nature chemical biology

ARTICLE

PUBLISHED ONLINE: 25 MARCH 2012 | DOI: 10.1038/NCHEMBIO.919

Rationally designed families of orthogonal RNA regulators of translation

Vivek K Mutalik[1-3,7,8], Lei Qi[4,8], Joao C Guimaraes[4,5], Julius B Lucks[4,6,7] & Adam P Arkin[1-4]*

Our ability to routinely engineer genetic networks for applications is limited by the scarcity of highly specific and non-cross-reacting (orthogonal) gene regulators with predictable behavior. Though antisense RNAs are attractive contenders for this purpose, quantitative understanding of their specificity and sequence-function relationship sufficient for their design has been limited. Here, we use rationally designed variants of the RNA-IN-RNA-OUT antisense RNA–mediated translation system from the insertion sequence IS10 to quantify >500 RNA-RNA interactions in *Escherichia coli* and integrate the data set with sequence-activity modeling to identify the thermodynamic stability of the duplex and the seed region as the key determinants of specificity. Applying this model, we predict the performance of an additional ~2,600 antisense-regulator pairs, forecast the possibility of large families of orthogonal mutants, and forward engineer and experimentally validate two RNA pairs orthogonal to an existing group of five from the training data set. We discuss the potential use of these regulators in next-generation synthetic biology applications.

Precise control of gene expression is at the core of any genetic engineering–dependent discipline. For most synthetic biology applications, it is necessary to express, at controlled levels, multiple genes possibly responsive to multiple internal and external signals. In a practical sense, the ability to rationally regulate the expression of multiple genes aids in optimization of biosynthetic pathways for industrial chemical production[1]. Elements that allow controlled exploration of different pathway-enzyme stoichiometries to maximize productivity or enable design of regulatory circuitry that balances enzyme activity to minimize toxic intermediates are valuable components of the metabolic engineer's toolbox[1]. These elements are even more critical in emerging applications, such as cell-based therapies and active biomaterials, that require more sophisticated regulatory circuitry implementing complex logic functions and memory[2]. It has been argued that a large compendium of diverse regulators that are orthogonal (that is, do not inadvertently cross-react), homogeneous (operate with similar kinetics, thermodynamics and other structural properties) and have predictable functionality is necessary to enable increasingly complex genetic circuit design[3–6].

Though there has been some success in mining such regulatory elements, most commonly promoters, from the vast library of function available in nature, each individual instance must itself be characterized in a specific context and needs further part and strain engineering to make it operate as desired in a given application (for example, to match a desired dynamic range or not to interfere with host machinery). An alternative approach is to engineer part variants from a common ancestor (part) that differ in designed ways from otherwise homogeneous operation. There has been an increasing effort to engineer such part libraries. Most efforts have focused on engineering promoters[7,8] and ribosome-binding sites[9] with desired activities. Models to predict the activities of new variants of these have shown some success[9,10]. Alternatively, there have been a number of efforts to engineer parts libraries that regulate different aspects of gene expression, such as transcription initiation with families of transcription factors[5], transcriptional elongation via antisense RNA–regulated attenuators[11] and leader-sequence expression regulated by unnatural amino acids[12]; translation initiation using orthogonal ribosomes[6,13], riboswitches[14] and antisense RNAs[15]; and mRNA stability via engineered ribozymes[16,17], stability hairpins[18], RNase sites[19,20] and protein-sensing RNA devices[21,22]. Though these are foundational first steps, only a few efforts have attempted to create models that link sequence to activity to allow forward design of new family members[9,13,16].

Regulatory mechanisms that are mediated by RNA molecules, however, might be expected to be among the most amenable to such modeling. The apparent simplicity in their base-pairing rules and the modularity of their structural components make them excellent substrates for design and attractive contenders for a standard platform for gene expression regulation in synthetic biology applications[11,15–17,20,23,24]. They also provide a powerful basis set of functions for gene expression engineering owing to their ability to sense biomolecules; regulate transcriptional elongation, translation initiation and mRNA degradation; and act in *cis* when arrayed on the same transcript or in *trans* to allow the construction of regulatory networks[11,15–17,20,23–25].

Though the versatility of RNA molecules has been recognized, there have been few studies on designing an entirely artificial (synthetic) riboregulator (cognate sense-antisense RNA) family or modifying a naturally occurring mechanism for use in gene silencing or activation purposes in *E. coli*[11,15,26–28]. In a seminal study[15], simple base-pairing design rules were used to create a family of two orthogonal sense-antisense RNA partners that regulate translation. The interaction between cognate pairs showed a positive correlation with RNA-RNA binding free energies, implying a thermodynamically driven, one-step pathway to a stable duplex. Recently, we reported a set of three orthogonal mutants of a natural antisense RNA–mediated transcriptional attenuation system and found a

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. [2]Joint Bioenergy Institute, Emeryville, California, USA. [3]QB3–California Institute for Quantitative Sciences, Berkeley, California, USA. [4]Department of Bioengineering, University of California, Berkeley, California, USA. [5]Department of Informatics–Computer Science and Technology Center, University of Minho, Campus de Gualtar, Braga, Portugal. [6]Miller Institute for Basic Scientific Research, Berkeley, California, USA. [7]Present addresses: BioFAB, Joint BioEnergy Institute, Emeryville, California, USA (V.K.M.); School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York, USA (J.B.L.). [8]These authors contributed equally to this work.
*e-mail: aparkin@lbl.gov

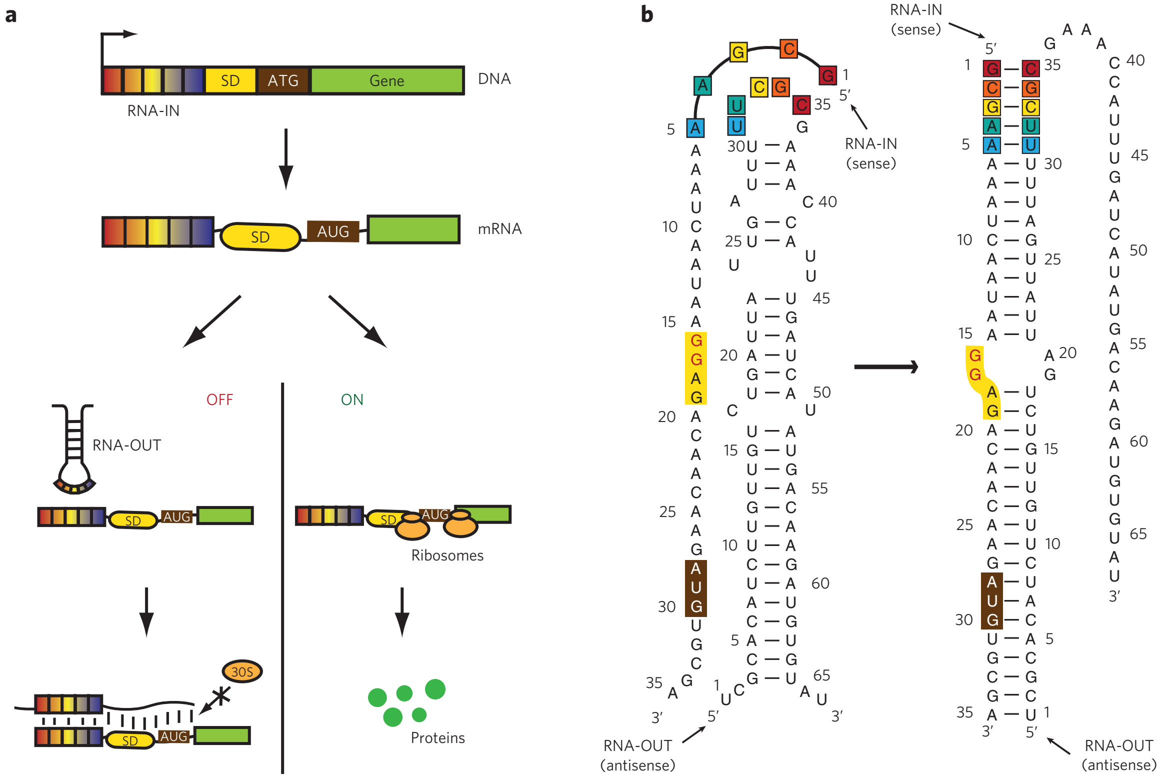

Figure 1 | Schematic of sense RNA-IN and antisense RNA-OUT interaction. (**a**) Schematic of antisense RNA–regulated translation repression of the RNA-IN-SFGFP translational fusion. Antisense RNA (RNA-OUT, indicated by stem-loop structure), binds specifically to RNA-IN-SFGFP mRNA, which blocks the Shine-Dalgarno (SD) sequence and AUG start codon, thereby inhibiting its translation. Molecular recognition and initial base pairing between sense and antisense RNA is indicated by colored boxes. In the absence of RNA-OUT, translation of RNA-IN-SFGFP mRNA proceeds normally to produce SFGFP. (**b**) RNA-OUT–RNA-IN hybridization. The entire region of RNA-OUT[35] used in the current work is not shown for clarity (**Supplementary Data Set 2**). Colored boxes indicate the core recognition region between interacting species. The mutated SD region in RNA-IN mRNA is marked as yellow, and the AUG start codon is colored brown.

poor correlation between target repression and the thermodynamic stability of the duplex, suggesting that other factors besides thermodynamics may be at play[11]. These results are consistent with earlier observations that the specificity and efficiency shown by antisense RNA control systems may be explained by interactions between different structural motifs or modules and not solely by complementarity in base-pairing (or hybridization free energies)[29,30]. Notably, in both of these instances, the strategy for creating orthogonal mutants was based on the design features observed in other similar systems, and mutants were more or less handcrafted.

If scalable design of expression with RNA regulators is to succeed, quantitative sequence-function relationships that elucidate the mechanisms behind orthogonally acting RNA regulators and thus enable their more rational design are needed. Sequence-activity relationships have proven useful in the annotation of genomes, making it possible to infer regulatory sites and other such features[31]. Similar quantitative studies have provided mechanistic insights into the operation of eukaryotic microRNAs (miRNAs) and small interfering RNAs in silencing target mRNAs and their off-target cross-talk[32,33]. In these cases, statistical modeling has been used to mine large compendiums of experimental data to uncover specificity determinants for RNA-RNA interaction[32,33]. These advances have aided the implementation of complex synthetic programs that could have an immense therapeutic value[34].

Here we present a minimal-complexity, maximally predictive data-driven sequence-activity model from quantitative assays of a mutant library derived from a well-known antisense RNA–mediated translation control system (RNA-IN–RNA-OUT (RNA-IN/OUT) system[35] of IS10) in *E. coli*. By model integration of hundreds of *in vivo* reporter assays, we identify the key determinants of antisense RNA interaction specificity. We then use the model to predict the performance of additional regulator pairs and experimentally validate these predictions by forward designing new orthogonal mutant pairs. Overall, this study demonstrates the utility of hypothesis-driven library construction to parameterize predictive models of regulatory element function. These models yield insight into the critical mechanisms of the activity of these elements and provide computer-aided design tools for the production of new elements of the antisense RNA family.

RESULTS

A minimal assay system for quantifying orthogonal mutants

Our translational regulators are derived from the copy-number control element from the insertion sequence IS10, wherein an antisense RNA (RNA-OUT) inhibits transposase expression[35]. RNA-OUT base pairs to the translation initiation region of the transposase mRNA (RNA-IN), thereby repressing translation both by preventing ribosome binding[36] and by promoting transcript degradation[37] (**Fig. 1a**). The 5′ end of the unstructured, unstable sense RNA-IN is complementary to the top of the loop domain and to one entire side of the stable RNA-OUT hairpin (**Fig. 1b**). Earlier studies have suggested that the 5 base pairs (bp) in the 5′ end of RNA-IN and in the loop domain of RNA-OUT determine the initiation of RNA duplex formation[35,38]. The loop domain of RNA-OUT contains a pyrimidine-uracil-nucleotide-purine (YUNR) motif and is predicted to promote specificity and rapid duplex formation with RNA-IN[39]. The first 3 bp between RNA-IN and RNA-OUT are G-C pairs, and the strength of hybridization free energy in this G-C–rich region seems to be critical for effective antisense interaction and molecular specificity[35]. We reasoned, therefore, that these

Table 1 | Selected sense and antisense mutants

Sense	Sequence	Antisense	Sequence	Nucleotide swaps	Tested/ possible
S1	GCGAA	A1	UUCGC	Wild type	1/1
S3	GGGAA	A3	UUCCC	1	4/5
S4	GCCAA	A4	UUGGC		
S5	GCGUA	A5	UACGC		
S6	GCGAU	A6	AUCGC		
S7	CGGAA	A7	UUCCG	2	2/10
S8	CCCAA	A8	UUGGG		
S17	CGCAA	A17	UUGCG	3	6/10
S19	CGGAU	A19	AUCCG		
S21	CCCAU	A21	AUGGG		
S22	CCGUU	A22	AACGG		
S23	GGCUA	A23	UAGCC		
S26	GCCUU	A26	AAGGC		
S27	CGCUA	A27	UAGCG	4	5/5
S28	CGCAU	A28	AUGCG		
S29	CGGUU	A29	AACCG		
S30	CCCUU	A30	AAGGG		
S31	GGCUU	A31	AAGCC		
S32	CGCUU	A32	AAGCG	5	1/1
S34	GGGUAAA	A34	UUUACCC	Extra 2[a]	4/24
S43	CCCAUAA	A43	UUAUGGG		
S49	CCCCGAA	A49	UUCGGGG		
S52	GGGGCAA	A52	UUGCCCC		

The recognition motif (5′ to 3′) of 23 sense RNA-IN and 23 antisense RNA-OUT RNAs are shown with mutated or inserted positions from wild-type species denoted by underlining. Total number of possible mutants from a particular base swap and the number of chosen mutants for experimental characterization are also shown. [a]Two extra nucleotides inserted within the core region of RNA-IN and RNA-OUT.

specificity-determining interactions could be manipulated to create families of mutually orthogonal variants of the native system. We use the term orthogonal family to describe a group (more than two members) of sense and antisense mutants that specifically interact with their cognate partners and show minimal interaction with their noncognate counterparts.

To simplify the engineering of the RNA-IN/OUT system, we identified a minimized and slightly modified regulatory region that is sufficient for >85% repression of the target (details are in **Supplementary Methods**). To assess the performance of the RNA-IN/OUT pair, we measured the percentage repression of RNA-IN–mediated super-folder green fluorescent protein (SFGFP)[40] fluorescence (constitutively expressed from a low-copy plasmid) in the presence of RNA-OUT (expressed from a high-copy plasmid) in *E. coli* TOP10 cells during exponential growth. We observed graded tuning of target repression at different induction levels of antisense RNA and about 90% repression of SFGFP fluorescence when RNA-OUT was fully induced (**Supplementary Results**). As this result agrees with previously reported data corresponding to much longer RNA-IN and RNA-OUT regions (used in conjunction with their endogenous promoters and a lacZ reporter system)[35,38], we conclude that our minimized system retains the desired activity. We confirmed that our minimized system is sensitive to changes in antisense and sense specificity by examining a reported specificity-altering mutation[35] (**Supplementary Fig. 4**).

Design of the mutant library

To engineer mutually orthogonal sense-antisense pairs, we considered complementary mutations at the 5 nucleotides (nt)—in all combinations—in the 5′ specificity region of RNA-IN and at the corresponding nucleotides in the loop of RNA-OUT (**Fig. 1b** and **Table 1**). This led to a set of 32 variants in sense RNA-IN and antisense RNA-OUT. We reasoned that the possible number of orthogonal pairs could also be increased by inserting nucleotides within the recognition motif of this system, thereby 'scaling up' the RNA-RNA interaction region. We therefore considered insertion of two extra nucleotides, A-T, G-C, T-A or C-G, between position 3 and 4 of RNA-IN (corresponding to complementary nucleotides at positions 32 and 33 of RNA-OUT; **Fig. 1b** and **Table 1**). We also hypothesized that compensatory mutations in the first 3 bp of the interaction region in these scaled-up mutants would extend the number of orthogonal pairs and possibly improve regulatory efficiency. This resulted in 24 additional RNA-IN/OUT paired mutations for a total library size of 56 (**Table 1** and **Supplementary Table 1**). This number may be further increased by considering all possible combinations of (single, double and so on) nucleotide insertions with different combinations at the first 3 bp.

Earlier studies have suggested that the RNA-IN/OUT interaction is thermodynamically favored over strand exchange from the imperfectly base-paired RNA-OUT hairpin to the perfectly base-paired RNA-IN/OUT duplex[41]; furthermore, our rationally constrained library of RNA-IN/OUT pairs is composed of mutants that have a 5-bp variable sequence region surrounded by a common flanking sequence (**Table 1** and **Supplementary Table 1**). We thus assumed that the specificity of interaction and repression efficiency in our library of mutants can be explained, to a large extent, by differences in their hybridization free energies. The cognate pairs would be expected to have lower hybridization energy than noncognate pairs. To predict to the first order which pairs in our virtual library would show the highest specificity of interaction and the lowest cross-talk with other members, we estimated the hybridization free energies using UNAfold software[42] for all 56 sense-antisense pairs in the library (total 3,136 interactions). As we expected, we found that the cognate sense-antisense partners, which formed as a group along the diagonal in **Supplementary Figure 5**, showed far more stable hybrids compared to noncognate partners. To maximize the chance of mutual orthogonality, we selected 23 candidates from the total of 56 library members via a clustering procedure (**Fig. 2a**, **Table 1** and **Supplementary Fig. 6**). That only 5 out of the 23 RNA-OUT mutants conserve the YUNR motif (**Supplementary Table 7**) also gave us a chance to test the importance of this motif in the functioning of the RNA-IN/OUT system.

Measurement and analysis of the mutant library

We generated the 23 RNA-IN and RNA-OUT mutant constructs, and each of the 529 possible pairs was cotransformed on separate plasmids into *E. coli*. Additionally, all RNA-IN plasmids were cotransformed with a nonsense (nonfunctional) RNA-OUT plasmid as a negative control (described in Methods and **Supplementary Methods**). The performance of RNA-IN/OUT pairs was quantified by measuring the fluorescence during the exponential phase of each strain, and percentage repression was calculated. The matrix of percentage repression for the 529 combinations of sense and antisense mutants is shown as a heat map in **Figure 2b** (additional data is in **Supplementary Results**). Most cognate sense-antisense pairs (along the diagonal on the heat map in **Fig. 2b**) show strong repression (>80%). Approximately 5% of all pairs achieve more than 70% repression, whereas about 75% of total pairs show less than 20% repression (**Supplementary Fig. 10**). Overall, we observed a wide range of percentage target repression, ranging from negligible repression (<5%) to over 90% repression, and the change in dynamic range (ratio between target expression in absence of antisense RNA and maximal antisense RNA) varied from zero to ten-fold (**Supplementary Tables 5** and **6**, respectively). We also found many examples of a single antisense

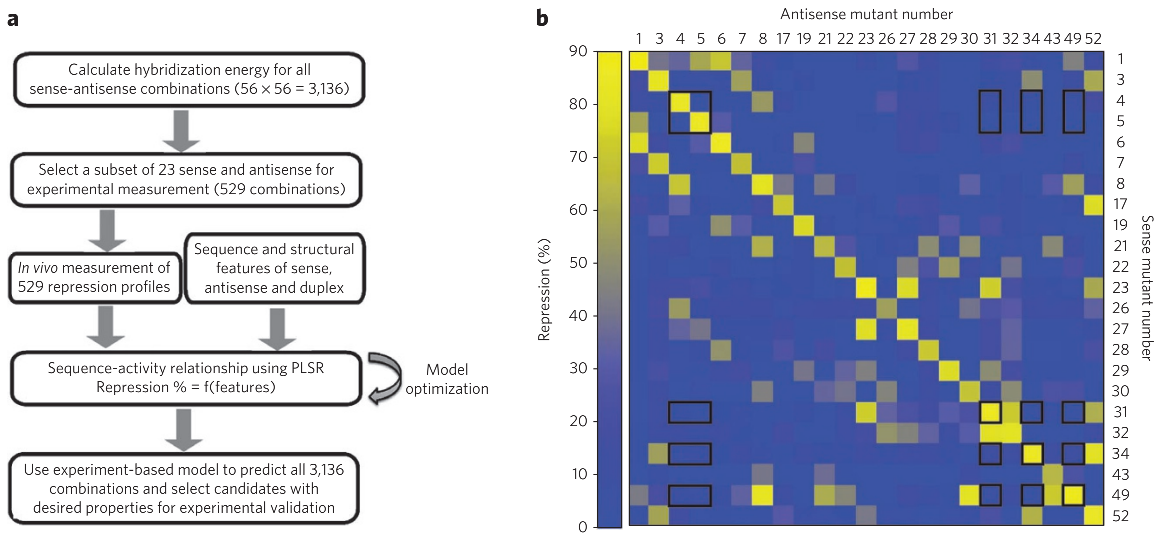

Figure 2 | A rationally designed library for finding orthogonal mutants in the RNA-IN/OUT system. (**a**) The quantitative framework to describe RNA-RNA interactions by integrating *in vivo* reporter assay data with sequence-activity modeling to identify specificity determinants and forward engineering of orthogonal families of translation regulators. The equation, Repression % = f(features), defines percentage repression as a linear function of different structural and thermodynamic features of sense, antisense and duplex RNAs (Methods). (**b**) Heat map of percentage repression profile of 23 RNA-IN mutants in presence of 23 antisense RNA-OUT mutants (total 529 data points; **Supplementary Table 5**). Cognate pairs are arranged diagonally and show maximum repression. Five mutually orthogonal pairs are shown as black boxes.

RNA repressing multiple sense targets and single sense targets being recognized by multiple antisense RNAs (**Supplementary Fig. 11**). One notable result is that more than 70% of cognate pairs showed repression higher than 75% and did not have a YUNR motif in the antisense RNA species, indicating that the YUNR motif is dispensable for the proper functioning of this system (**Supplementary Table 7**).

To determine how the energetics of the sense-antisense RNA interaction correlates with the experimental percentage repression, we plotted the calculated hybridization free energy for all 529 interactions against the experimental percentage repression (**Fig. 3a**). We observed that hybridization of both cognate and noncognate pairs with a free energy more than −41 kcal mol^{-1} is not active in repression, whereas most hybridizations with a free energy less than −46 kcal mol^{-1} showed stronger repression (closer to 85%). These results indicate that there is a critical threshold free energy needed for the propagation of the initial pairing interaction leading to a stable duplex formation, which causes efficient repression of target mRNA. Similar results have been reported for interaction of miRNAs with their targets in HeLa cell lines[43]. We did not observe a strong correlation between target repression and the accessibility of the recognition motif (unfolding free energy of sense and antisense RNA). This indicates that the applied mutational strategy may interfere minimally with the structure of both interacting RNAs.

Validating the orthogonality of mutant pairs

Using the experimentally determined percentage repression data to quantify target and nontarget specificity, we can identify groups or families of RNA-IN/OUT variants expected to function orthogonally when placed in the same cell. Further, identifying noncognate partners that show substantial cross-talk aids in determining base-pairing features that impart promiscuity. Thus, the definition of mutual orthogonality depends on thresholds of repression and cross-reactivity percentages for cognate and noncognate pairs, respectively, that we deem acceptable for a specific application.

The total number of observed mutants for different family sizes demonstrating an 80% repression threshold and 10% or 20% cross-reactivity with other members of the family (and orthogonal family) is shown in **Figure 3b**. At an 80% repression threshold and 10% cross-reactivity, we have more than ten families of mutually orthogonal pairs and triplets and one family of four orthogonal mutants, whereas at 20% cross-reactivity, we have more than 20 families made up of two, three and four mutually orthogonal mutants and five families of five mutants (**Supplementary Table 9**). A more detailed depiction of the number of mutants at different thresholds of percentage repression and cross-reactivity are shown in **Supplementary Figure 12**.

To test whether the orthogonality and repression profile of these mutants is retained with a sequence-divergent gene of interest, we fused five mutually orthogonal sense mutants (shown in **Fig. 3c**) to the fluorescent protein mRFP1 (52% sequence identity to SFGFP) and assayed them in the presence of corresponding antisense RNAs. The observed percentage repressions were quantitatively equivalent to that observed using SFGFP, demonstrating the modularity of the sense region and the efficiency of antisense RNA (**Supplementary Fig. 13**). To demonstrate the mutual orthogonality among members of an orthogonal family in the same cell, we picked five sense-antisense pairs (shown in **Fig. 3c**) and characterized every combination of two pairs in a single cell (described in Methods and **Supplementary Fig. 14**). Here, the sense partner of each pair was translationally fused to either SFGFP or mRFP1, and repression was quantified in the presence of different combinations of the cognate antisense RNAs expressed from a different plasmid (**Supplementary Methods** and **Supplementary Fig. 14**). The results support the possibility that our library produced a large number of mutually orthogonal and modular regulatory variants that retain their specificity characteristics within the same cell (**Supplementary Figs. 15** and **16**).

Selection of a sequence-function relationship model

The correlation between the hybridization free energy and percentage repression suggests that free energy is a good, though not perfect, predictor of interaction specificity (**Fig. 3a**). To find other features that determine the specificity of interaction between RNA-IN and RNA-OUT, we pursued modeling of the sequence-function relationship for the *in vivo* experimental data set.

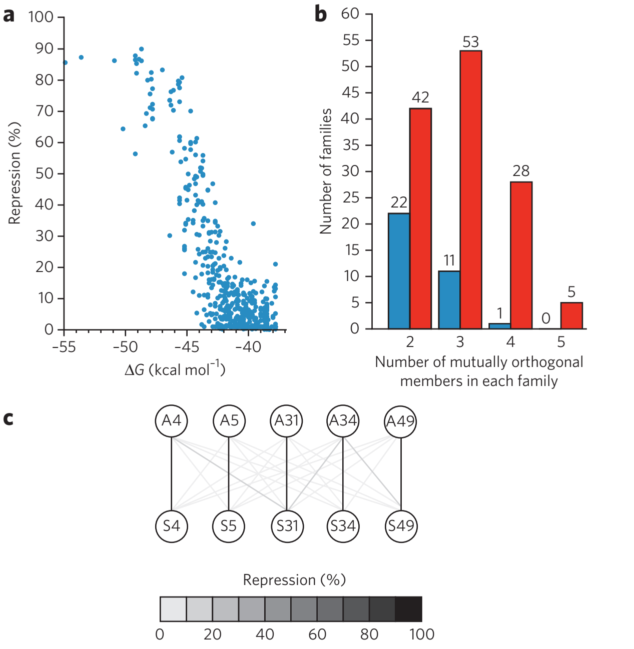

Figure 3 | Repression characteristics and estimation of total number of orthogonal pairs in the experimental data set. (**a**) The scatter plot of calculated hybridization free energy (Δ*G* kcal mol^{-1}) as a function of experimental percentage repression. (**b**) Estimation of the number of families made up of two, three, four and five mutually orthogonal members at different thresholds of percentage repression and cross-talk. Two representative data sets are shown here: data for ≥80% repression and ≤10% cross-talk (blue bars) and data for ≥80% repression and ≤20% cross-talk (red bars). (**c**) The observed network of interaction between mutually orthogonal cognate and noncognate sense and antisense RNAs (mutants 4, 5, 31, 34 and 49). The shading of links indicates the percentage repression (black for 100% repression and light gray for almost no repression).

From an inspection of the predicted sense and antisense RNA secondary structures and the form of the duplex, we selected a short list of 31 possible features that might explain the observed patterns of repression (**Supplementary Data Set 2** and **Supplementary Table 4**). To select the most important feature explaining the repression data, we applied partial least-squares regression (PLSR) with stepwise feature selection and outlier detection[44] (described in Methods and **Supplementary Methods**). The analysis, after detecting and discarding 8 outlier interactions (out of 529 interactions), identified just two features that explain the 86% variation in the data after ten-fold cross-validation: the hybridization energies of the entire 37-bp interaction region and of a 5-bp seed region (**Fig. 4a**). The model suggests that the initial nucleation event at the G-C–rich 5-bp seed region and the subsequent helix progression is thermodynamically driven and determines the efficient repression of the target mRNA. These results recapitulate early studies that pointed out the importance of the 5-bp interaction region in determining the copy-number control performance of the RNA-IN/OUT system[35].

The unexplained 14% variance in the repression data may be due to other features or factors that are not included in this work, such as the *in vivo* concentrations of interacting RNAs, which influence the efficiency of antisense RNAs[41]. The eight outlier RNA interactions removed from the training model, which had high residual variance and high leverage, did not have any obvious properties that explained the observation. We speculate that the peculiarity of the structure or the *in vivo* stability of duplexes of these outlier interacting pairs may be the reason for their unpredictable performance. More detailed biochemical studies are needed to pursue these hypotheses. However, the model, trained on the remaining 521 pairs, has sufficient explanatory power to support the design of new pairs.

Validation of model predictions

To validate the predictive capability of the model and forward engineer new orthogonal mutants, we used the model to predict percentage repression for all of the 56 mutant pairs that we initially considered (described in Methods, **Fig. 2a** and **Supplementary Table 1**). This yielded a total of 3,136 percentage repression predictions, including the 529 experimentally tested pairs (**Supplementary Fig. 18** and **Supplementary Table 10**). We estimated the total possible number of mutually orthogonal pairs in the 56 RNA-IN and RNA-OUT variants from these predictions for different family sizes at different thresholds of percentage repression and cross-talk (**Fig. 4b** and **Supplementary Fig. 20**). At 80% threshold percentage repression and 10% cross-reactivity, we have more than 300 families of mutually orthogonal pairs, triplets and quadruplets; more than

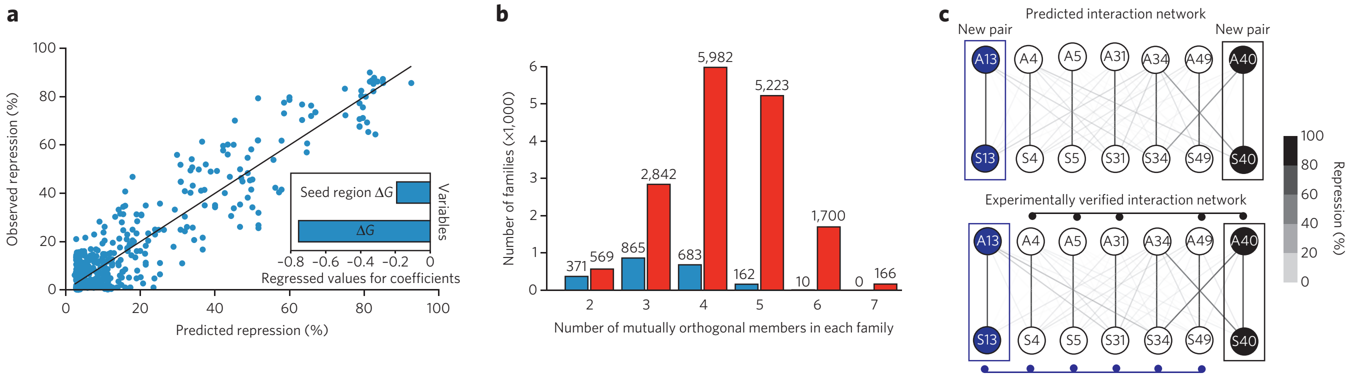

Figure 4 | The sequence-activity PLSR model and experimental validation of model predictions. (**a**) The PLSR model results are shown as a scatter plot of predicted versus observed percentage repression. The coefficient of determination (a measure of the quality of the model, $R^2 = 0.87$, $P < 0.0001$) using a training set of 521 interactions is shown along with the ten-fold cross-validated model (a measure of the predictive ability of the model, cross-validated (cv) $R^2 = 0.86$) Inset, weighted regression coefficients for the final two predictors. These include Δ*G*, hybridization energy of the sense-antisense RNA duplex; seed region Δ*G*, hybridization energy of duplex in seed region. (**b**) The estimated number of families with two to seven mutually orthogonal members in the computationally predicted repression profile of the 56 pairs. Two representative data sets are shown here for repression ≥80%: data for ≤10% cross-talk (blue bars) and ≤20% cross-talk (red bars). (**c**) The predicted (top) and experimentally verified (bottom) network of interaction between mutually orthogonal cognate and noncognate sense and antisense RNAs (mutants 13 and 40 in addition to mutants 4, 5, 31, 34 and 49). The shading of links indicates the percentage repression (black for 100% repression and light gray for almost no repression). Model-predicted and experimentally verified enlarged orthogonal families are indicated with connections between orthogonal members, and new pairs are shown in blue and black.

150 families of 5 mutants; and 10 families of 6 orthogonal mutants (**Fig. 4b**). At 20% cross-reactivity, we have more than 1,000 families made up of 3, 4, 5 and 6 mutually orthogonal mutants and more than 150 families of 7 mutants.

To experimentally validate a subset of these predictions, we forward engineered two sense and antisense RNA pairs (mutant 13 and mutant 40 shown in **Supplementary Table 1** and **Supplementary Fig. 18**) predicted to have a desired strong threshold of percentage repression and minimal cross-talk with the experimentally discovered family of five orthogonal pairs, thus expanding the size and number of orthogonal families (**Fig. 4c**). We characterized these four forward-engineered mutants in the presence of their cognate and noncognate partners, and the resulting number of interactions was more than 50 (**Supplementary Fig. 21**). As predicted, sense and antisense mutants 13 and 40 yielded an altered and expanded family made up of six and five mutually orthogonal mutants, respectively (**Fig. 4c**). The mutant 13 sense-antisense cognate pair showed specific interaction, whereas they showed <20% cross-talk with five noncognate partners. Similarly, the mutant 40 sense-antisense pair showed less than 10% cross-talk with four noncognate partners and specific interaction with each other (**Fig. 4c** and **Supplementary Fig. 21**). In addition to validating orthogonality, the experimental results clearly demonstrate the ability of the model to reliably and quantitatively predict a wide range of percentage repression shown by more than 50 interactions (**Supplementary Fig. 21**). This tool thus provides a powerful avenue to rationally design and forward engineer new orthogonal members to an existing mutually orthogonal RNA-IN/OUT pair family (**Supplementary Data Set 5**).

DISCUSSION

The translational controllers we developed here, and the model that enables the design of effective and orthogonal regulator pairs, provide an effective platform for RNA-based regulation of translation. The platform allowed the discovery of a large number of different-sized families of mutually orthogonal regulators that could, theoretically, be used in the same cell. The model enabled the engineering of new families of five and six mutually orthogonal translational regulators showing consistent and predictable performances (**Fig. 4c**). To our knowledge, these are the largest orthogonal families constructed from a single regulatory mechanism based on a predictive model (in contrast to commonly used screen-based approaches). In addition, we forecast a possibility of finding thousands of mutually orthogonal families made up of two, three, four or five regulatory members differing in specificity and cross-talk properties (**Fig. 4b**, **Supplementary Data Set 5** and **Supplementary Fig. 20**). We also demonstrate that the specificity of interaction is preserved by inserting extra nucleotides or scaling up the core interaction region, thereby opening up an avenue to allow the search of additional orthogonal regulators in the enlarged sequence space.

Analysis of the library also provides insight into the mechanisms of specificity and activity in the RNA-IN/OUT system. For example, earlier studies[35] showed that single complementary mutations at the third and fourth nucleotide at the 5′ end of RNA-IN and at the corresponding nucleotides in the loop region of RNA-OUT alter the sequence specificity of the antisense pairing reaction with their wild-type counterparts. In the present work, in addition to recapitulating these results, we observe that combinatorial complementary nucleotide swaps (1–5 nucleotides) show the mutants cleanly altered specificity of interaction with each other and with their wild-type partners, indicating the importance of the initial base-pairing region in helix progression and stable duplex formation.

Further, our results demonstrate that the ubiquitously observed YUNR motif in antisense RNA systems seems to be nonessential for retaining the specificity and efficiency of interaction in the RNA-IN/OUT system *in vivo*. In a few specific cases, antisense mutants with no YUNR motif have the same threshold of percentage repression as the wild type (for example, A6, A5, A23, A31, A52; **Supplementary Table 7**). We speculate that having a stable stem structure with a U-A–rich region around the antisense loop provides the necessary flexibility to accommodate variations in the core recognition motif (**Fig. 1b**). Having this arrangement may also allow bases following U to be directed outward and hence be available to interact with RNA-IN, thereby retaining the efficiency and specificity of interaction in the absence of the YUNR motif. Though we have not studied the performance of loop-region mutants in the context of alterations in the U-A–rich motif at the base of the RNA-OUT loop region, we speculate, on the basis of this work and earlier studies[35,38], that the U-A–rich motif determines the rapid helix progression after the initial base-pairing at the G-C–rich motif (**Fig. 1**). This speculation is in agreement with our scaled-up mutant data, which shows that insertion of two extra nucleotides between the G-C–rich 3-nt core region and A-U–rich downstream region is well tolerated. A few mutants that do not show a strong interaction (threshold of percentage repression) with their cognate partners (for example, A7, A21, A22, A30 and A43 in **Supplementary Table 7**) also do not have a YUNR motif, and further characterization is needed to understand whether concentration or structural features influence their performance. These results indicate that the YUNR motif as a principal design feature may not be universally applicable to all synthetic RNA regulators, and more elaborate studies are needed to show whether there are any structural dependencies for this motif to function as a key determinant of interaction.

Our model selection procedure suggests that the hybridization energy of the entire duplex and that of the 5-bp duplex seed region are the important features determining the sense-antisense RNA interaction specificity. The importance of the seed region suggests that the initial nucleating events and interaction strength of this region in the propagation of the base-pairing reaction are critical to specificity and repression. Earlier work suggested that only the first 3 bp of the RNA-IN/OUT duplex and their free energy of base pairing may be important for this nucleating event[41]. The seed region has also been found to be a key determinant of activity in eukaryotic miRNA studies[32,33] and recently has been recognized in bacterial small RNAs[45]. The role of the 37-bp duplex hybridization energy in explaining the majority of *in vivo* repression data suggests that initiation and propagation of the nucleating event and strand displacement of intramolecular RNA-OUT base pairs with the stable IN-OUT duplex are, however, the main determinants of repression activity. We speculate that hybridization free energy is determinative of interaction specificity only for systems in which an unstructured RNA 5′ end initiates binding within the loop of its counterpart and possibly forms a duplex by a single-step strand-exchange pathway (as exemplified in this work and in a study involving a synthetic antisense RNA system[15]). This does not seem to be the case with those systems in which sense and antisense RNAs initiate loop-loop kissing-complex formation followed by the interaction at a distal site to overcome the topological limitation of helix progression[11]. In such cases, it is the kinetics rather than the thermodynamics of the interaction that seems to be important, and the specificity seems to be the consequence of interactions between intricate three-dimensional structures of the interacting RNAs[11,29,30]. It is reassuring that the RNA-IN/OUT system falls into the simpler class, enabling design of new variants that function equivalently and orthogonally to the wild-type element. These insights can be readily used for designing structurally analogous but sequence-independent antisense RNAs and corresponding sense RNAs using the same orthogonal core seed regions found in the present work. However, the experimental data–driven modeling approach presented here can also be applied for engineering other RNA species (that rely on different modes of interaction). We speculate that to do so may require characterization of a large panel of mutants before building a predictive sequence-activity model with a large number of sequence and structural features.

By describing the thresholds of repression and cross-reactivity percentages for cognate and noncognate pairs, respectively, we can quantify mutual orthogonality and begin to define specification for parts acceptable for a particular application. Such an analysis is of immense value when these RNA regulators are integrated into synthetic genetic circuits and are required to function in a predictable way to avoid unwanted cross-talk with other parts in the circuit. Some applications (for example, a circuit regulating cell death) demand a very precise and specific regulation of part components, whereas in some applications this requirement may not be that stringent. It is likely, for example, that as the number of orthogonal pairs to be used together in a cell increases, there will be an increasingly stringent requirement for a low percentage of cross-talk as the apparent nonspecific repression of a sense target might be affected by the sum of the concentrations of all cross-talking antisense molecules. However, the combination of orthogonal and promiscuous variants (for example, a single antisense RNA repressing multiple sense targets and, alternatively, single sense targets recognized by multiple antisense RNAs; **Supplementary Fig. 11**) can be useful either in designing synthetic circuits as signal propagation and/or signal integration modules[23,46] or for building a hierarchy of interlinked functional modules to understand the network structure and function[46]. Such bottom-up engineering of gene expression circuits using well-characterized parts (with different thresholds of percentage repression and cross-talk) into interlinked networks reminiscent of natural systems can yield insights into organization principles of circuit design and, more generally, how evolution may have shaped these complex regulatory networks.

We envision that both the large families of orthogonal regulators and the approach used to discover them are important for a variety of next-generation synthetic biology applications[4]. This includes the use of orthogonal translation repressors to perform different modes of RNA computations in bacteria[23] and regulate expression of different genes in operons[19], to establish the hierarchical order of regulation[46], to modulate protein abundance and thereby impart pathway balance[1,19], to explore the network architecture and understand the function of natural small-RNA regulatory networks by rewiring or building smaller synthetic circuits and to understand the evolutionary design principles behind regulatory architectures. Even though the design specification for these various applications may not be very well defined or realized, having a compendium of well-characterized parts to meet the diverse specification needs saves a lot of resources compared to *ad hoc* approaches[47] and provides a robust platform for further technological development (for example, increasing the dynamic range of repression or engineering the RNA components presented here to be sensitive to environmental signals). The characterization and standardization of genetic parts is particularly important because the real potential of using RNA-based components in genetic circuits derives from their particular designability, homogeneity and scalability compared to protein regulators[11,22–24]. Therefore, we believe that the use of RNA components described in this work in combination with available transcriptional and post-transcriptional regulatory components and other recently developed orthogonal systems[5,6,11,14,15,22,48,49] should provide an unsurpassed platform for flexibly tailoring delays, rapid response times, controlled variability in gene expression or other interesting and useful dynamic properties[46] that are sensitive to internal or external signals.

METHODS

Strains, plasmids and growth conditions. Strains used in this study are listed in **Supplementary Data Set 1**. *E. coli* strain TOP10 (Invitrogen) was used for plasmid construction purposes and for fluorescence measurements. All strains were grown in LB (DIFCO) medium or MOPS EZ rich medium (Teknova) at 37 °C with shaking. Medium was supplemented with 100 μg ml^{-1} carbenicillin, 34 μg ml^{-1} chloramphenicol or both as required.

The construction details of the sense and antisense plasmid library are presented in the **Supplementary Methods**. The sense constructs constitutively express a variant of sense RNA-IN translationally fused to SFGFP (VKM84) or mRFP1 (VKM85) from a low-copy vector (pSC101 replication origin, chloramphenicol-resistance marker) derived from the pBbS6c vector series[50]. The antisense construct expresses a variant of antisense RNA-OUT under the control of the IPTG-inducible $P_{\mathrm{LlacO\text{-}1}}$ promoter from a high-copy vector (VKM87, ColE1 replication origin, ampicillin resistance marker, pBbE6a vector series[50]). A control vector (VKM22) was also constructed that lacks the RNA-OUT region and yields a ~50-nt nonsense transcript derived from the double terminator of the pBbE6a vector series[50]. The sense RNA-IN variant sequences (+1 to +40 from the transcription start site) and antisense RNA-OUT variant sequences (+1 to +115 from the transcription start site) are presented in **Supplementary Table 1**, and plasmids used in this study are listed in **Supplementary Table 2**.

***In vivo* assays using the plate reader and flow cytometer.** Frozen TOP10 cells carrying a RNA-IN fusion plasmid with a plasmid expressing either RNA-OUT or nonsense RNA (control vector, VKM22) were grown overnight (~16 h) in MOPS medium with appropriate antibiotics at 37 °C with shaking. The following day, overnight cultures were diluted 1:25 into a fresh medium with appropriate antibiotics and 1 mM IPTG and were grown until cultures reached logarithmic growth phase. The measurements of attenuance (*D*) at 600 nm and fluorescence (arbitrary fluorescence units; for SFGFP, excitation at 480 nm and emission at 510 nm; for mRFP1, excitation at 560 nm and emission at 610 nm) were performed in the Tecan Sapphire 2 spectrophotometer by diluting the log-phase culture (0.5–0.6 D_{600}) 1:2 in PBS buffer (pH 7.4). The percentage of repression was calculated by taking into account the background-subtracted fluorescence of RNA-IN-SFGFP or RNA-IN-mRFP1 in the presence and absence of plasmid expressing antisense RNA (**Supplementary Methods**). The measurement of fluorescence in the flow cytometer was performed by diluting the log-phase culture 1:50 with PBS buffer (pH 7.4) containing 200 μg ml^{-1} kanamycin. The plate was analyzed immediately using a Partec CyFlow Space flow cytometer with a RobbyWell 96-well-plate auto sample loader (Forward scatter (FCS), side scatter (SSC); 488 nm excitation, 520 nm band pass emission filter for SFGFP and 590 nm excitation, 610 nm band pass emission filter for mRFP1). For each strain, a minimum of 50,000 cellular counts were collected using SSC as the cell trigger. The flow cytometer data was analyzed as detailed in the **Supplementary Methods**.

Multivariate PLSR model. To compile different sequence, thermodynamic and accessibility features that explain the base-pairing specificity with respect to percentage repression, we used the predicted sense-antisense RNA duplex structures as a guiding reference (**Supplementary Data Set 2** and http://genomics.lbl.gov/supplemental/RNA-IN-OUT-Mutalik-etal-2011/). We shortlisted 31 different features belonging to single and duplex RNA sequences (**Supplementary Table 4** and **Supplementary Methods**).

To build a predictive model based on the sequence and thermodynamic features of sense, antisense and duplex RNA species and their corresponding percentage repression profiles, we used the PLSR approach[44]. PLSR is a method for relating two data matrices, response variable *Y* to multiple predictors *X*, by a linear multivariate model. PLSR finds components of *X* so that they approximate *X* and correlate with *Y*. This is the method of choice for handling multicollinearity among *X* values when two or more predictor variables are highly correlated and hence robust estimation of regression coefficients by simple multiple linear regression method is difficult. The following equation shows the linear relationship between response variable and predictors:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j3} \ldots\ldots + \beta_t x_{jt} + \varepsilon_j$$

where y_j is a column vector containing the percentage repression data for a total of 529 data points for all 23 sense and antisense pairs; x_{j1}, x_{j2}, x_{j3}… are predictors in matrix X (total of 31 predictors or features); β_1, β_2, β_3… are regression coefficients, β_0 is the regression constant, ε_j is an error term, the value of *j* ranges from 1 to 529; and the value of *t* ranges from 1 to 31. Thus, the matrix X contains computed normalized predictors and column vector Y contains the experimentally determined percentage of repression (**Supplementary Methods** and **Supplementary Data Set 3**). We used the Unscrambler X10 (CAMO software) for PLSR model (PLSR1) building, outlier detection and calculation of regression coefficients. All models were built by applying the standard data preprocessing procedures. To eliminate features with insignificant (*P* value >0.05) regression weights (those that do not contribute to the model), we used stepwise regression for iterative feature elimination procedure and downselected the two most significant (*P* value <0.05) predictors for the final model (**Supplementary Methods**). Candidates with high residual *y* variance and high leverage were deemed outliers and were removed from the final model. To test whether the model overfitted the data, ten-fold cross-validation was performed. We used the model trained on the experimental data set to predict percentage repression of all 56 pair combinations (**Supplementary Data Set 4**). The predicted percentage repression from the model for the 56-pair combination is given in **Supplementary Table 10**.

Received 12 October 2011; accepted 27 January 2012;
published online 25 March 2012

References

1. Keasling, J.D. Synthetic biology for synthetic chemistry. *ACS Chem. Biol.* **3**, 64–76 (2008).
2. Ruder, W.C., Lu, T. & Collins, J.J. Synthetic biology moving into the clinic. *Science* **333**, 1248–1252 (2011).
3. Lucks, J.B., Qi, L., Whitaker, W.R. & Arkin, A.P. Toward scalable parts families for predictable design of biological circuits. *Curr. Opin. Microbiol.* **11**, 567–573 (2008).
4. Purnick, P.E. & Weiss, R. The second wave of synthetic biology: from modules to systems. *Nat. Rev. Mol. Cell Biol.* **10**, 410–422 (2009).
5. Zhan, J. *et al.* Develop reusable and combinable designs for transcriptional logic gates. *Mol. Syst. Biol.* **6**, 388 (2010).
6. Rackham, O. & Chin, J.W. A network of orthogonal ribosome x mRNA pairs. *Nat. Chem. Biol.* **1**, 159–166 (2005).
7. Alper, H., Fischer, C., Nevoigt, E. & Stephanopoulos, G. Tuning genetic control through promoter engineering. *Proc. Natl. Acad. Sci. USA* **102**, 12678–12683 (2005).
8. Deuschle, U., Kammerer, W., Gentz, R. & Bujard, H. Promoters of *Escherichia coli*: a hierarchy of *in vivo* strength indicates alternate structures. *EMBO J.* **5**, 2987–2994 (1986).
9. 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).
10. Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C. & Wold, S. Quantitative sequence-activity models (QSAM)—tools for sequence design. *Nucleic Acids Res.* **21**, 733–739 (1993).
11. Lucks, J.B., Qi, L., Mutalik, V.K., Wang, D. & Arkin, A.P. Versatile RNA-sensing transcriptional regulators for engineering genetic networks. *Proc. Natl. Acad. Sci. USA* **108**, 8617–8622 (2011).
12. Liu, C.C., Qi, L., Yanofsky, C. & Arkin, A.P. Regulation of transcription by unnatural amino acids. *Nat. Biotechnol.* **29**, 164–168 (2011).
13. Chubiz, L.M. & Rao, C.V. Computational design of orthogonal ribosomes. *Nucleic Acids Res.* **36**, 4038–4046 (2008).
14. Dixon, N. *et al.* Reengineering orthogonally selective riboswitches. *Proc. Natl. Acad. Sci. USA* **107**, 2830–2835 (2010).
15. Isaacs, F.J. *et al.* Engineered riboregulators enable post-transcriptional control of gene expression. *Nat. Biotechnol.* **22**, 841–847 (2004).
16. Carothers, J.M., Goler, J.A., Juminaga, D. & Keasling, J.D. Model-driven engineering of RNA devices to quantitatively program gene expression. *Science* **334**, 1716–1719 (2011).
17. Win, M.N. & Smolke, C.D. Higher-order cellular information processing with synthetic RNA devices. *Science* **322**, 456–460 (2008).
18. Carrier, T.A. & Keasling, J.D. Library of synthetic 5′ secondary structures to manipulate mRNA stability in *Escherichia coli*. *Biotechnol. Prog.* **15**, 58–64 (1999).
19. Smolke, C.D. & Keasling, J.D. Effect of copy number and mRNA processing and stabilization on transcript and protein levels from an engineered dual-gene operon. *Biotechnol. Bioeng.* **78**, 412–424 (2002).
20. Babiskin, A.H. & Smolke, C.D. A synthetic library of RNA control modules for predictable tuning of gene expression in yeast. *Mol. Syst. Biol.* **7**, 471 (2011).
21. Saito, H. *et al.* Synthetic translational regulation by an L7Ae-kink-turn RNP switch. *Nat. Chem. Biol.* **6**, 71–78 (2010).
22. Culler, S.J., Hoff, K.G. & Smolke, C.D. Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. *Science* **330**, 1251–1255 (2010).
23. Benenson, Y. RNA-based computation in live cells. *Curr. Opin. Biotechnol.* **20**, 471–478 (2009).
24. 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).
25. Georg, J. & Hess, W.R. *cis*-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 286–300 (2011).
26. Engdahl, H.M., Hjalt, T.A. & Wagner, E.G. A two unit antisense RNA cassette test system for silencing of target genes. *Nucleic Acids Res.* **25**, 3218–3227 (1997).
27. Man, S. *et al.* Artificial *trans*-encoded small non-coding RNAs specifically silence the selected gene expression in bacteria. *Nucleic Acids Res.* **39**, e50 (2011).
28. Nakashima, N. & Tamura, T. Conditional gene silencing of multiple genes with antisense RNAs and generation of a mutator strain of *Escherichia coli*. *Nucleic Acids Res.* **37**, e103 (2009).
29. Thomason, M.K. & Storz, G. Bacterial antisense RNAs: how many are there, and what are they doing? *Annu. Rev. Genet.* **44**, 167–188 (2010).
30. Wagner, E.G.H., Altuvia, S. & Romby, P. Antisense RNAs in bacteria and their genetic elements. *Adv. Genet.* **46**, 361–398 (2002).
31. Sahota, G. & Stormo, G.D. Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes. *Bioinformatics* **26**, 2672–2677 (2010).
32. Rajewsky, N. microRNA target predictions in animals. *Nat. Genet.* **38** (suppl. 1) S8–S13 (2006).
33. Thomas, M., Lieberman, J. & Lal, A. Desperately seeking microRNA targets. *Nat. Struct. Mol. Biol.* **17**, 1169–1174 (2010).
34. 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).
35. Kittle, J.D., Simons, R.W., Lee, J. & Kleckner, N. Insertion sequence IS10 anti-sense pairing initiates by an interaction between the 5′ end of the target RNA and a loop in the anti-sense RNA. *J. Mol. Biol.* **210**, 561–572 (1989).
36. Ma, C. & Simons, R.W. The IS10 antisense RNA blocks ribosome binding at the transposase translation initiation site. *EMBO J.* **9**, 1267–1274 (1990).
37. Case, C.C., Simons, E.L. & Simons, R.W. The IS10 transposase mRNA is destabilized during antisense RNA control. *EMBO J.* **9**, 1259–1266 (1990).
38. Jain, C. IS10 antisense control *in vivo* is affected by mutations throughout the region of complementarity between the interacting RNAs. *J. Mol. Biol.* **246**, 585–594 (1995).
39. Franch, T., Petersen, M., Wagner, E.G., Jacobsen, J.P. & Gerdes, K. Antisense RNA regulation in prokaryotes: rapid RNA/RNA interaction facilitated by a general U-turn loop structure. *J. Mol. Biol.* **294**, 1115–1125 (1999).
40. Pédelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C. & Waldo, G.S. Engineering and characterization of a superfolder green fluorescent protein. *Nat. Biotechnol.* **24**, 79–88 (2006).
41. Jain, C. Models for pairing of IS10 encoded antisense RNAs *in vivo*. *J. Theor. Biol.* **186**, 431–439 (1997).
42. Markham, N.R. & Zuker, M. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.* **453**, 3–31 (2008).
43. Doench, J.G. & Sharp, P.A. Specificity of microRNA target selection in translational repression. *Genes Dev.* **18**, 504–511 (2004).
44. Wold, S., Sjostrom, M. & Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemometr. Intell. Lab.* **58**, 109–130 (2001).
45. Papenfort, K., Bouvier, M., Mika, F., Sharma, C.M. & Vogel, J. Evidence for an autonomous 5′ target recognition domain in an Hfq-associated small RNA. *Proc. Natl. Acad. Sci. USA* **107**, 20435–20440 (2010).
46. Beisel, C.L. & Storz, G. Base pairing small RNAs and their roles in global regulatory networks. *FEMS Microbiol. Rev.* **34**, 866–882 (2010).
47. Endy, D. Foundations for engineering biology. *Nature* **438**, 449–453 (2005).
48. Tabor, J.J., Levskaya, A. & Voigt, C.A. Multichromatic control of gene expression in *Escherichia coli*. *J. Mol. Biol.* **405**, 315–324 (2011).
49. An, W. & Chin, J.W. Synthesis of orthogonal transcription-translation networks. *Proc. Natl. Acad. Sci. USA* **106**, 8477–8482 (2009).
50. Lee, T.S. *et al.* BglBrick vectors and datasheets: A synthetic biology platform for gene expression. *J. Biol. Eng.* **5**, 12 (2011).

Acknowledgments

The authors would like to thank G. Cambray, E. Gogol, C. Liu and J. Skerker for comments on the manuscript. V.K.M. was supported by British Petroleum under contract number LB08004883 at the Joint BioEnergy Institute. A.P.A. and L.Q. acknowledge support from the Synthetic Biology Engineering Research Center under US National Science Foundation grant number 04-570/0540879. J.C.G. acknowledges financial support by the Portuguese Fundação para a Ciência e a Tecnologia (SFRH/BD/47819/2008). J.B.L. acknowledges the financial support of the Miller Institute for Basic Scientific Research. This work conducted by the Joint BioEnergy Institute was supported by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy under Contract No. DE-AC02-05CH11231.

Author contributions

V.K.M. conceived of the study, designed and performed experiments, built the computational model. L.Q. designed and performed experiments. J.C.G. designed and built the computational model. J.B.L. provided reagents and key insights. A.P.A advised at all levels of the project. V.K.M. and A.P.A. wrote the manuscript. V.K.M., L.Q., J.C.G., J.B.L. and A.P.A. interpreted results, discussed and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available online at http://www.nature.com/naturechemicalbiology/. Reprints and permissions information is available online at http://www.nature.com/reprints/index.html. Correspondence and requests for materials should be addressed to A.P.A.