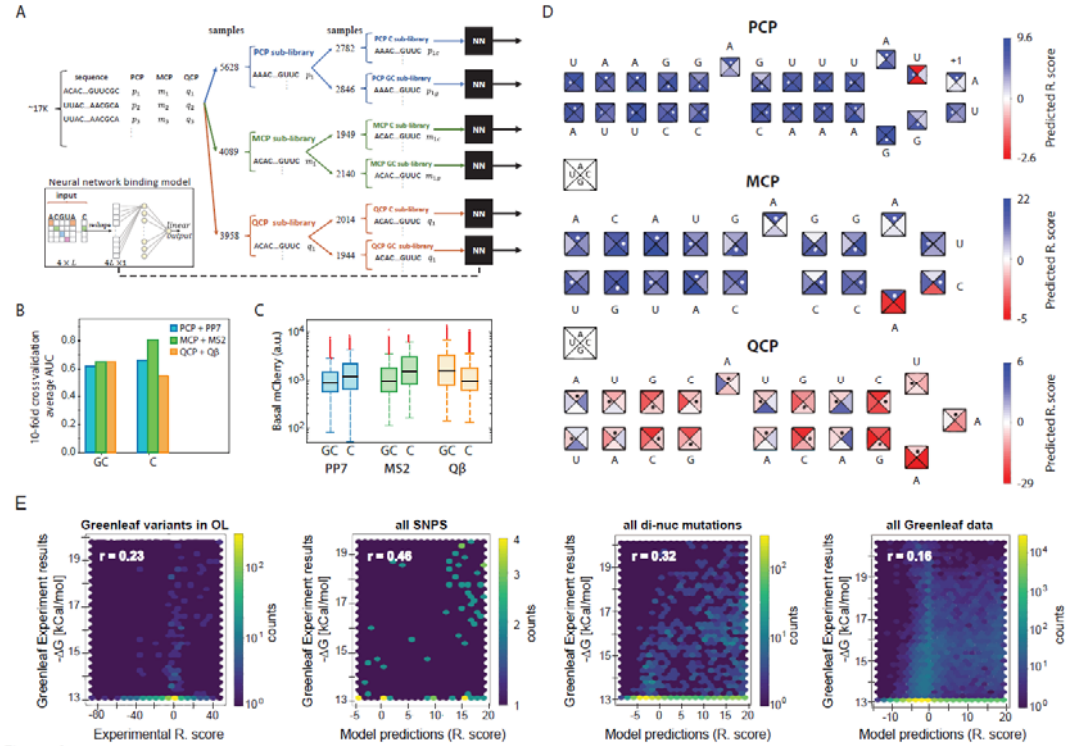
A
samples
samples
PCP sub-library
MCP sub-library
QCP sub-library
NN
Neural network binding model
B
10-fold cross validation average AUC
PCP + PP7
MCP + MS2
QCP + Qβ
GC
C
C
Basal mCherry (a.u.)
GC C
PP7
MS2
Qβ
D
PCP
MCP
QCP
Predicted R. score
E
Greenleaf variants in OL
r = 0.23
all SNPS
r = 0.46
all di-nuc mutations
r = 0.32
all Greenleaf data
r = 0.16
Greenleaf Experiment results -ΔG [kCal/mol]
counts
Experimental R. score
Model predictions (R. score)

Figure 3

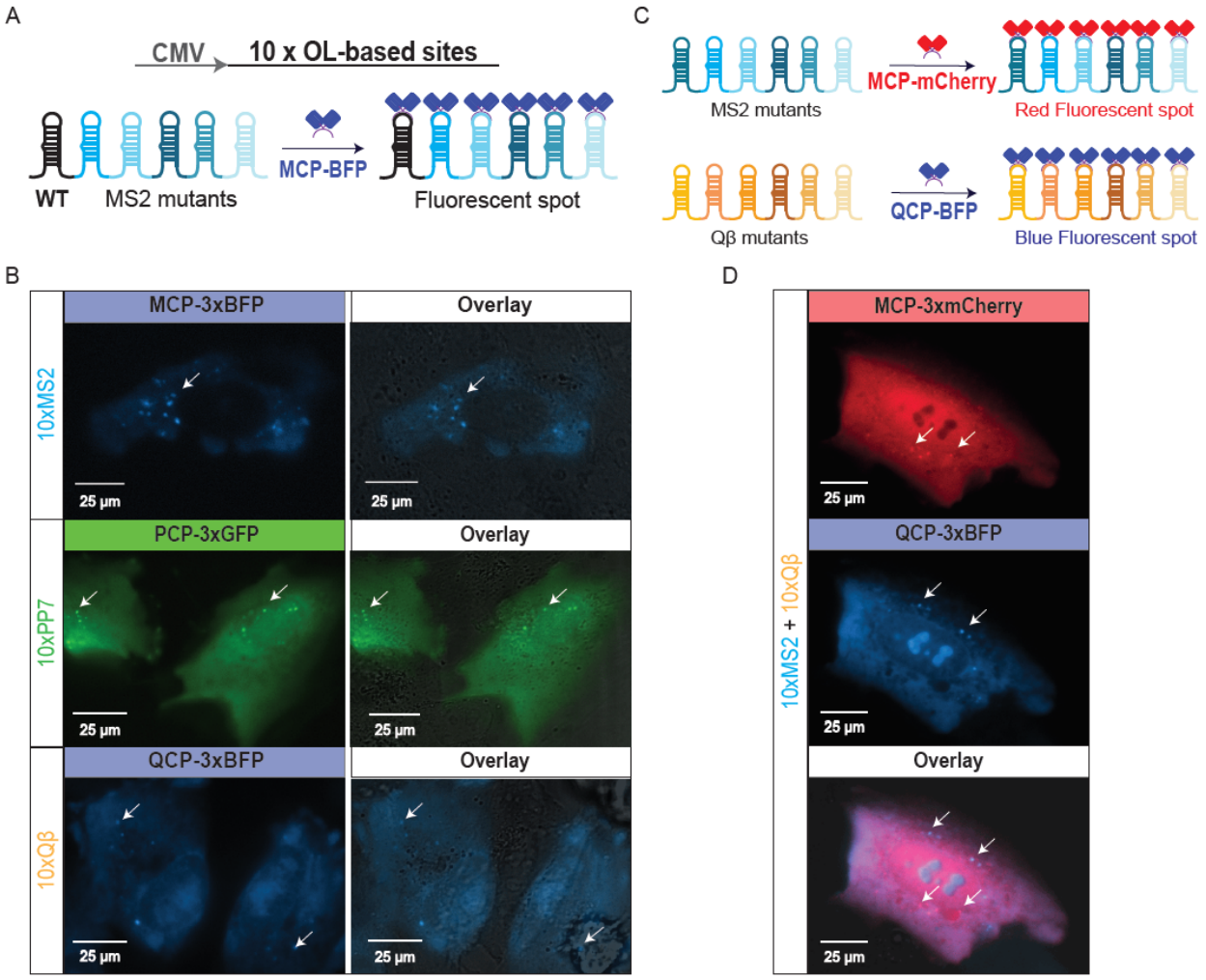
A
CMV
10 x OL-based sites
WT
MS2 mutants
MCP-BFP
Fluorescent spot
C
MS2 mutants
MCP-mCherry
Red Fluorescent spot
Qβ mutants
QCP-BFP
Blue Fluorescent spot
B
10xMS2
MCP-3xBFP
Overlay
10xPP7
PCP-3xGFP
Overlay
10xQβ
QCP-3xBFP
Overlay
25 µm
D
10xMS2 + 10xQβ
MCP-3xmCherry
QCP-3xBFP
Overlay
25 µm

Figure 4

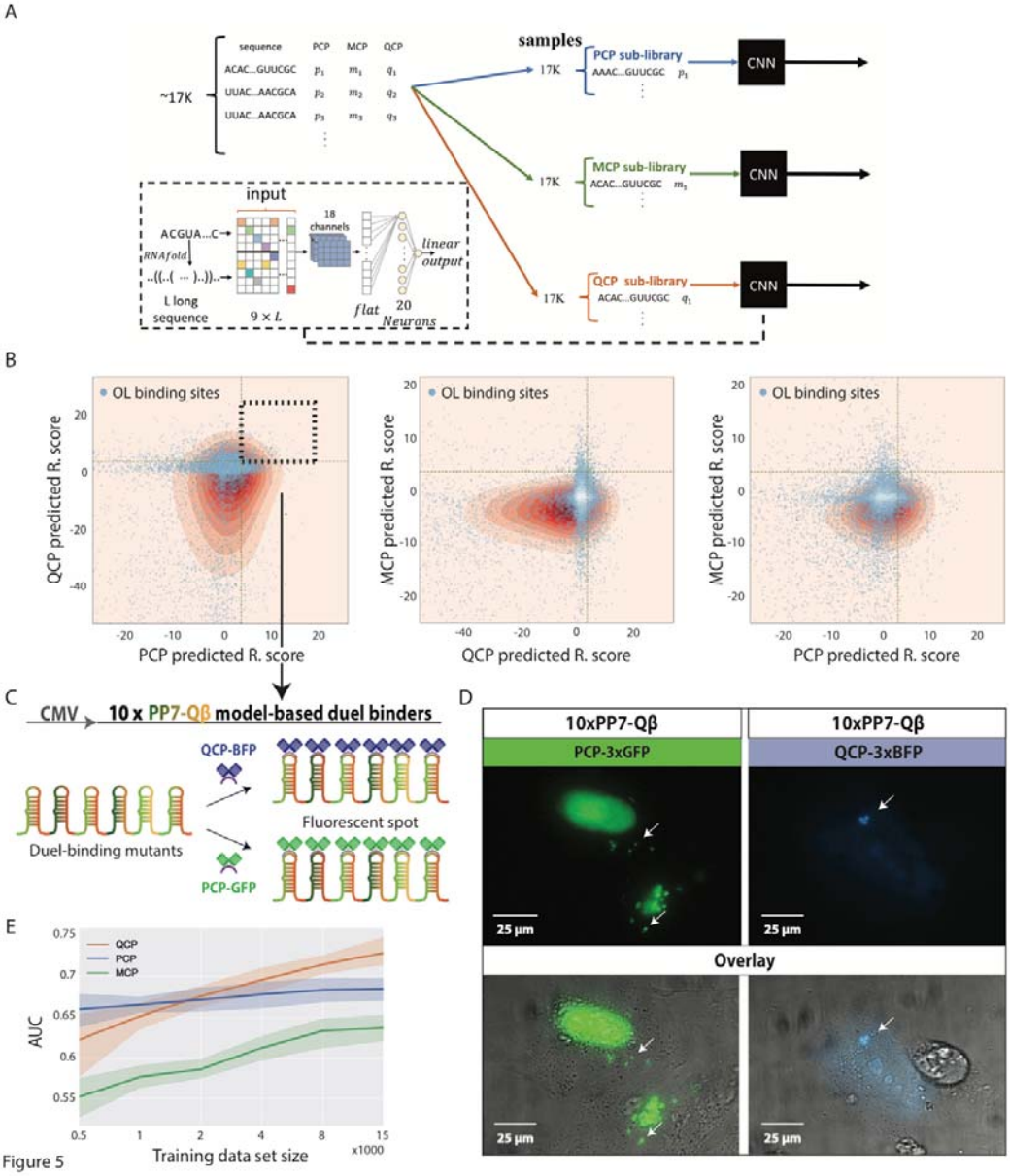
A
samples
~17K
sequence
PCP
MCP
QCP
ACAC...GUUCGC
UUAC...AACGCA
PCP sub-library
MCP sub-library
QCP sub-library
17K
CNN
input
ACGUA...C
RNAfold
L long sequence
9 × L
18 channels
flat
20 Neurons
linear output
B
OL binding sites
QCP predicted R. score
PCP predicted R. score
MCP predicted R. score
C
CMV
10 x PP7-Qβ model-based duel binders
QCP-BFP
PCP-GFP
Duel-binding mutants
Fluorescent spot
D
10xPP7-Qβ
PCP-3xGFP
QCP-3xBFP
Overlay
25 µm
E
QCP
PCP
MCP
AUC
Training data set size
x1000

Figure 5