Computational Design of Tunable Kinetic RNA Biosensors

David Sparkman-Yager

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington 2021

Reading Committee:

James M. Carothers, Chair

Jesse Zalatan

Christopher Thachuk

Program Authorized to Offer Degree:

Molecular Engineering

University of Washington

Abstract

Computational Design of Tunable Kinetic RNA Biosensors

David Sparkman-Yager

Chair of the Supervisory Committee:

James M. Carothers

Department of Chemical Engineering

Bacterial metabolism is made up of complex gene networks that can be engineered to produce valuable chemicals from renewable resources such as sugar. However, these complex networks require sophisticated genetic engineering to optimize the production of these high-value compounds. For most biosynthetic pathways, it remains impossible to predict what combinations of gene expression levels will give rise to the most productive bacteria. Therefore, to identify the most productive pathway variants it is necessary to develop biosensors that can give a visible readout of how much of the target chemical each cell is producing. In nature, bacteria's solution to measuring the concentrations of important chemicals is a class of RNA based biosensors called "riboswitches". Riboswitches bind their target chemical and modulate gene expression levels in response. However, despite numerous design and selection methods published to date, there exists no common methodology to design high-performance, tunable, chemical-responsive RNA biosensors from first principles. By combining design lessons learned from nature with novel kinetic RNA folding predictions we present a new class of engineered RNA biosensors with increased activation ratios and tunable ligand sensitivities, both of which are essential for applying biosensors to real-world problems. We first demonstrated that our novel molecular architecture can be

applied to the regulation of self-cleaving ribozymes *in vitro*, resulting in activation ratios exceeding 200-fold. By rationally modifying a single domain within the biosensors we were able to tune the sensitivity to its target molecule by more than 100-fold. We then demonstrated that the same architecture can be applied to regulate protein expression levels in *E. coli* with unprecedented sensitivity. Finally, we use our kinetic design rules to engineer a CRISPR-based transcriptional activation system, resulting in a completely novel biosensor, for an industrially-relevant chemical, able to sense the production of the target within an engineered bacterium.

In loving memory of Teresa Sparkman and Sofie Sparkman-Yager

I wouldn't be here without you

I can't believe I'm here without you

Table of Contents

Introduction

Industrial Biotechnology

Bacterial metabolism is made up of complex gene networks that can be engineered to produce valuable chemicals from renewable resources such as sugar, or cellulosic feedstocks. However, these complex networks require sophisticated genetic engineering to optimize the production of these high-value compounds. For most biosynthetic pathways, it remains impossible to predict what combinations of gene expression levels will give rise to the most productive bacteria. Therefore, to identify the most productive pathway variants it is necessary to develop biosensors that can give a visible readout of how much of the target chemical each cell is producing[1]. In nature, bacteria's solution to measuring the concentrations of important chemicals is a class of RNA based biosensors called "riboswitches"[2–4].

Inspired by nature, ligand-responsive RNA switches have emerged as exciting tools with applications as biosensors and genetic controllers in diverse fields such as biomedicine, metabolic engineering, and synthetic biology[5–7]. However, despite numerous design and selection methods published to date, there exists no methodology to design high-performance, tunable, ligand-responsive RNA switches from first principles[8,9]. As computational power rapidly becomes cheaper and more available, computational switch design becomes increasingly attractive compared to currently available time-, money- and skill-intensive laboratory screening methods. By combining the RNA design lessons learned from nature with novel RNA folding predictions it should become possible to identify RNA switches with increased dynamic ranges (ratio of ligand-induced signal to background signal) and tunable ligand sensitivities, both of which are essential for applying RNA switches to real-world problems.

Bacterial riboswitches

It is well known that chemical reactions can proceed via thermodynamic or kinetic reaction control, wherein the reaction time-scale dictates whether the more stable product, or the product more rapidly formed, dominates. The role of kinetic control mechanisms in tuning the sensitivities and selectivities of biological information processing functions has long been recognized in natural systems and, more recently, kinetic mechanisms providing non-equilibrium binding responses have been observed in naturally-occurring RNA genetic regulators[10–14]. In many of these systems kinetic mechanisms

coordinate co-transcriptional RNA folding with metabolite binding and can enable outputs that are highly-sensitive to targeted ligand concentrations.

Kinetic mechanisms providing non-equilibrium responses have been observed in many naturally-occurring riboswitches[14–17]. Riboswitches are a class of cis-regulatory ligand-responsive RNA switch that are ubiquitous in bacteria and control gene expression through diverse biochemical mechanisms such as transcription termination and translation initiation. Those riboswitches that display kinetic control often possess a common molecular architecture in which the ligand-binding aptamer domain is upstream of the expression platform that it regulates. Combined with the directionality of transcription, this defines a transcriptional 'binding window' for ligand association that opens with the folding of the aptamer, and closes with the transition to an inactive conformation. Thus, changing the duration of the binding window provides a route for tuning the dynamics of the response. In general, lengthening the binding window promotes aptamer-ligand association, thereby increasing the sensitivity of the riboswitch to lower metabolite concentrations. For example, the *E. coli* thiamine pyrophosphate (TPP) riboswitches exhibit co-transcriptional time delays as large as 40 seconds, conferring concentration response set-points that can differ by more than an order of magnitude[13].

While natural riboswitches display kinetic co-transcriptional ligand-binding behavior, engineering such behavior in synthetic RNA switches has proved challenging. Although thermodynamic RNA secondary-structure design, based on minimum free energy (MFE) structural predictions, has been utilized to identify sequences capable of folding into multiple structure-states, synthetic riboswitches exhibiting kinetic control have only been identified by chance[15–18]. To our knowledge, it has not yet been demonstrated that kinetically-controlled RNA devices can be rationally designed.

Organization of Thesis

This thesis is organized into 5 chapters, each covering a particular focus of my research while in the Carothers lab as a graduate student, between Fall 2014 and Spring 2021. Taken together this thesis demonstrates a novel method, inspired by nature, to generate genetically-encoded biosensors able to measure the intracellular concentration of target chemicals, and presents a new paradigm for the

computational generation of tunable kinetic RNA biosensors.

Chapter 1 describes the development of a molecular architecture and computational workflow for the *in silico* engineering of Kinetic-Aptazymes. Through *in vitro* co-transcriptional cleavage assays, a set of computational design rules are identified that enable the robust *in silico* identification of functional RNA switches. High performance switches and tunable ligand sensitives are demonstrated.

Chapter 2 describes the engineering of antisense-ribosome binding site (AS-RBS) riboswitches, wherein the presence of the small molecule theophylline controls the expression of a fluorescent protein in *E. coli*. Borrowing from nature, transcriptional pause sites are incorporated into the switches to achieve unprecedented levels of sensitivity for their target molecule.

Chapter 3 is a manuscript on which I was a co-first author, published in *Current Opinion in Biotechnology*. It provides an introduction to bacterial CRISPR activation (CRISPRa) system that is the focus of the final 2 chapters of my thesis. This chapter describes the potential for CRISPRa as a tool for bacterial metabolic engineering. It also explores the potential for the dynamic regulation of CRISPRa activity through various mechanisms.

Chapter 4 describes our work to predict the activity of scaffold RNAs (scRNAs) from their sequence alone. Utilizing our computational tools, we develop the Wayfinder algorithm to predict the activity of full-length and truncated scRNAs. Subsequently, we compare the Wayfinder algorithm to the state of the art for guide activity prediction tools. Finally, we determine the sequence and structure conservation of the Cas-9 binding handle, opening the door for the engineering of ligand-responsive scRNAs.

Chapter 5 describes our work to engineer ligand-responsive scRNAs. First, we engineer a scRNA able to respond to the small molecule theophylline. Next, we perform *in vitro* selection to identify a novel RNA aptamer to a human milk oligosaccharide (HMO), a critical component of human breast milk. Then we engineer a CRISPRa-regulated metabolic pathway for the biosynthesis of our HMO product (HMO-p) in *E. coli*. Finally, using our novel aptamer, we engineer HMO-p-responsive scRNAs able to sense the production of HMO-p from our engineered metabolic pathway.

Note: for contractual obligations to confidentiality, the particular HMO target, as well as the enzymes that

are necessary to produce it, have been anonymized.

Chapter 1: Kinetic-Aptazymes (K-As)

Introduction

Aptazymes are a class of synthetic RNA switch combining a ligand-binding aptamer domain with an autocatalytic self-cleaving ribozyme domain. As they do not rely on other biomolecules to function, they are attractive for their utility as multi-host genetic controllers, where RNA backbone cleavage can be utilized by the host biochemistry in a variety of ways. Aptazymes have been implemented to dynamically regulate gene expression levels in bacteria, yeast, mammalian cells, and viruses[6,19–21]. Another beneficial result of aptazymes' independence from other biomolecules is that their kinetic properties can be readily assayed *in vitro* through changes in the length of the RNA molecules, as characterized by denaturing Urea-PAGE[22].

Based on ribozyme cleavage rates, and the background hydrolysis rate of RNA, aptazymes can theoretically possess dynamic ranges of greater than 10^7-fold[23,24]. However, despite numerous design and selection methods, the best dynamic ranges identified to date within physiological buffer conditions are < 10^2-fold, due to high background cleavage, low induced cleavage, or both[8]. One potential reason for this shortcoming of conventionally-designed thermodynamic aptazymes is their fundamental competition between aptazyme dynamic range and their concentration sensitivity. This means it is only possible to generate thermodynamic aptazymes that are sensitive to their target, or possess large dynamic ranges, but not both[25,26]. This becomes especially problematic when trying to design switches responsive to ligands with low affinities for their cognate aptamer, where the concentration necessary for significant actuation lies above the limit of solubility (or toxicity for a host organism). Another potential reason for the shortcoming of conventional aptazymes is the modification of the component parts. In order to couple the functions of binding and cleavage, the aptamer domain is inserted in place of one of the ribozyme's interaction loops in the hopes of causing the ribozyme domain to reversibly misfold. As these loops are critical for forming tertiary contacts known to dramatically speed ribozyme cleavage, the maximum cleavage rate, and therefore aptazyme performance, is degraded even when the target ligand is bound[27,28]. Additionally, as the sequences are overlapping, there is no guarantee that the 3D structures of the two domains will be compatible with each other, adding additional unknown variables in the design process.

Numerous strategies have been employed to design aptazymes to date, including semi-rational design, thermodynamic secondary-structure design, *in vitro* selection, and *in vivo* screening[7,9,19,29]. The most successful strategy to date has been *in vivo* selection, wherein libraries containing thousands of variants are screened for their ability to provide ligand-responsive genetic output (usually fluorescence) within a desired genetic context[8]. While computational design holds perhaps the greatest promise long-term, computationally designed aptazymes suffer from poor predictability, with only a small fraction of designed devices proving functional, and possessing small dynamic ranges when they do[30]. When computational design strategies are employed, they are rarely generalizable or systematic, and usually involve varying the sequence composition of a single stem[31]. An additional problem with the rational/computational design of aptazymes for genetic control is their integration into a novel context. Even aptazymes possessing excellent *in vitro* performance are often rendered non-functional when placed into a genetic context, as the surrounding RNA sequence can interact with the aptazyme and hinder its function[32]. Identifying molecular design rules that apply just as well to an aptazyme within a larger RNA molecule, as they do to an isolated aptazyme, would represent a tremendous step forward for RNA design as a whole. A final, significant limitation of conventional aptazymes is the lack of control over the ligand concentration to which the aptazyme responds. For nearly any application, if the aptazyme sensitivity and desired ligand concentration are mismatched, the aptazyme becomes useless. Thus, to fully realize the utility of ligand-responsive switches to the extent that nature has, it will be necessary discover a route to rationally tune the ligand sensitivity of an aptazyme without breaking the switching capability.

One critical aspect of the cellular production of RNA molecules is that of co-transcriptional folding[33]. As the RNA polymerase produces an elongating transcript, that transcript can begin folding before the entire RNA molecule has been produced. Despite being an integral feature of naturally-occurring RNA switches, co-transcriptional folding has long been viewed as an impediment to synthetic RNA design, as the resulting kinetic traps often invalidate thermodynamic structural predictions on relevant biological timescales[32]. A kinetic trap occurs when an RNA molecule becomes stuck in an energetically-suboptimal conformation, due to a slow transition rate to the minimum free-energy (MFE) structure. However, despite this perceived impediment to RNA design, co-transcriptional folding allows

the free-energy landscape to evolve, and be programmed, as a function of time and sequence length, resulting in tunable kinetic responses in addition to those dictated solely by thermodynamic ensemble behaviors[34].

As a general principle, increasing the length of an RNA molecule dictates an increased stability of its minimum free energy (MFE) structure, as well as decreased rates of interconversion between its various global folds. Thus, an elongating transcript provides a unique opportunity to utilize small temporally-resolved inputs early in transcription to dictate large changes in the global RNA structures present on biologically-relevant timescales. While thermodynamic RNA structure predictions have been successfully used to design numerous types of short functional RNAs, longer RNA sequences such as mRNA transcripts are known to possess kinetic traps that prevent those sequences from reaching equilibrium on relevant timescales[35,36]. Thus, to predict the function of genetically-encoded RNA switches it is necessary to predict what structures those molecules adopt on the short and intermediate timescales dictated by co-transcriptional folding.

To date, there have been numerous efforts to utilize computational algorithms to predict the secondary-structures that an RNA molecule will adopt along its co-transcriptional folding trajectory[37–39]. Each has its own strengths and weaknesses, which have limited their application to RNA switch design. Problematically, many published algorithms are trained on a data set in order to get good predictions, limiting the kinetic information that can be extracted from individual transcription and refolding steps. Many are difficult to implement, and even more difficult to modify to suit one's own purposes. To this end, we developed the MFEpath algorithm for the course-grained screening for sequences able to rapidly fold during transcription into their MFE structure.

Methods

Transcription template assembly

Preparation of templates for *in vitro* transcription experiments is an important aspect of the K-A design-build-test-learn workflow. One particular difficulty faced in the synthesis of such templates is their length. Templates for K-A devices range from 150-300 nucleotides long. Oligonucleotides of up to ~150 bases can be commercially synthesized using solid state phosphoramidite chemistry, however, small

inefficiencies in the base coupling reactions result in very small yields as the lengths increase. The solution to this length limitation is to enzymatically assemble larger sequences from a collection of short fragments. Integrated DNA Technologies (IDT) uses such assembly to produce its 200-500 base long 'gBlock' fragments for $79. Unfortunately, they are unable to synthesize sequences possessing large regions of complementary or repeated sequence, which K-As sometimes possess. Additionally, the turnaround time is ~1 week.

To enable the rapid, cost-effective assembly of our desired K-A templates we developed a recursive computational algorithm to automate the generation of fragments that can be subsequently assembled through PCR (Figure S3). Primers are designed from the outside in (the reverse order of how they will be assembled), while the initial template is produced through primer extension. Each primer set is designed to have the same annealing temperature to aid in the parallel assembly of multiple devices. Primers are automatically screened to avoid stable hairpins, stable homo-dimer formation, and stable hetero-dimer formation. Once the primers are designed, they can be ordered (delivered next-day), and can then be assembled in a series of short overlap extension PCR reactions, taking ~10 minutes each. Thus, for a standard K-A, devices can be ordered one day, and then assembled and characterized the next.

MFEpath algorithm

Algorithmically, MFEpath breaks down coarse grained co-transcriptional RNA folding into a series of binary operations. For each 5' subsequence of the RNA molecule, the previous structure can either transition to the MFE substructure, or the next base is added and the structure remains unchanged (except for rapid local structural rearrangements). To determine which of these two events happens faster, their rates (approximated using structural rearrangement barriers as Arrhenius-like activation energies) are calculated using the ViennaRNA algorithm Findpath and a previously described relationship between these barriers and the interconversion rate (Equation 3). Upon addition of the final base, the height of the rearrangement barrier between the penultimate structure and the MFE structure of the full sequence is used to determine the rate of folding into the desired structure. This rate can then be compared to the rate of actuation to determine whether or not the device is expected to be functional on a

relevant timescale. As MFEpath does not depend on any specific algorithm it should be extensible as new and more efficient RNA structure/barrier prediction algorithms become available, allowing MFEpath to remain relevant into the foreseeable future.

Preliminary implementations of the MFEpath algorithm possessed a significant limitation with respect to the total evaluation time for long, kinetically-trapped, sequences. The barrier-height analysis algorithm currently employed (Findpath) increases the execution time dramatically as sequence length increases, and as the barriers between the evaluated states increase (a hallmark of sequences that don't follow the MFE structures). As co-transcriptional RNA folding is a pathway-dependent process, a dead end elimination algorithm, that terminates as soon as the pathway deviates from the desired one, is especially well suited for this application, as it saves significant computational time. Therefore, we have implemented MFEpath with a number of checkpoints (some for general properties, and some specific to the correct folding of the K-A structure states) that will terminate the simulation if failed.

Computational K-A design

In order to identify sequences capable of adopting the desired co-transcriptional structures we initially screen the population to ensure that the desired structure states exist and possess the appropriate relative free energies (Figure S6). First, everything 5' through the aptamer (s2) is screened for the ability to form the proper aptamer fold necessary for binding the ligand in the MFE structure. Next, each nucleotide of the Timer is added (s3) and screened for the presence of the aptamer structure in the MFE to ensure that once the aptamer domain folds, that it remains folded. Next, the sequence through the toehold target (s4) is considered, and screened for the ability to form a stable toehold-target duplex. Next, the entire sequence (s5) is considered and screened such that when the aptamer is present, the ribozyme is too (and vice versa). Then the entire sequence is again considered (s6) and folded to make sure that neither the aptamer nor ribozyme domain active structures appears in the MFE. This is important, as the presence of the ribozyme domain in the MFE would indicate that the K-A is active regardless of aptamer binding state, and because the presence of the aptamer in the MFE would indicate that binding the target ligand will not selectively stabilize the ribozyme. Finally, the ligand-binding

stabilization energy, as calculated from the aptamer dissociation constant, is compared to the difference in free energies of the s5 and s6 states to ensure that if the ligand binds, that the ligand stabilization is sufficient to make s5 the most stable state, instead of s6. Co-transcriptional folding analysis is then performed using MFEpath to ensure that the cutoffs described above are passed.

In vitro co-transcriptional cleavage analysis

In order to characterize the relevant kinetics of designed K-As it was necessary to develop an assay that mimicked cellular RNA production. The two critical components of cellular production are co-transcriptional folding and low free magnesium concentration. Details of the assay can be seen in previous publication[22]. In brief, DNA templates containing the K-A downstream of a bacteriophage T7 RNA polymerase were incubated with T7 RNA polymerase for up to 45 minutes. By matching the concentration of $MgCl_2$ to the concentration of NTPs, the free concentration of Mg^{2+} is expected to be between 0.5 and 1 mM, comparable to what is observed in nature. Reaction aliquots were quenched at four time points by mixing the reaction with a formamide-EDTA solution. Reaction time points were analyzed on an 8% denaturing 7.5 M Urea-PAGE gel. Gel bands were quantified using ImageJ with rolling ball background subtraction of radius 50[40].

Biphasic cleavage fitting

To capture the contributions of the two putative reaction pathways, co-transcriptional cleavage data was fit using a biphasic cleavage function (Equation 1). This 3-parameter function assumes that the rapidly-cleaving burst fraction (f_{burst}) cleaves at rate $k_{burst,}$ and the slow fraction (f_{slow}=1-f_{burst}) cleaves with rate k_{slow}. All fitting was performed using weighted least-squared regression using the Scipy package in Python. As assay signal (and therefore certainty in f_{clv}) increases as a function of time, time point duration was used for weighting of residuals. To determine the appropriate threshold for distinguishing rapid and slow cleavage, we performed an F-test for selecting either a monophasic or biphasic model for each (alpha= 0.05) of the no-, and max-ligand assays for all 50 K-A devices. For each of the 38 conditions (out of 100) for which the biphasic model was statistically superior, the k_{burst} rate was > 0.2 min^{-1}, and the

kslow rate was < 0.2 min^{-1}. For this reason, the cleavage rate 0.2 min^{-1} was used as a lower cutoff for k_{burst}, and an upper cutoff for k_{slow}, for all subsequent fitting and analysis for the entire set of 100.

In order to prevent fitting/experimental error to return non-physiological rate constants, all fitting was performed with a lower bound on cleavage rate of 1e-7 min^{-1} (rate of spontaneous RNA cleavage in buffer), and an upper bound of 5 min^{-1} (maximal cleavage rate of hammerhead ribozymes). To estimate confidence intervals for the calculated variables, percentile bootstrapping was used to yield 95% confidence intervals from 1000 re-sampled data points. Estimation of experimental error yielded a value of 1.15 % f_{clv}, by comparing the four timepoints of the 0 mM pAF condition of the pAF10-100nt device assayed on separate days.

Equation 1. $$f_{unclv.} = f_{burst} * \left(\frac{1-e^{-k_{burst}*t}}{k_{burst}*t}\right) + (1 - f_{burst}) * \left(\frac{1-e^{-k_{slow}*t}}{k_{slow}*t}\right)$$

Results and Discussion

In order to attain the tunable ligand sensitivity and large dynamic ranges (DRs) necessary for future applications, we sought to implement the kinetic control observed in natural riboswitches for the computational design of aptazymes. To do so we first had to identify a molecular architecture able to provide robust access to a co-transcriptional ligand-binding window. Inspired by natural riboswitches, as well as recent aptazyme literature, we placed the aptamer domain upstream of the ribozyme domain. In doing so it becomes possible to bind the ligand before the ribozyme domain is transcribed, providing a temporal window in which ligand-binding and cleavage are not competing. Unlike conventional aptazymes, which combine the aptamer and ribozyme domains with a randomized communication module, K-As utilize variable sequence upstream of the aptamer, and between the aptamer and ribozyme domains, to encode the desired structural transitions (Figure 1). This has a number of potential advantages: The first is the preservation of the tertiary structure of the component parts. Because none of the internal positions of either aptamer or ribozyme are mutated in order to encode structural transitions, the parts should maintain their optimal parental characteristics such as aptamer K_d and ribozyme cleavage rate. This should increase the likelihood of identifying solutions to the aptazyme design problem, by eliminating cases where the 3D structures of the two domains are incompatible.

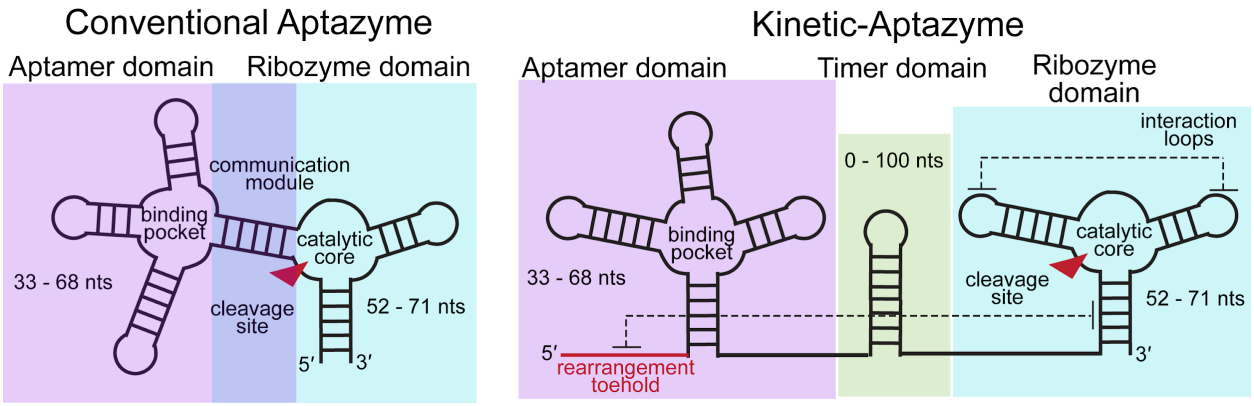


Figure 1. Comparison of the conventional strategy for assembling aptazymes with the novel Kinetic-Aptazyme (K-A) molecular architecture. Instead of encoding structural transitions in a hybrid stem (communication module), the K-A architecture allows transitions to be encoded in the intervening sequence, preserving parental part performance, and enabling the possibility of co-transcriptionally decoupling the ligand-binding and cleavage activities.

If implemented correctly, the K-A molecular architecture should allow the desired structure states to be accessed co-transcriptionally, and the relative population of the active and inactive pathways to be determined by the concentration of ligand present during the co-transcriptional ligand binding window. In the absence of ligand, the elongating K-A undergoes rapid structural rearrangement into an inactive state lacking correctly-folded aptamer and ribozyme domains (Figure 2A). In the presence of ligand, however, the aptamer domain is thermodynamically stabilized and kinetically trapped, allowing the rapid folding, and cleavage, of the ribozyme domain. Unlike conventional aptazymes, in which the background cleavage rate is dictated by the relative stability of rapidly equilibrating ON and OFF states, the background rate of K-A cleavage should be dictated by the slow kinetics of conversion from the S6 to S5 states post-transcriptionally. Thus, if the height of the B3 barrier is large, and the rate of no-ligand structural rearrangement (B2) is fast relative to transcription, K-As should be able to attain extremely low background cleavage. Combined with a small B1 barrier, which should allow the rapid formation of the catalytic ribozyme, large DRs should be within reach.

To create the diversity necessary for screening, the K-A architecture divides the mutable nucleotide positions into three regions of variable length, containing complementary sequence to the functional domains (Figure S4). This complementary sequence provides thermodynamic incentive for the K-A to fold into an alternative more stable structure in which both the aptamer and ribozyme domains are

misfolded. Each of the three domains targets a different region, and tuning their respective length modulates the thermodynamic incentive to complex with their target. By combinatorially screening all possible lengths of the three regions it should be possible to generate a *in silico* pool containing such solutions for any aptamer or ribozyme domain (Figure 2B).

In order to ensure that our designed K-A candidates fold along the desired trajectories we break the screening into two stages (described in greater detail in the methods below): multi-state thermodynamic screening, and kinetic co-transcriptional screening using our novel MFEpath algorithm. By screening for the presence, and relative stability, of multiple target states within the thermodynamic ensemble we are able to ensure that the device predominantly remains inactive at equilibrium, and that the thermodynamic stability provided by ligand-binding will bias the ensemble towards a catalytically-active fold. By subsequently screening the devices for their ability to rapidly reach the target structure states during transcription, we ensure that the general thermodynamic switch properties screened previously are accessible within an elongating transcript.

Containing up to 54 variable nucleotide positions, the K-A architecture contains $>10^{32}$ sequences for each aptamer/ribozyme pair, dramatically higher than RNA pools that can be commercially synthesized ($\sim10^{17}$). Based on the successful identification of aptazymes from similarly designed pools containing 10^6 sequences, the K-A molecular architecture likely contains many sequences that would perform as desired[19]. However, when the K-A candidate pool is generated with these semi-rational complementary sequences, as opposed to completely random sequences, the search space is reduced to a more manageable size ($\sim10^6$), while dramatically increasing the odds of finding a solution. For a single combination of aptamer and ribozyme domain we determined the odds of identifying a solution to our thermodynamic objective functions is ~1 in 57,000 within the random pool, as opposed to ~1 in 6 for the complementary pool. This ~10,000-fold increase in search efficiency allows a computational pool of 10^6 sequences to potentially contain as many solutions as a randomly-generated pool of 10^{10} sequences, which is significantly greater than what is screenable with current *in vivo* methods.

In order to test the viability of the molecular architecture for designing diverse aptazymes, we utilized two different aptamer domains: the well studied theophylline aptamer, which binds the methylated xanthine derivative theophylline, and the pAF4z1d3 aptamer, which binds the functionalized amino acid *p*-

aminophenylalanine (pAF)[41,42]. We also used three different ribozymes: The S. mansoni, sTRSV1, and PLMVd hammerhead ribozymes[6]. We analyzed the devices utilizing an *in vitro* co-transcriptional cleavage assay designed to mimic cellular production, characterizing the cleavage kinetics at various concentrations of ligand[22]. We designed, built, and characterized 50 different K-As containing combinations of the various aptamer and ribozyme domains. In doing so we were able to identify devices, incorporating each of the 5 domains, with dynamic ranges greater than 29, and as high as 240. This demonstrates that the K-A molecular architecture (and our automated computational design algorithms) can utilize diverse input components to robustly identify functional aptazymes.

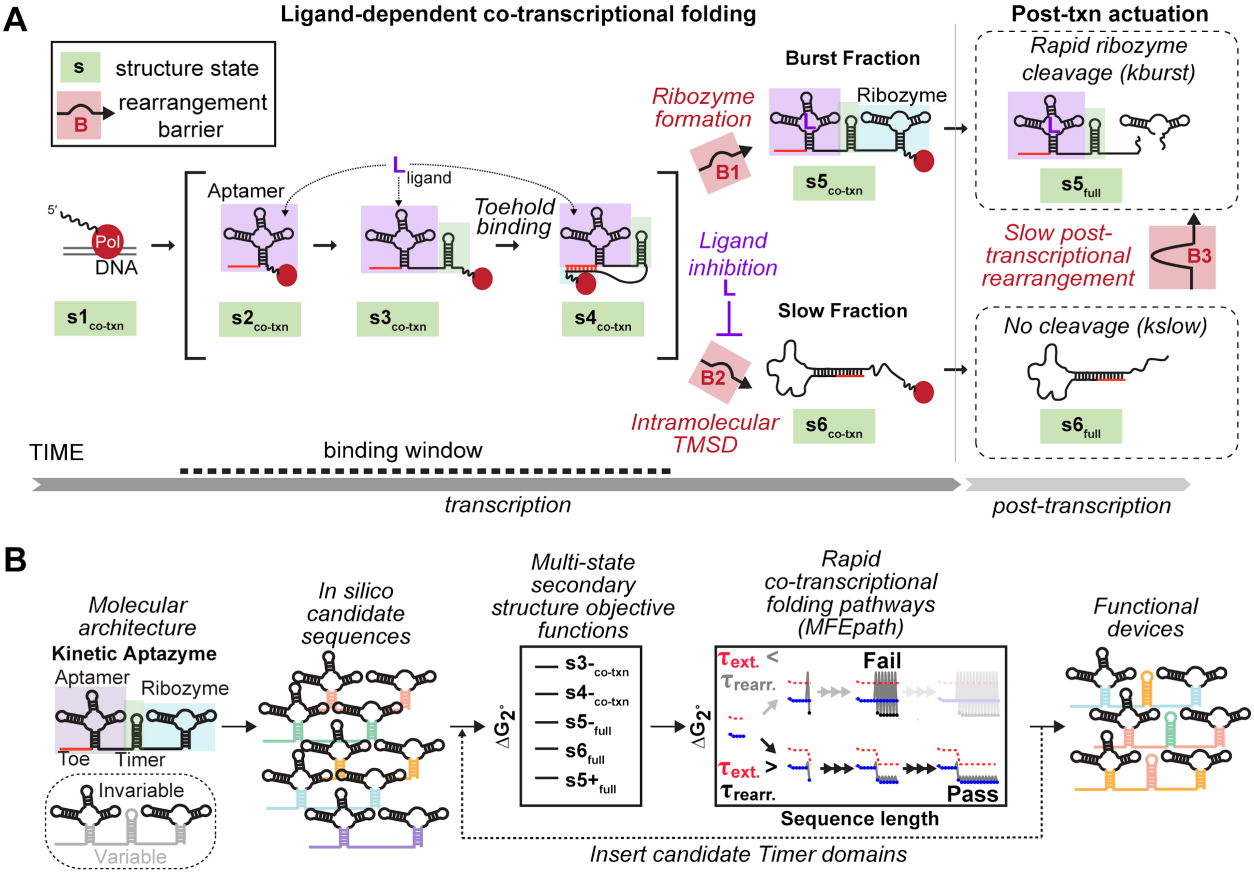


Figure 2. Kinetic-Aptazymes for engineering kinetically-controlled binding and actuation. A) Kinetic-aptazyme (K-A) switch fate is decided co-transcriptionally. After the aptamer domain is transcribed, if no ligand is present, the toehold binds its target at the 5′ end of the ribozyme domain, and the K-A undergoes rapid, intramolecular toehold-mediated strand displacement into an inactive conformation (B2). In the presence of ligand, the aptamer is stabilized, allowing the K-A to fold rapidly into the catalytically-active structure ensemble (B1). Large predicted B3 barriers should result in slow structural interconversion from the S6 to the S5 state post-transcriptionally therefore very low background

cleavage. B) In order to identify co-transcriptionally-functioning switches a molecular architecture must first be chosen that defines the variable and constant regions. Next, candidate sequences are evaluated using thermodynamic objective functions. Then, surviving sequences are analyzed for rapid co-transcriptional folding, using a novel dead end elimination algorithm that efficiently ignores sequences that fall into significant kinetic traps (MFEpath). Optionally, candidate Timer domains are introduced into functional devices, and reevaluated using the thermodynamic and co-transcriptional objective functions.

Timer domain creates 'binding window'

One essential feature of the K-A molecular architecture is the co-transcriptional ligand binding window. The binding window is the period of time after the transcription and folding of the aptamer domain, but before the K-A has made a fate decision. This window closes when the K-A, if unbound by its target ligand, structurally rearranges into a state that is neither able to bind its target nor able to cleave (S6). To ensure that the designed K-As have time for the aptamer domain to properly fold, and the target ligand to associate, we incorporated additional sequence referred to as the 'Timer' domain. The Timer domain, placed between the aptamer and ribozyme, is designed to be an orthogonal sequence element that does not contribute to the relative energetics of the designed states. However, by providing additional sequence between the two other domains, it should extend the binding window by the length of time it takes the Timer to be transcribed.

So long as the ribozyme domain is able to rapidly adopt the active conformation during transcription, on a faster timescale than that of the intrinsic ribozyme cleavage rate, we would expect K-As that bind their target co-transcriptionally would display rapid cleavage at the rate of their parental ribozyme. Thus, we expect that K-A molecules that bind their target ligand will be able to cleave at a rapid 'burst' rate, while those that do not will undergo fast structural rearrangement to the inactive S6 state, and only cleave at a 'slow' rate limited by large-scale structural rearrangement to S5 post-transcriptionally. In order to capture the contributions of the two expected reaction pathways, *in vitro* co-transcriptional cleavage data was fit using a biphasic cleavage function (see below for more details), wherein the burst fraction describes the relative abundance of the population of RNA molecules cleaving rapidly.

We hypothesized that increasing the interdomain separation between the aptamer and ribozyme with a Timer domain would aid in co-transcriptional ligand binding, and observed this to be the case. In fact, Timer domains appear to be essential for achieving rapid co-transcriptional actuation, as only K-As

containing Timer domains demonstrated ligand-inducible burst phase cleavage. For example, the Theo1-0nt K-A, which has no Timer domain, displays low background cleavage rate, no burst phase kinetics, and a moderate DR (Figure 3A-C). However, the Theo1-15nt K-A, which is an identical sequence with only a 15 nt Timer domain added, displays burst phase kinetics while maintaining the low background cleavage, resulting in a significantly increased DR of 237. We calculate DR as the ratio of k_{avg+} over k_{avg-}, which is calculated as in Equation 2. This holds true broadly, as for the twenty K-As with low background cleavage, only those K-As containing Timer domains had significant burst fractions in the presence of the ligand (Figure 3E). It is unclear why the differences are so stark, though it is possible that the additional time is necessary for the aptamer to fold into the ligand-competent state, or that there are specific geometric constraints on an elongating RNA molecule that require a minimal interdomain separation to allow the aptamer domain to be solvent accessible prior to structural rearrangement.

Equation 2. [43] $$k_{avg.} = (f_{burst} * k_{burst}) + ((1 - f_{burst}) * k_{slow})$$

It is important to note that the presence of the Timer domain alone does not appear sufficient to gain access to co-transcriptional ligand binding and burst phase kinetics. Three of the devices containing Timers displayed no significant induced burst, suggesting that other factors, such as the rate of folding of the ribozyme domain likely play a role in K-A function. For example, a long-lived folding intermediate between s4 and s5 could slow the effective cleavage rate of the K-A below our threshold for burst cleavage, despite binding the target ligand co-transcriptionally.

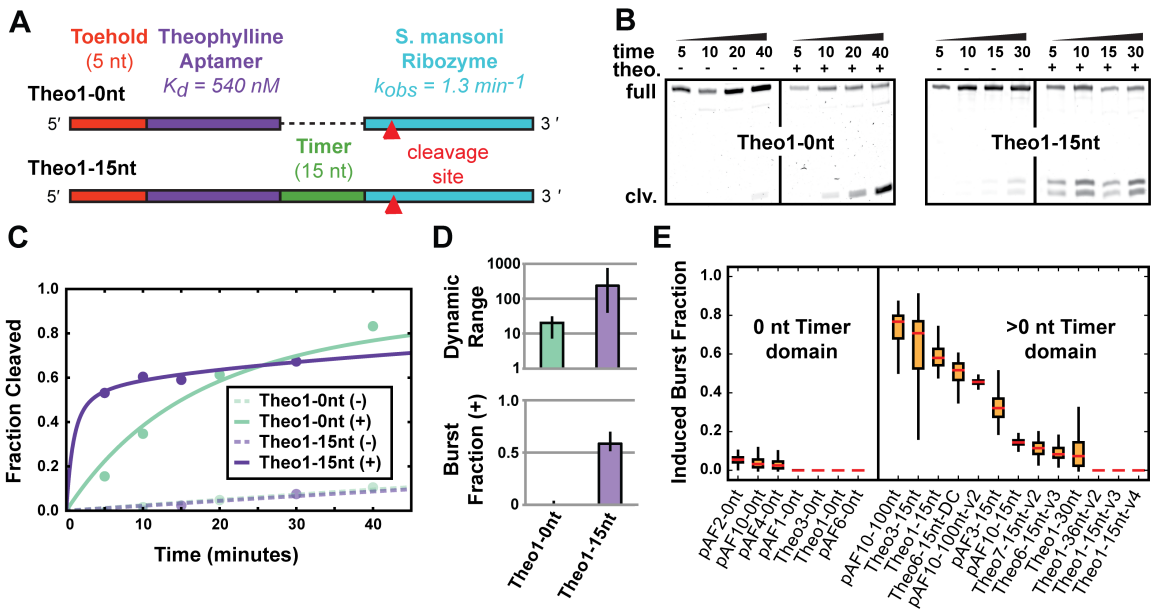


Figure 3. The Timer domain enables rapid, biphasic ligand-dependent function. A-D) A theophylline-responsive K-A designed with 0-nt Timer exhibits slow, monophasic, ligand-dependent *in vitro* co-transcriptional cleavage (Theo1-0nt). The addition of a minimal 15-nt Timer results in a K-A exhibiting rapid, biphasic, ligand-dependent cleavage and a dynamic range (kobs+/kobs-) of 237. E) For the 20 K-As with low background (UBF< 0.1, k_{avg-} < 0.01 min^{-1}), a Timer domain is necessary to access ligand-dependent burst phase cleavage.

MFEpath for screening RNA co-transcriptional folding

In order to design RNA aptamer-based switches that can function when produced *in situ*, or *in vivo*, it is necessary to be able to predict the relevant three dimensional structures that an elongating RNA molecule will adopt co-transcriptionally. As direct time-resolved prediction of three-dimensional structures of macromolecules (on the seconds scale) is currently computationally infeasible, it is necessary to abstract RNA three-dimensional structures to rapidly computable secondary structures. Due to the hierarchical folding of RNA, the secondary structure that an RNA molecule adopts dictates its accessible 3D folds, and it follows that a lack of the functional 2D structure precludes the formation of the functional 3D structure[44,45]. This allows the use of 2D objective functions to drive the screening for functional 3D structures.

Although the directional transcription of RNA molecules complicates structure prediction, it also provides an opportunity to encode kinetic control through a series of rapid nucleotide addition and structural rearrangement steps. This enables the exciting engineering prospect that the ligand-binding

and actuation reactions can be separated and tuned independently. Although there have been several algorithms developed to predict the co-transcriptional folding trajectories of RNA molecules, they predominantly are either unable to produce quantitatively accurate folding timescales, cannot be applied to long sequences, or are insufficiently transparent to allow for the type of quantitative analysis desired for our design-build-test-learn cycle[46]. To fill this need, we have created the MFEpath algorithm and computational framework for the predictable design of functional multi-state RNA devices. MFEpath works by screening RNA sequences for rapid co-transcriptional folding trajectories using secondary-structure prediction and Arrhenius-like interconversion kinetics (Figure 4A). The Arrhenius equation relates the activation energy of a reaction to exponential changes in reaction rate, and has been previously applied to the rates of RNA secondary-structural transitions. While empirical relationships between RNA structural rearrangement rate and barrier height have been described, and numerous barrier-prediction algorithms exist in the literature, there is a noted lack of broader computational tools for designing kinetically-functioning RNA devices using those calculated rate constants[39]. While there are more algorithmically complex tools for the prediction of ensemble co-transcriptional folding, we find that MFEpath finds the ideal balance of computational efficiency and output granularity to be optimal for screening rapidly-folding kinetic RNAs[47].

Once K-A candidates satisfying our thermodynamic objective functions were identified, we screened them for the ability to rapidly transition between the desired states during transcription. The candidate sequences possessed diverse *in vitro* co-transcriptional cleavage kinetics, as well as diverse MFEpath-predicted co-transcriptional folding characteristics. In order to predict coarse-grained folding trajectories during RNA transcription, MFEpath predicts the time to rearrange to the next MFE substructure using Arrhenius-like interconversion barrier heights ($\Delta G^{\ddagger}$), which correspond to the $\Delta\Delta G$ between the starting structure and the least stable structure along their refolding pathway. If rearrangement is calculated to be faster than the addition of the next nucleotide, structural rearrangement is allowed. If calculated to be slower, the transition is disallowed, the next base is added, and the analysis is performed for the new substructures. After the final base is added, the last barrier height ($\Delta G^{\ddagger}_{final}$) is used to predict the time needed for the RNA to convert from the co-transcriptional structure to the post-

transcriptional structure. This analysis is performed both for the folding trajectory in the presence and absence of the target ligand.

In order to determine functionally-relevant screening cutoffs for the $\Delta G^{\ddagger}$ values, we calculated the apparent $\Delta G^{\ddagger}$ that would result in structural rearrangement kinetics of the same rate using Equation 3. Assuming an elongation rate for T7 RNA polymerase of 230 nt/s, we calculated that $\Delta G^{\ddagger}$ values of < 6.4 kcal/mol would occur faster than the addition of the next nucleotide. Assuming an upper limit on hammerhead ribozyme cleavage rate of 5 min^{-1}, we calculated that $\Delta G^{\ddagger}$ values < 11.1 kcal/mol would result in structural rearrangements faster than ribozyme cleavage[48]. We then utilized the calculated $\Delta G^{\ddagger}_{final}$ values to predict function within our kinetically-characterized K-As (Figure 4B), as they should indicate how long a transcribed RNA will take to fold into its MFE structure post-transcription, and therefore whether the K-As will undergo the rapid structural rearrangements necessary to display high DRs. For example, MFEpath predicts the Theo5-0nt K-A to become kinetically trapped into the catalytically-active state regardless of the presence of the ligand, due to a large $\Delta G^{\ddagger}_{final}$. As expected, this device displays extremely high background cleavage, and therefore a negligible DR of 1.1. In contrast, in the absence of ligand, MFEpath predicts the Theo1-30nt K-A will rapidly rearrange into the inactive conformation co-transcriptionally and, as predicted, the K-A displays low background cleavage of 0.01 min^{-1}. However, in the presence of ligand, MFEpath predicts that the formation of the ribozyme domain will occur more slowly than the rate of ribozyme cleavage, and thus no rapid cleavage will be observed. This too holds true, as the device demonstrates a modest increase in the rate of non-burst cleavage, resulting in a DR of 5.6. Finally, MFEpath predicts that the Theo1-15nt K-A will rapidly reach the desired structure states for both the ligand-dependent and -independent pathways. The device displays a low background cleavage rate, as well as rapid, ligand-dependent, burst phase cleavage resulting in an unprecedented DR of 237. These results illustrate how in order to attain large DRs, K-A devices must possess small folding barriers for both the ligand-dependent and -independent folding trajectories.

Equation 3. [39]

$$\tau = 10^{\left(\frac{8}{11} * \Delta G^{\ddagger}\right) - 7}$$

To validate that our calculated threshold values are accurate for our experimental conditions, we analyzed all potential $\Delta G^{\ddagger}_{final}$ cutoffs for the ability to split K-As into ‘pass’ and ‘fail’ categories. K-As

possessing $\Delta G^{\ddagger}_{final}$ values below the cutoff should refold more rapidly than the competing reaction (i.e. transcript elongation or ribozyme cleavage). As we expect that the optimal barrier thresholds for screening K-As will be the barrier heights corresponding to the rate of the competing reaction, we should be able to map the best-performing barrier cutoff to the most physiologically-accurate barrier height. In order to compare cutoff values, we define 'performance increase' as the ratio of the median cleavage rate (k_{avg+} or k_{avg-}) for devices that pass the threshold, to the median of those that fail. For the plus-ligand screening, the one of the main peaks in performance increase occurs precisely at our *a priori* prediction (Figure 5A). For the minus-ligand screening, the maximum performance enhancement occurs slightly below our chosen value, which may either indicate a slight inaccuracy of our chosen threshold, or may suggest that there are additional factors that must be accounted for when predicting rearrangement from the s4 to s6 states co-transcriptionally (Figure 5B).

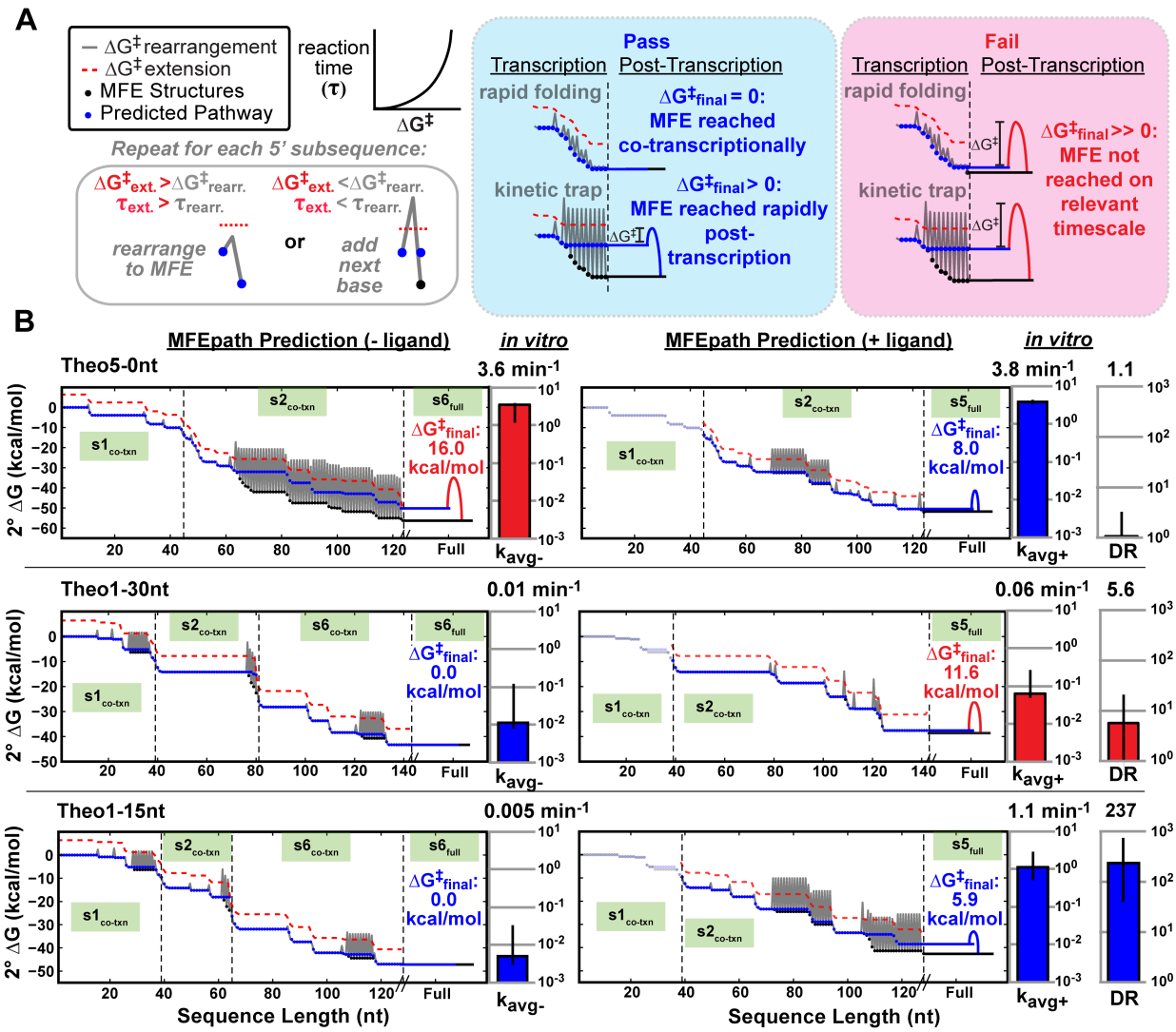


Figure 4. Reliable multi-state co-transcriptional folding design enables K-A engineering. A) In order to predict coarse-grained folding trajectories, MFE path predicts the time to rearrange to the next MFE substructure from Arrhenius-like interconversion barriers. If rearrangement is expected to be faster than the addition of the next nucleotide, structural rearrangement is allowed. If slower, the transition is disallowed, and the next base is added, and the analysis is performed for the new substructures. After the final base is added, the last barrier height is used to predict the time needed for the K-A to convert from the co-transcriptional structure to the post-transcriptional MFE structure. B) Selected K-As demonstrate the significance of final output barrier. In the absence of ligand, a low final barrier indicates a co-transcriptional interconversion to the OFF state, resulting in low background cleavage. A high barrier results in rapid cleavage and a loss of potential DR. In the presence of ligand, a high barrier indicates that the cleavage rate is limited by structural interconversion. A low barrier indicates that the interconversion occurs on a faster timescale than cleavage, and the intrinsic cleavage rate of the ribozyme becomes limiting. When both barriers are low, very large dynamic ranges become possible.

The calculated thresholds hold true when applied to the devices in aggregate as well. The K-As that possess smaller -ligand $\Delta G^{\ddagger}_{final}$ values (B2* barrier) for aptamer deformation have smaller uninduced burst fractions (UBFs) (Spearman rho=0.34 , p=0.02), and all K-As whose values lie above 6.4 possess large UBFs (Figure 5D). For K-As with low background cleavage (to remove devices that attain burst cleavage due to very small B3 barriers), the devices that possess smaller +ligand $\Delta G^{\ddagger}_{final}$ values (B1 barrier) have larger IBFs (Spearman rho= -0.74 , p=0.004), and all whose values lie above 11.1 possess negligible IBF values compared to those with values lying below (Figure 5C). It is important to note that although many of the K-As that pass MFEpath's -ligand screening possess small burst fractions in the absence of ligand, there remain a sizable number of such K-As that possess large UBF values. This suggests that there are additional factors that dictate the B2 barrier that should describe our system.

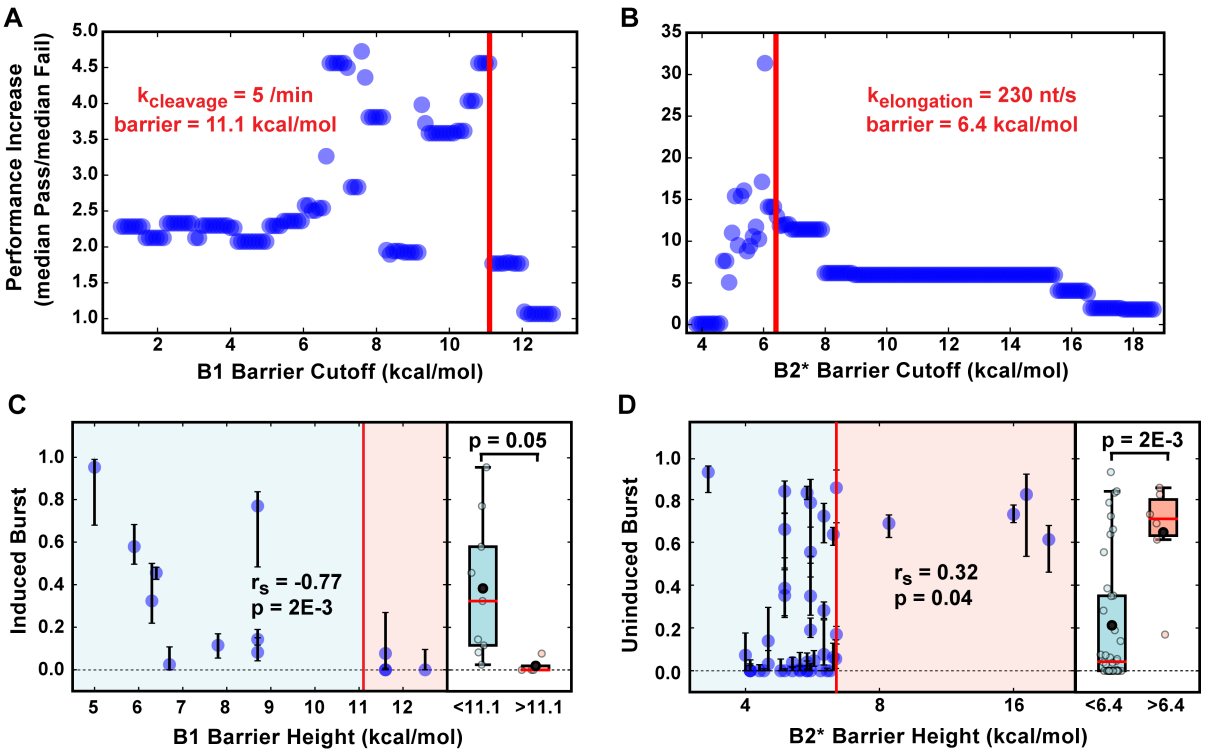


Figure 5. Thresholds for structural rearrangement. A) Performance increase (the ratio of the median cleavage rate (kavg +) for devices that pass the threshold, to the median of those that fail) is at a maximum at our predicted value, implying our *a priori* predicted barrier height (11.1 kcal/mol) is ideal for screening K-As. Dashed line represents the predicted barrier of the rate of nucleotide addition. B) We define performance increase as the inverse of part A, with respect to induced cleavage (kavg -). The

maximum is near our implemented elongation barrier (6.4 kcal/mol), though slightly below. This suggest our estimate is good, but that there may be additional factors to consider. C) Predicted values for the B2* of aptamer deformation partially explain observed uninduced burst fraction (UBF). D) K-As with Timers and low background (UBF< 0.1, k_{avg-} < 0.01 min^{-1}) only display significant burst phase kinetics when B1 is below our predicted threshold.

Toehold-mediated strand displacement to reduce leakage

The most significant impediment to large-DR aptazymes is undesired cleavage in the absence of target ligand[49]. More specifically, rapid, undesired cleavage is virtually incompatible with functional K-As. As seen above, although MFEpath's barrier height predictions allow for identification of K-As with a lower probability of possessing large uninduced burst fractions (UBF), they alone clearly do not explain the UBF in all cases. One likely reason for this is that even though a threshold is used as a pass-fail criterion for structural rearrangement, a predicted barrier height (and therefore reaction rate) nearly identical to the threshold, would result in a PASS within the MFEpath algorithm, but would result in an ~50/50 split between molecules that structurally rearranged, versus those that became kinetically trapped. Thus, in order to ensure extremely low UBF values, a structural rearrangement significantly faster than nucleotide addition is likely necessary.

Toehold mediated strand displacement (TMSD) is a well-known molecular mechanism in the field of DNA nanotechnology that can accelerate the rate of intermolecular strand exchange by up to 10^6-fold[50]. It has recently been utilized, with great success, to increase the effectiveness of trans-acting genetic RNA 'toehold switches'[5,51–53]. By implementing the TMSD mechanism to accelerate the intramolecular structural rearrangement from state s4 to state s6 it may be possible achieve extremely low background signal, and therefore unprecedented dynamic ranges. The K-A molecular architecture is capable of utilizing the TMSD mechanism, as the P1 aptamer stem is analogous to the initial duplex, and the 5' end of the ribozyme domain acts as the invading strand (Figure 6A). Thus, the K-A's rearrangement toehold initiates structural rearrangement by binding to its target.

There are three quantitative predictors of traditional intermolecular TMSD: 1. Stability of the toehold-target duplex. 2. The barrier height of the steps of the displacement reaction. 3. The concentration of the two species[54,55]. By analogy, the effectiveness of intramolecular TMSD to enhance

the rate of structural rearrangement in our system should be predictable from the stability of the toehold-target duplex, the free energy barrier height for the structural rearrangement, and relative volume that the toehold and target domains can explore.

To investigate whether K-A UBF could indeed be predicted from the analogous TMSD parameters, we began by calculating the expected stability of the toehold-target duplex (Figure 6B). To do so, we utilized the ViennaRNA folding package's RNAeval algorithm with constraint folding to evaluate the stability of the toe-target duplex[56]. In order to account for frustrating structure formed internally to the toehold, any bases predicted to be base-paired prior to duplex formation were considered unable to contribute to the toe-target duplex. Toehold-target duplex stability displays a significant correlation with K-A UBF (Spearman rho = 0.3 p = 0.05), but as expected does not satisfactorily explain all of the data.

Next, we characterized the barrier heights for the structural rearrangement that occurs post-duplex formation. As structural rearrangement from the active to inactive folding trajectory can occur at more than one transcriptional step, the B2 rearrangement barrier is the combination of all such possible barriers. Rearrangement barriers are only considered for steps in which the toehold and target are not engaged in frustrating structure, and therefore have a duplex stability less than zero kcal/mol. Barriers were combined utilizing Equation 4. This composite B2 barrier has a highly significant rank correlation with the observed UBF (Spearman rho = 0.57, p = 5E-5), and displays a stark threshold response just above 1.5 kcal/mol (Figure 6C). Although all K-As with a B2 barrier smaller than 1.5 kcal/mol display UBF values less than 0.1, there remains some unexplained variation.

Equation 4. $$B2 = ln\left(\sum_{n=branch}^{end} e^{-B2_n}\right)$$

We reasoned that if the K-A structural rearrangement was proceeding via TMSD, that its rate would be proportional to the effective concentration of the toehold and target domains. As the TMSD reaction is intramolecular, the effective concentration should be inversely proportional to the 3-dimensional volume that the two domains can explore. Assuming that single-stranded RNA acts as a flexible linker, this volume should be proportional to the cube of the length of single-stranded RNA linker between the toehold and target. To predict this length, we implemented a simple algorithm for the coarse-grained estimation of the maximum linear distance between two nucleotides within a structured RNA. In effect, the algorithm counts the number of unstructured bases between the two, while skipping over any

self-contained helical elements. For the K-As possessing small B2 values, the remaining variation in UBF value correlated with this predicted distance (Spearman rho = 0.58 p = 0.006) (Figure 6E). For example, the Theo3-30nt K-A and the Theo3-30nt-v2 K-A are identical except for the sequence and structure of their 30 nt Timer domain. Theo3-30nt is expected to form a hairpin within its Timer domain that brings the toehold and toehold target into close proximity prior to structural rearrangement (Figure 6D). By contrast, the Timer domain of Theo3-30nt-v2 is expected to remain unstructured, allowing the toehold and target to explore a large volume, and decreasing the effective concentration of the two species. Thus, it appears that intramolecular TMSD is a viable and predictable mechanism for the enhancement of structural rearrangement, and reduction of ligand-independent burst phase cleavage.

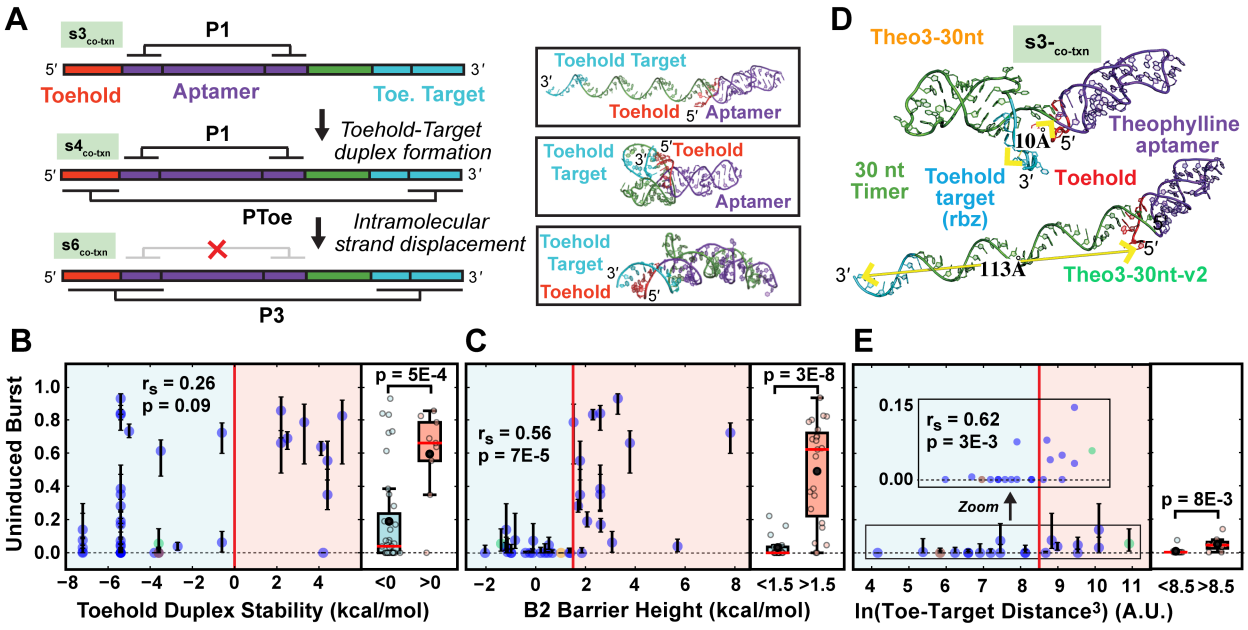


Figure 6. Intramolecular toehold-mediated strand displacement (TMSD) reduces leak. A) Cartoon mechanism for intramolecular TMSD. B) Devices predicted to form stable toehold-target duplexes display significantly reduced uninduced burst fraction (UBF) relative to those not expected to form favorable contacts (supplementary information). C) The B2 barrier, which combines the barrier heights of all possible transitions from ON to OFF pathways after the branch point, displays a threshold below which devices have UBFs below 0.1. D) Representative 3-dimensional predictions of the toe-target distances for identical K-A's containing Timer domains with different amounts of predicted structure. Structures generated from MFEpath secondary-structures and 3-D structures produced by RNAcomposer[57]. E) For devices below the threshold in part D, devices with small predicted toe-target distance display lower UBFs

than those with larger predicted distances. 3-D images produced using PyMol.

Predictable K-A design

Ultimately any computational methodology to design aptazymes will only be useful if it can be relied upon to consistently produce devices with the large DRs necessary for downstream applications. To that end, we looked at the impact that the various design metrics described above have on the identification of high-DR K-As (Figure S5B). Interestingly, despite providing access to ligand-dependent burst phase cleavage, the addition of a Timer domain does little for the overall success of a device within our data set. This may be in part because the increased toehold-target distance that often accompanies a Timer domain, thus increasing the UBF observed. As expected, toehold stability (< 0 kcal/mol) plays a very large role in predicting the function of a candidate K-A. This effect becomes even more significant when B2 (which incorporates toehold stability) is utilized for screening (B2 < 1.5 kcal/mol). Interestingly, despite increasing the mean of the screened populations, through increased burst phase cleavage, a small B1 barrier (< 11.1 kcal/mol) decreases the median DR observed. This result may be somewhat idiosyncratic, as many of the devices with large B1 barriers also happened to have very low B2 barriers, and no measured UBF. K-As possessing toehold-target *natural log(linear distance^3)* values of less than 8.5 (unitless) have improved mean and median values, though the differences are not overwhelming without first ensuring that the K-A possesses a small B2 barrier. When all of the screens are implemented simultaneously, the population possesses a median DR of ~10, and an extremely high mean DR of ~40. These are dramatic improvements over the population of K-As that fails even one of the criteria ($p = 0.02$). This seems to suggest that the screens are indeed synergistic, and will aid in the identification of high-DR K-As in the future.

The main area in which the prediction of K-A behavior could improve is in the description of the B3 barrier height (Figure S5A). As the upper limit on cleavage is ~5 min^{-1}, a value only slightly faster than +ligand cleavage of the fastest K-As, the main available avenue for accessing DRs approaching the theoretical limit is to dramatically reduce the ligand-independent cleavage rate. As shown in Equation 2, k_{avg} is the weighted average of the burst and slow rates multiplied by their relative fractions. As it appears that TMSD has allowed us to design K-As with extremely low UBFs, the only avenue remaining to

decrease k_{avg-} is to decrease k_{slow-}. Within our interpretation of the K-A system, the height of the B3 barrier should dictate this k_{slow-} rate constant. Indeed, we observe that there is a statistically significant rank correlation between the computationally predicted B3 barrier height and the k_{slow-} rate. However, it is apparent that the current calculations for B3 barrier will not be sufficient, as K-As with predicted B3 heights varying by 6 kcal/mol display the same k_{slow-}.

The limitations of B3 prediction likely arise from two main issues: barrier height algorithm limitations, and structure state selection difficulty. As findpath, the algorithm implemented in MFEpath, only considers direct refolding pathways (those that only contain base pairs in either the initial or final structure) it is likely to overestimate the barrier heights for real pathways, which usually undergo indirect refolding[58,59]. Also, although our initial K-A design identifies the most thermodynamically stable structure that contains the ribozyme (s5), it is possible that the correct B3 barrier height would be one from s6 to a less-stable, but more rapidly accessed, structure that also contains the ribozyme. An additional current limitation is that the *in vitro* cleavage assay cannot statistically differentiate between cleavage rates slower than 10^{-3} min^{-1}. Thus, to validate extremely low k_{slow-} values, the duration of the assay itself will have to be extended. However, the ability to consistently design K-As with DRs of 1000 would represent a significant improvement to the state of the art for most biosensing applications.

K-A ligand sensitivity tuning

One critical aspect of producing ligand-responsive switches is ensuring that they respond at concentrations that are relevant for subsequent applications. Designing switches that show switching behavior below cellularly-toxic, or insoluble, ligand concentrations has proved problematic in the past, and may be a principal reason that more aptazymes have not been identified to date[15]. As such, it is critical that any methodology for the design of such switches allows the sensitivity to be rationally tuned. It has been suggested that the variable-length hairpins observed between the aptamer domain and expression platform in natural riboswitches may exist in order to serve this purpose[60]. By increasing the amount of time that the aptamer is available co-transcriptionally, through additional time the polymerase spends transcribing the hairpin, they may in turn increase the riboswitch's sensitivity to ligand.

To interrogate the impact of Timer domain length, and therefore the duration of the binding window, on K-A ligand-sensitivity we designed four additional K-As based on the pAF10-0nt device (Figure 7A). The K-As contain identical sequences to pAF-10-0nt except for their Timer domains, which vary in both their length and sequence, and all possess DR values greater than 9. Two devices containing different 15 nt-long Timers were designed, as well as two with different Timers 100 nt-long. To determine the impact that these Timers had on device sensitivity, we performed the described co-transcriptional cleavage assays at six different concentrations of pAF for each device. To evaluate the ligand sensitivities of multiple devices we chose to compare the half maximal effective concentration (EC_{50}), with respect to pAF. Our *a priori* EC_{50} value predictions were calculated from Equation 5, which assumes that ligand binding is a pseudo-first order irreversible process, that halts when the ligand binding window closes. In Equation 5, k_{on} is the experimentally-determined association rate constant for the independently assayed aptamer-ligand pair, [L] is the concentration of ligand, k_{pol} is the literature value for the *in vitro* elongation rate of T7 RNA polymerase at 37C, and nt is the nucleotide length of the inserted Timer domain.

Equation 5.
$$EC_{50} = -\frac{\ln(0.5)*k_{pol}}{k_{on}*nt_{timer}}$$

With the exception of the parental device, which does not have a Timer domain and therefore is not expected to bind co-transcriptionally, all of the K-As possess B1 barriers well below the cutoff of 11.1 and are therefore expected to display ligand-dependent burst phase kinetics. As not all of the pAF10 K-As displayed burst phase kinetics, and as very high k_{burst} values can possess large errors due to manual pipetting limitations, we decided to agnostically select whichever parameter (k_{avg+}, or IBF) provided the best r^2 value for each K-A when fit to the 2-parameter binding Equation 6. The fit value for the maximum signal was used to normalize the data, which was subsequently fit to a 1-parameter binding Equation 7.

Equation 6.
$$signal = \frac{signal_{max}*[L]}{[L]+EC_{50}}$$

Equation 7.
$$f_{bound} = \frac{[L]}{[L]+EC_{50}}$$

We observed that as the length of the inserted Timer increased, so too did the measured EC_{50} for the K-A, resulting in fit EC_{50} values spanning more than two orders of magnitude (8.4 mM to 4.4 M) (Figure 7B). Excitingly, the EC_{50} values observed for pAF10-100nt and pAF10-100nt-v2 of 9.3 mM and

8.4 mM both fall within three-fold of our *a priori* expectation of 3.3 mM (Figure 7C). While somewhat less accurate, the EC_{50} values observed for pAF10-15nt and pAF10-15nt-v2 of 40 mM and 142 mM fall near the prediction of 22 mM. No prediction for the EC_{50} of the pAF10-0nt K-A was made as, without a Timer domain, it is expected to bind pAF through post-transcriptional thermodynamic routes for which no straightforward EC_{50} estimation methods exist.

It is relevant that our *a priori* EC_{50} predictions were lower than observed for the four K-As displaying burst phase kinetics. This suggests one of three main possibilities: 1. The characterized aptamer-ligand association rate measured in isolation is higher than within a K-A. This is certainly possible as weak, transient, interactions with the rest of the nucleotides may reduce the availability of the aptamer for binding. 2. The literature value for T7 elongation rate is lower than the actual elongation rate in our experiment. While this is possible, the literature value utilized is already on the high end of those reported. 3. The binding window is shorter than the transcription time of the Timer domain. This again is very likely, as MFEpath predicts that Timer domain is already partially transcribed by the time the aptamer domain becomes properly folded.

Although the idea of ‘controlling any gene with any molecule’ is an extremely ambitious and likely unattainable goal, we believe that the work done to date illustrates the potential RNA switches hold in approaching that aim. We believe that the performed and proposed research will provide significant advances to RNA design in a number of critical ways. First, any broadly-applicable computational strategy to design RNA switches is a major advance not only for the ability to design high performance RNA devices, but additionally for the lack of experimental expertise it demands from the end users. By moving the design labor from researchers to ever-cheaper computational resources, K-As will be available to scientists who lack the technical proficiency (or resources) to perform the otherwise-necessary cellular screening experiments. Additionally, the design rules provide an excellent starting place for the computational design of other types of RNA switches. The extremely low background cleavage rates enabled by the intramolecular TMSD mechanism should enable applications in which leaky background signal cannot be tolerated, as in the described dCas9-based system. The demonstrated ability to tune the EC_{50} of K-As utilizing Timer domains provides another significant step forward for the field of RNA design,

as it provides a framework for both the quantitative *a priori* predictions of switch sensitivity, as well as the rational tuning of switches custom-tailored to their application.

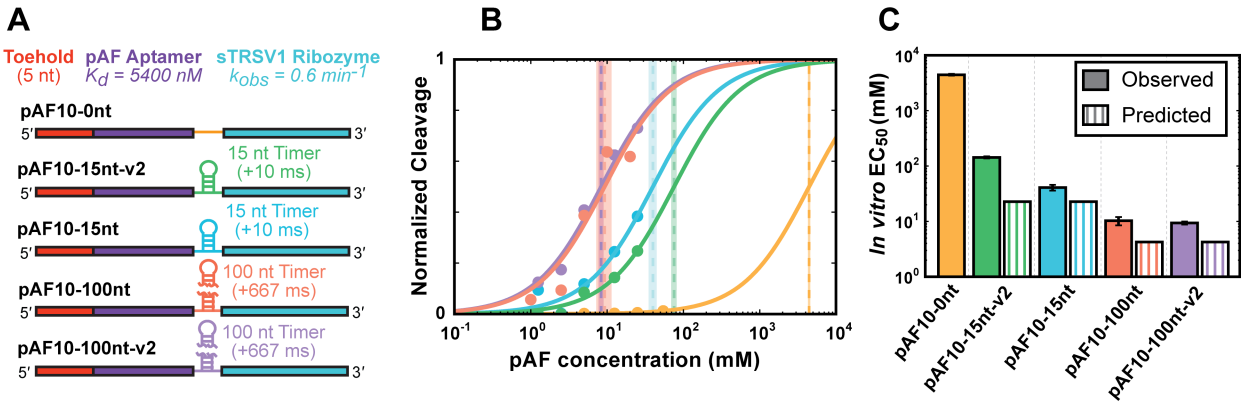


Figure 7. Longer Timer domains increase the ligand sensitivity of K-As. A) Five functional (DR > 9) K-As were designed from one parental device (pAF10-0nt), varying only in the length and sequence of the Timer domain. B) Normalized output vs. ligand concentration is fit to a standard 1:1 binding curve. EC_{50} values, indicated by vertical lines, decrease as Timer length increases. C) Fit and predicted EC_{50} values for the five devices. Error bars represent the standard deviation of the fits shown in B). Predicted EC_{50} values, derived from known rate constants and the length of the Timer domain agree well with observed values.

One class of genetic controller our lab has previously demonstrated in *E. coli* is that of an aptazyme-regulated expression device (aRED)[6]. The mechanism utilized in an aRED is that of variable RNA degradation rates. When a ribozyme (or aptazyme) cleaves in the 5'-UTR of bacteria, the downstream gene is then terminated with a 5'-hydroxyl group, instead of a 5'-triphosphate[61]. This has implications for the degradation rate, as an exonuclease recognizes and removes 5'-triphosphate groups. In the absence of such a group, the mRNA is degraded instead through endonucleolytic pathways instead. This slower degradation results in up to a 6-fold increase in the half-life of the RNA, and therefore steady-state protein expression level. While preliminary efforts to incorporate K-As into an aRED showed promising results, subsequent analysis yielded unsatisfying and contradictory responses. One significant confounding factor is the adjacency of the aptazyme to the ribosome binding site. As ribosome binding site structure is known to be one of the primary factors in determining prokaryotic translation rates, any changes in ribosome binding site (RBS) structure that occur as a result of ribozyme cleavage or structural

rearrangement are likely to have additional, unintended, impact on protein expression levels. Considering that the fold change of protein levels in response to an aRED is ~6, and the fold change of protein levels in response to changes in RBS structure are several orders of magnitude, it is very possible that the unintended effect will have a greater impact than the intended one.

Another piece of evidence that aptazymes may not be the ideal biochemical mechanism to utilize for genetic control is their surprising absence in nature. While riboswitches that dynamically regulate gene expression levels are ubiquitous in natural bacteria, and a natural ribozyme that uses a small molecule as a co-factor has been characterized, aptazymes where the self-cleaving ribozyme's cleavage activity is controlled by the binding state of an RNA aptamer domain have yet to be found[62]. While there are several hypotheses as to why they are not more common, it appears that natural systems have found that other mechanisms are preferable, such as riboswitches that control the folding of a transcriptional terminator or RBS.

Supplementary Materials

Small-molecule ribosensors (performed by Cassandra Burke)

To determine whether MFEpath provides useful information about the co-transcriptional folding trajectories of other types of RNA molecules, we applied it to the design of small-molecule responsive RNA ribosensors. Kinetically-controlled ribosensors are comprised of *in vitro* selected RNA aptamers for molecular recognition coupled through a Timer domain to a trans-acting TMSD toehold to generate FRET-based fluorescent outputs. Like K-As, the ligand binding occurs during the binding window, and as such is sensitive to the specific co-transcriptional folding pathway the molecule adopts (Figure S1A).

To validate that ribosensors bind their target ligand co-transcriptionally, our lab performed co-transcriptional *in vitro* assays in which the T7 RNA polymerase was rapidly digested by proteinase K. Compared to reactions in which no proteinase K was added, the rate of fluorescence generation was reduced to background levels. To demonstrate that the binding window can be elongated by reducing the polymerase elongation rate (in addition to Timer length as demonstrated with K-As), our lab reduced the concentration of NTPs in the transcription reaction, previously demonstrated to achieve the desired rate reduction, and observed a two-fold increase in the EC_{50} of the ribosensors.

One of the key metrics of MFEpath is the pathway convergence, which is the fraction of the positions in the folding trajectory in which the predicted structure is the MFE of the current subsequence. For multiple reasons, we expect MFEpath's predictions to be most accurate when the MFEpath convergence is high. When this metric was used to screen pAF-responsive ribosensors, the fraction of screened devices that were functional increased from < 60% to 100% (Figure S1B). These results are extremely encouraging for MFEpath as a broadly-applicable tool for the screening of rapid co-transcriptional RNA folding.

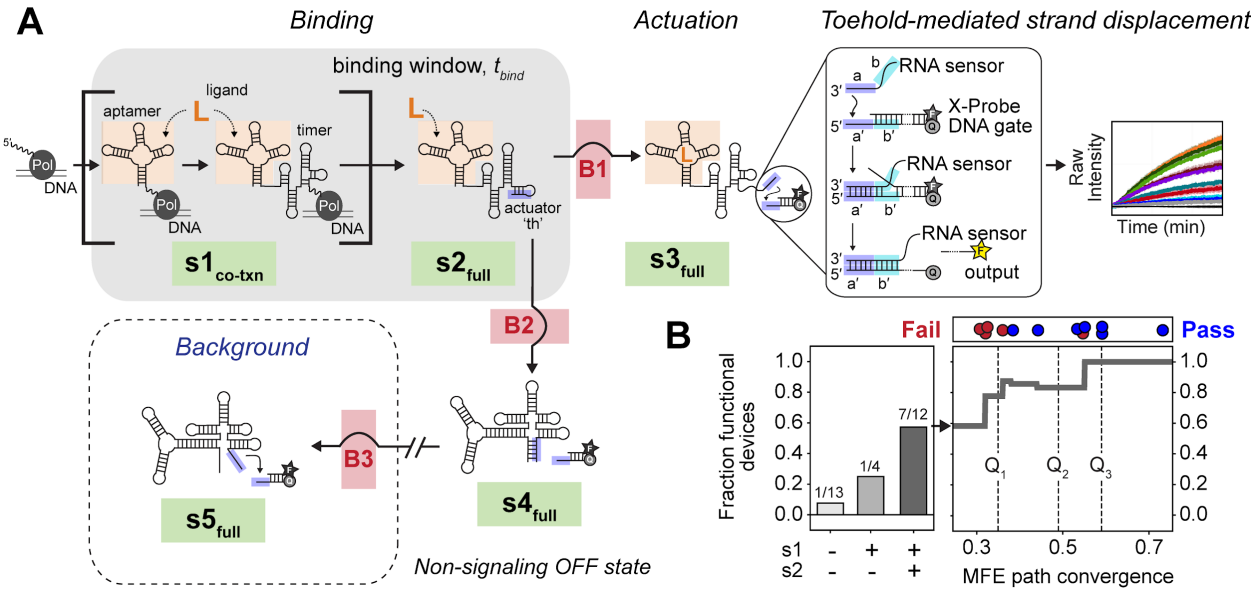


Figure S1. MFEpath aids in the prediction of functional Kinetically-controlled ribosensors. A) Kinetically-controlled ribosensors convert co-transcriptional ligand-binding to TMSD-mediated unquenching of fluorescent outputs. Like K-As, their performance is expected to be dictated by three main structural conversion barriers. B) The fraction of pAF-responsive ribosensors that are functional is plotted against the predicted accessibility of the $s1_{co\text{-}txn}$ and $s2_{full}$ structure states (light to dark gray bars) and MFE path convergence; functional (blue) and non-functional devices (red) are indicated.

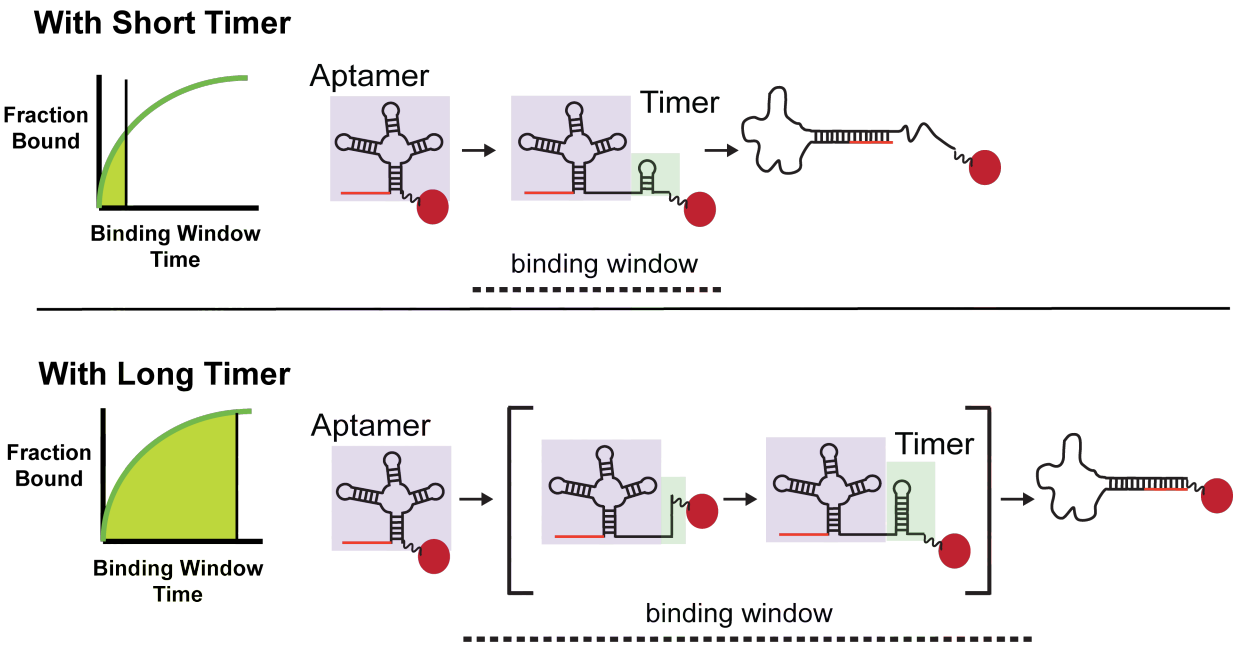


Figure S2. Timer domains increase the ligand binding window. As the length of the Timer domain increases, so too does the time it takes to transcribe it. This results in a greater fraction of the aptamers bound when the binding window closes, and therefore increased sensitivity to the target molecule.

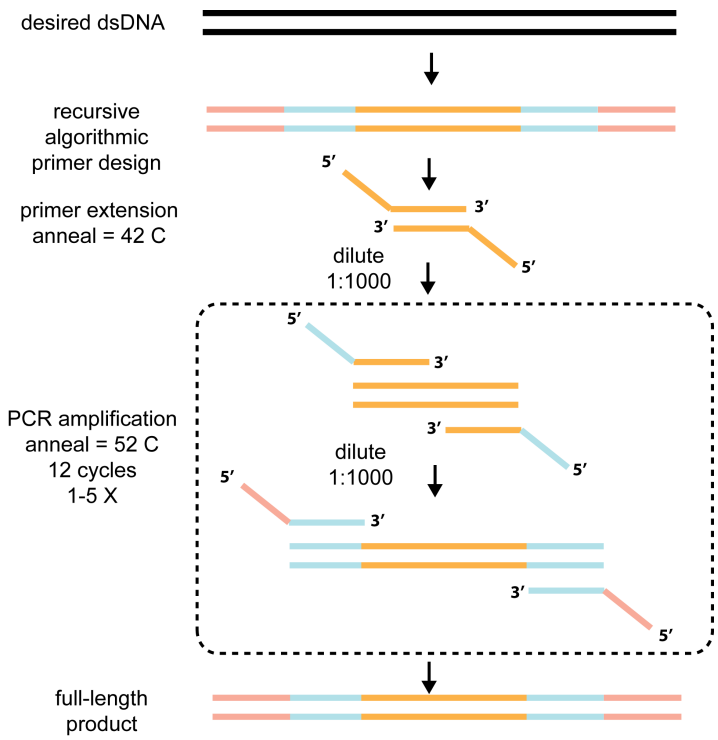


Figure S3. Assembly scheme for K-A DNA transcription templates. First, the desired sequence is input into a recursive algorithm that generates a series of short assembly primers. The full-length sequence is assembled through a series of primer- or overlap-extension PCR reactions and template dilution. The automated process yields primers that possess an annealing temperature of 52 C, which enables numerous sequences to be batched together in a single thermocycler.

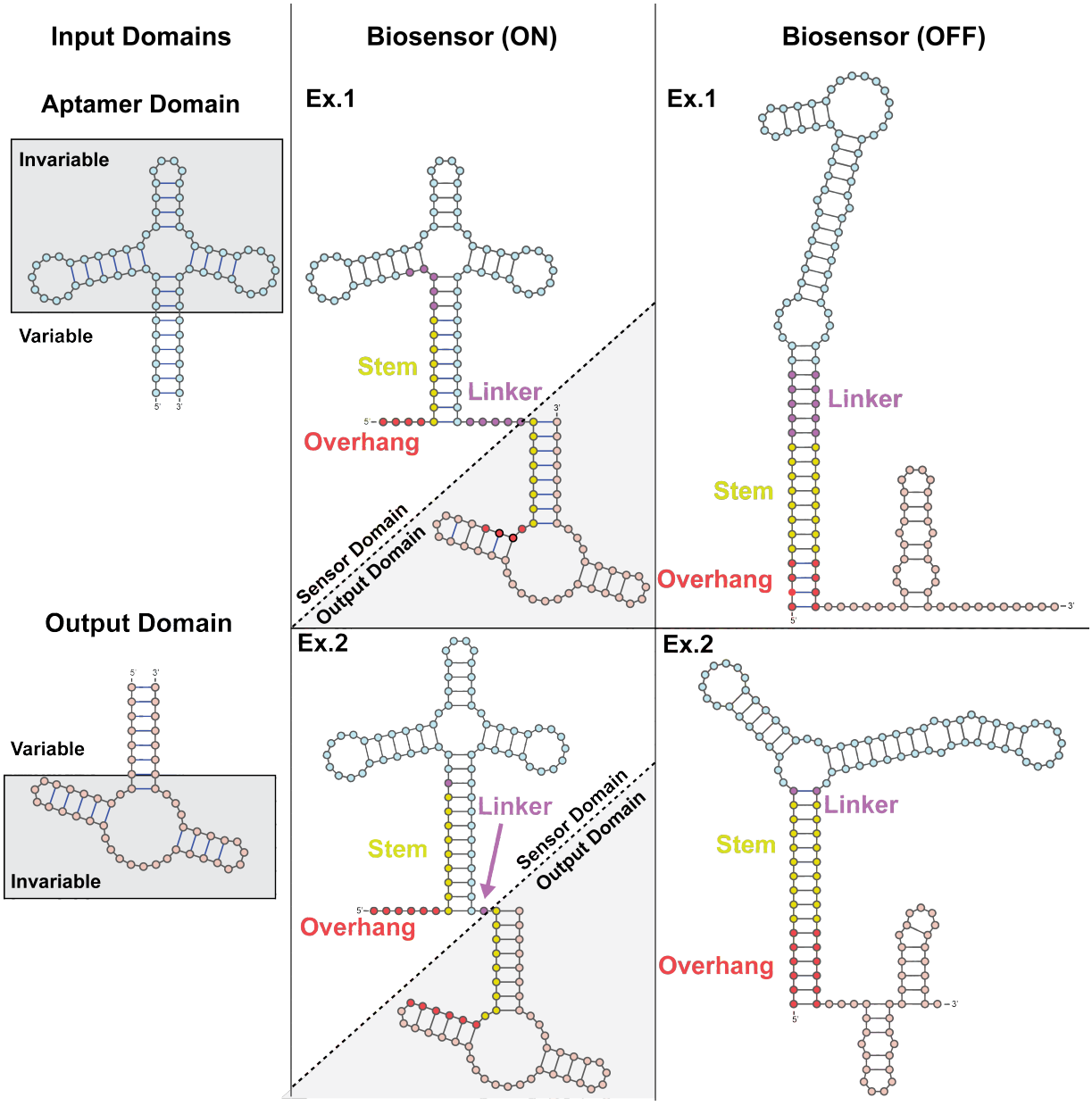


Figure S4. Examples of the kinetic biosensor molecular architecture. Variations in the length of the Overhang, Stem, and Linker sequences couple the structures of the Sensor Domain and the Output Domain.

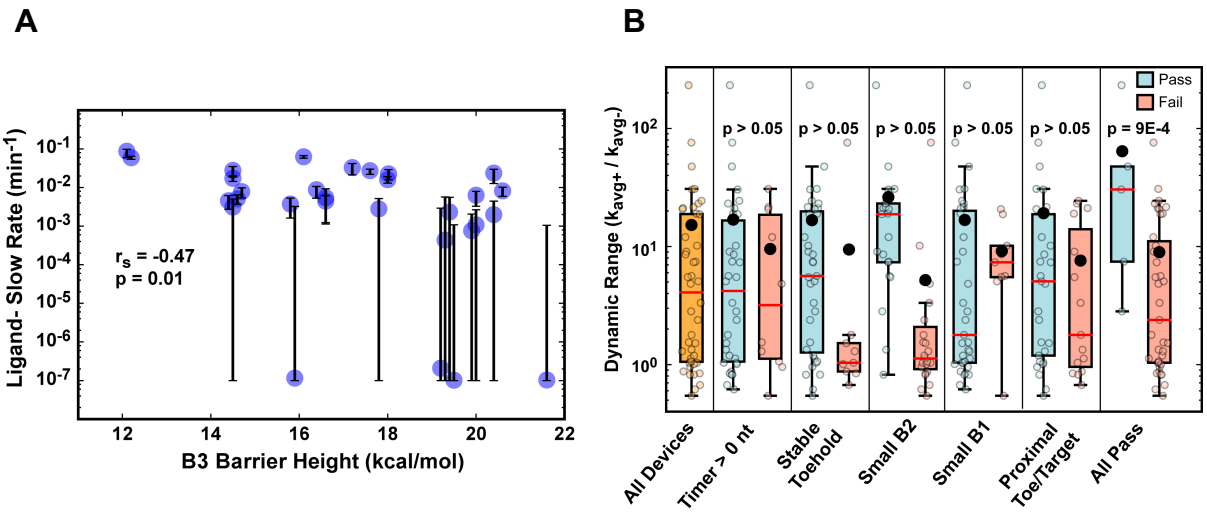


Figure S5. Computational screening metrics predict K-A activity. A) Ligand independent slow rate decreases as predicted B3 height increases. Error bars represent bootstrapped 95% confidence intervals of the fit *in vitro* data. Rate constants below 10^{-3} min^{-1} cannot be determined with statistical significance, limiting prediction of extremely low rates. B) Impact of K-A screening steps on the performance of the resulting populations. Red lines represent the median of the population, while green circles represent the means. Although all screening steps improve the mean dynamic range of the surviving sequences, p-values are > 0.05 for all means except when all screening steps are implemented (p = 0.02).

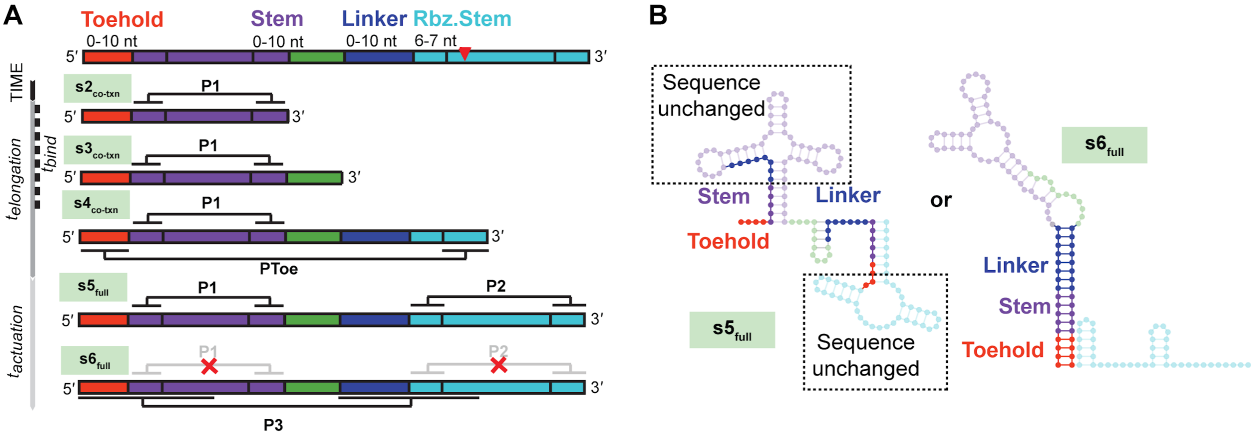


Figure S6. Structure states analyzed in the screening of K-A candidates. A) Analyzed during the thermodynamic stage of K-A design. Sub-sequences are screened for their ability to stably form target structure states, mimicking co-transcriptional folding. B) The three regions of complementary sequence (toehold, linker, and stem) utilized in the design process to create a selectively-inactive MFE structure.

Name	Aptamer	Ribozyme	Sequence
pAF_sTRSV1_1020,4	*pAF*	sTRSV1	GGGUGAGCAUUGCAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUGCAAUGCCUGUUAUGCAAUGCUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAGCAUUGC
pAF_sTRSV1_1200,5	*pAF*	sTRSV1	GGGUGAACGCGCAAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUUGCGCGUCUGUUAUUGCGCGUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAACGCGCA
pAF_sTRSV1_155	*pAF*	sTRSV1	GGGAAUGCCUCAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUGAGGCUCGAGAAAGAAUUCUGUUAUGAGGCAUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAUGCCUC
pAF_sTRSV1_1676,5	*pAF*	sTRSV1	GGGUGAGCAUUCAAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUUGAAUGCUGUUAUUGAAUGCUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAGCAUUCA
pAF_sTRSV1_1676,5-279	*pAF*	sTRSV1	GGGUGAGCAUUCAAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUUGAAUGCUUCUCAACAAGAUGAUGUGAAGUGGCUUUAAUGUGCGCGUGGGAAUAUUUCCCCCGUGAUUGGUUAUAGUUAAGCGUGACAGCGCGGAAUCGCGACGCUCUGUUAUUGAAUGCUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAGCAUUCA
pAF_sTRSV1_1676,5-8	*pAF*	sTRSV1	GGGUGAGCAUUCAAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUUGAAUGCUGUAAAUCUAUCUCCAAGAGACAAGGUGUGAUAAUGGGAGCAAUGGCCCUUUCAACAUUGCUACUGUCGAGGUGAUAGCACAACUAGACCCCGGUUGGCCUGUUAUUGAAUGCUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAGCAUUCA
pAF_sTRSV1_309,2	*pAF*	sTRSV1	GGGAAGUGAACAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUGUUCACUUGUUAUGUUCACUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAGUGAAC
pAF_sTRSV1_309,2-878	*pAF*	sTRSV1	GGGAAGUGAACAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUGUUCACUUGAUGCUAAGCAGCUUUGCAUUCGGCAGGGUCUUCUAGGUCCGACGCAGCAGAGGAUAAAAUGGUAUCCGUUAGGGCUAAUGAUGCUUAGAUUGACAUUAUGUUAUGUUCACUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAGUGAAC

pAF_sTRSV1_673,9	*pAF*	sTRSV1	GGGAAGAUUCUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUAGAAUCUGUUAUAGAAUCUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAGAUUCU
pAF_sTRSV1_673,9-961	*pAF*	sTRSV1	GGGAAGAUUCUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUAGAAUCUCACACUAGGUGUACCUACCCUGAUGUCGCGUACACGUUCAAGCCCGGACGAAAGACUCAACUUCGUCCGCGUAUCUCCACUAAAUACGGGCGACUAAAAUGUUAUAGAAUCUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAGAUUCU
pAF_sTRSV1_673,9-966	*pAF*	sTRSV1	GGGAAGAUUCUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUAGAAUCUACAUGAGAUGACAAAGAUCGCACGAUGGGGGAGAUACAAACAACCGCGCCUAAAUAUACACGAUCCAUAUGAAGGCUCCUGACCACCGUAUGCGCUGACAGUUAUAGAAUCUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAGAUUCU
pAF_sTRSV1_693,9	*pAF*	sTRSV1	GGGAAGGUUCUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUAGAACCUGUUAUAGAACCUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAGGUUCU
pAF_sTRSV1_693,9-697	*pAF*	sTRSV1	GGGAAGGUUCUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUAGAACCUGUAUGCUCAUUCAACCAAUUGAGAUGAAACACUAUUUAUCUAUCACCAAGAGGAUAACGCUUACUCGUAGUGAGGCAGGGGCCCAACGCGUACGCACACUGUUAUAGAACCUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAGGUUCU
pAF_sTRSV1_76,14-875	*pAF*	sTRSV1	GGGAGUAACAGAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUCUGUUAAGGAGGCCUUUCAUCCAACAUACGAGGGUAUUAAUAAAUGUGUGCGCUGUUGGAUUUGAGAGAACGAACCCCGGUCGAAGGCAGCCCAUGGCUAAGCGAAUCUGUUACUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAGUAACAG
pAF_sTRSV1_82,1	*pAF*	sTRSV1	GGGUGAAUGCCCCAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUGGGGCAUCUGUUAUGGGGCAUUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAAAUGCCCC
pAF_sTRSV1_90,12	*pAF*	sTRSV1	GGGUGACUCUCGUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUACGAGAGCUGUUAUACGAGAGUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAACUCUCGU

pAF_sTRSV1_90,12-157	*pAF*	sTRSV1	GGGUGACUCUCGUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUACGAGAGGUAAUAGACUAGUCGUUCUUCAUCGUAUCAAGAAUUAACGUAAAAGUAUCGAUACAUUGAGGCUAAUUCUUGUAACGAUGACUACCGAAGAGGCGGUUAGCUGUUAUACGAGAGUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAACUCUCGU
pAF_sTRSV1_90,12-38	*pAF*	sTRSV1	GGGUGACUCUCGUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUACGAGAGCGUAAGUAACUUAAUCUGUUAUACGAGAGUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAACUCUCGU
pAF_sTRSV1_90,12-54	*pAF*	sTRSV1	GGGUGACUCUCGUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUACGAGAGGUAACAGGCUCCCGCCGGAUAGAAAGAGUACCGGACGAUAAAAACCGGCCGGUCGGAUUUAAACGUCAUGUACUGUUUCUAUUCCGCGGGGUCCAGAAAGCUGUUAUACGAGAGUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAACUCUCGU
pAF_sTRSV1_90,12-90	*pAF*	sTRSV1	GGGUGACUCUCGUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUACGAGAGGACAUGAAUCACGUUCUGUUAUACGAGAGUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAACUCUCGU
Theo_S_man_1038,3	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUCGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC
Theo_S_man_1038,3-108	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUUAGAAGCCAGGUAGGCGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC
Theo_S_man_1038,3-155	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUAGGGCGUAUCGCCUUCGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC
Theo_S_man_1038,3-160	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUUAAAUGCCGCAUGACCGAUGGGUCAUUCGGCGUUCGCGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC

Theo_S_man_1038,3-17	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUAAUACACCGCGUGAAAUAUAUGUCACACAGCGUACUCGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC
Theo_S_man_1038,3-22	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUUAUUACACUAAACAAUUCUUUCUGCUAUUACGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC
Theo_S_man_1038,3-56	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUAACCAAAAUUGGUUUCGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC
Theo_S_man_1038,3-56(2)	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUUGAAGGUAAAUUGUACGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC
Theo_S_man_1038,3-76	Theophylline	*S. Mansoni*	GGGACACAUCGAUACCAGCCGAAAGGCCCUUGGCAGCGAUGUUAGCGCUAGACCGAAUUAAUAUACACCGUGCGAUGUGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACACAUC
Theo_S_man_1158,2	Theophylline	*S. Mansoni*	GGGAAUCUACGAUACCAGCCGAAAGGCCCUUGGCAGCGUAGACGUAGAUUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAAAUCUAC
Theo_S_man_1158,2-72	Theophylline	*S. Mansoni*	GGGAAUCUACGAUACCAGCCGAAAGGCCCUUGGCAGCGUAGAAGCUUAUGACGGAGGCGUAGAUUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAAAUCUAC
Theo_S_man_365,2	Theophylline	*S. Mansoni*	GGGAAGACACGAUACCAGCCGAAAGGCCCUUGGCAGCGUGUCGUGUCUUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAAAGACAC
Theo_S_man_365,2-43	Theophylline	*S. Mansoni*	GGGAAGACACGAUACCAGCCGAAAGGCCCUUGGCAGCGUGUCGAGGUAUAGUAUGACGUGUCUUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAAAGACAC
Theo_S_man_519	Theophylline	*S. Mansoni*	GGGAUGACCCGAUACCAGCCGAAAGGCCCUUGGCAGCGGGUUCGGGUCAUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAAUGACCC

Theo_S_man_821,7	Theophylline	*S. Mansoni*	GGGUCCAGGACCAGUAGAUACCAGCCGAAAGGCCCUUGGCAGCUACUGGUCCUCUACUGGUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAACCAGUA
Theo_S_man_365,2-132	Theophylline	*S. Mansoni*	GGGAAGACACGAUACCAGCCGAAAGGCCCUUGGCAGCGUGUCAAUGGGCUGUUAAACUUUCAUAAUAUUUGCGUGUCUUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAAAGACAC
Theo_S_man_365,2-141	Theophylline	*S. Mansoni*	GGGAAGACACGAUACCAGCCGAAAGGCCCUUGGCAGCGUGUCGUAUUGUGGGUUAUGGUUGCUUUAAUUGUUGUGUCUUCCUGGAUUCCACUGCUUCGGCAGGUACAUCCAGCUGAUGAGUCCCAAAUAGGACGAAAAGACAC
Theo_PLMVd_29-14	Theophylline	PLMVd	GGGACAGAAAACCCUAGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCUAGGGUUUGAAUAACAGUAUUUCAUCUAGGGUUUUCUGUGCUAAGCACACUGACGAGUCUCUGAGAUGAGACGAAAAAACCCUA
Theo_PLMVd_29-14_clamp	Theophylline	PLMVd	GGGACAGAAAACCCUAGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCUAGGGUUUGAAUAACAGUAUUUCAUCUAGGGUUUUCUGUGCUAAGCACACUGACGAGUCUCUGAGAUGAGACGAAAAAACCCUAAGGGUU
Theo_PLMVd_1-19	Theophylline	PLMVd	GGGACAGAAUACCUUUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCAAAGGUAUUAAUGCGACGCACUGCAAAGGUAUUCUGUGCUAAGCACACUGACGAGUCUCUGAGAUGAGACGAAAAUACCUUU
Theo_PLMVd_1-19_clamp	Theophylline	PLMVd	GGGACAGAAUACCUUUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCAAAGGUAUUAAUGCGACGCACUGCAAAGGUAUUCUGUGCUAAGCACACUGACGAGUCUCUGAGAUGAGACGAAAAUACCUUUGGUAUU
Theo_PLMVd--29-18	Theophylline	PLMVd	GGGACAGAAAACCCUAGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCUAGGGUUUGAAAAUAAUAUCAUCAUCUAGGGUUUUCUGUGCUAAGCACACUGACGAGUCUCUGAGAUGAGACGAAAAAACCCUA
Theo_PLMVd--29-25	Theophylline	PLMVd	GGGACAGAAAACCCUAGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCUAGGGUUUUUUCGAUAAGCGAACAUCUAGGGUUUUCUGUGCUAAGCACACUGACGAGUCUCUGAGAUGAGACGAAAAAACCCUA
Theo_PLMVd--1-14	Theophylline	PLMVd	GGGACAGAAUACCUUUGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCAAAGGUAUCCGGAAAAUUUGGGGCAAAGGUAUUCUGUGCUAAGCACACUGACGAGUCUCUGAGAUGAGACGAAAAUACCUUU

Theo_PLMVd--10-19	Theophylline	PLMVd	GGGACAGACUUCCUUAGAUACCAGCAUCGUCUUGAUGCCCUUGGCAGCUAAGGAAGAAACUAAAUUUAUCUAAGGAAGUCUGUGCUAAGCACACUGACGAGUCUCUGAGAUGAGACGAAACUUCCUUA
pDSY9A_IVT	*pAF*	sTRSV1	GGGACGACGACAGGCACCCGAACUCCGCGUCCCAGGGGUGACUCUCGUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUACGAGAGCUGUUAUACGAGAGUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAACUCUCGUCACUUGCGAAAGAGGAGAAAUACUAGAUGAGCAAAGGAGAAGAACUUUU
pDSY9B_IVT	*pAF*	sTRSV1	GGGACGACGACAGGCACCCGAACUCCGCGUCCCAGGGGUGACUCUCGUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUACGAGAGCGUAAGUAACUUAAUCUGUUAUACGAGAGUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAACUCUCGUCACUUGCGAAAGAGGAGAAAUACUAGAUGAGCAAAGGAGAAGAACUUUU
pDSY9C_IVT	*pAF*	sTRSV1	GGGACGACGACAGGCACCCGAACUCCGCGUCCCAGGGGUGACUCUCGUAUAACAGGUGAUCAGUAGCCUGUACAGCUUCGGCUGCGUCCUACUAUACGGACUAUACGAGAGGUAAUAGACUAGUCGUUCUUCAUCGUAUCAAGAAUUAACGUAAAAGUAUCGAUACAUUGAGGCUAAUUCUUGUAACGAUGACUACCGAAGAGGCGGUUAGCUGUUAUACGAGAGUCACCGGUAACCGGUCUGAUGAGUCCGUGAGGACGAAACUCUCGUCACUUGCGAAAGAGGAGAAAUACUAGAUGAGCAAAGGAGAAGAACUUUU

Table S1. Sequences of K-As tested in this work.

Chapter 2: Antisense-RBS Riboswitches Regulate Gene Expression in *E. coli*

Introduction

Given the shortcomings of aptazymes as genetic controllers in bacteria, we decided to apply our molecular architecture and computational screening workflow to a new output domain better suited to regulating gene expression levels in *E. coli*. One of the most common mechanisms employed by natural riboswitches to control gene expression levels is to regulate the accessibility of the ribosome binding site (RBS) to incoming ribosomes[63]. The 16S ribosomal subunit, which is composed of RNA, binds to the RBS through the base-pairing of complementary sequence. As a result, the occlusion of the RBS by competing base-pairs within its mRNA molecule is known to have a dramatic impact on how effectively the RBS and ribosome are able to associate[64]. In turn, this regulates the rate at which the associated mRNA molecule is translated into protein, and the resulting steady-state protein concentration. This mechanism results in translation initiation rates that vary by several orders of magnitude, making them attractive for high performance biosensors.

While RBSs are an attractive output domain for synthetic aptamer regulation, they are not readily compatible with our described molecular architecture for the engineering of kinetic RNA biosensors. While the molecular architecture is designed to regulate the folding of structured RNA output domains that possess a closing stem, RBSs are nearly completely unstructured in their most active form. This represents a significant incompatibility, as the previously characterized objective functions and quantitative design metrics identified in the K-A system would no longer apply. To address this issue, we converted the RBS sequence into an antisense-RBS sequence by appending a 5' extension to the RBS that is the reverse-complement of the wild-type sequence (Figure 1A). This creates an RNA domain with a closing stem, suitable for use in our molecular architecture, while also inverting the signal of the riboswitch, such that translation levels are maximized when the aptamer doesn't bind its target and minimized when it does bind its target (Figure 1B).

Due to a complex interplay of thermodynamic structure ensembles, and refolding kinetics, one particular struggle with the design of RNA biosensors acting under thermodynamic control is the difficulty of predicting the concentration at which one would expect the switch to respond. For this reason, it is possible that many functional RNA-based biosensors have been deemed non-functional due to a

mismatch between the biosensor's actual EC_{50}, and the concentration that the researcher is able to assay the biosensor's performance under, whether due to solubility or other mechanistic incompatibility. This lack of predictability of response is exacerbated when the candidate biosensor is being expressed within a cell where cellular uptake of a molecule added extracellularly, and cellular metabolism of said molecule, create additional confounding factors that make the validation of novel biosensor design strategies increasingly difficult. For this reason, the theophylline aptamer has been a popular choice for the validation of new genetically-encoded biosensors. Although there is uncertainty regarding the quantitative relationship between extracellular and intracellular theophylline concentrations, it has been validated that theophylline can enter the cell and is not readily degraded by bacterial metabolism[65]. However, despite these advantages, the biosensors that have been reported in the literature routinely have EC_{50} values only slightly below the concentration where theophylline becomes toxic. In order to identify high performance biosensors, attaining high sensitivity to theophylline is therefore a high priority.

To validate this new application of our molecular architecture, and to benchmark our biosensors against those reported in the literature, we decided to first engineer a theophylline-responsive AS-RBS riboswitch. While our kinetic biosensor molecular architecture allows us to make *a priori* predictions about biosensor sensitivity prior to being experimentally characterized, it gives us no additional insight into the concentration of theophylline present intracellularly. Therefore, in order to give our biosensors the best chance to sense the potentially very low concentrations of theophylline within the cell, we decided to take a lesson from nature and implement a transcriptional pause site within the Timer domain of our candidate biosensor. The TPP riboswitch family possesses a range of sensitivities to TPP, with EC_{50} values ranging by over an order of magnitude. As in our AS-RBS riboswitches, the TPP riboswitch from the *ThiC* gene in *E. coli* regulates RBS accessibility such that translation levels are maximized when the aptamer doesn't bind its target and minimized when it does bind its target. Interestingly, between the aptamer and RBS, in the region analogous to the Timer domain in our kinetic biosensors, the riboswitch contains a hairpin that has been validated as a transcriptional pause site that causes the RNA polymerase to stall, with a half-life of nearly a minute, before continuing to transcribe the rest of the mRNA[66–69]. As would be expected from a kinetic biosensor, this transcriptional pause activity has been demonstrated to increase the sensitivity of the biosensor to its target molecule. We reasoned that in the correct folding context, this *ThiC*

transcriptional pause site could be incorporated into the Timer domain of our AS-RBS riboswitches to significantly increase their sensitivity to theophylline and increase their overall performance (Figure 1C).

Here we show that our kinetic biosensor molecular architecture can also be applied to the design of translation-controlling AS-RBS riboswitch constructs that function within *E. coli*. Furthermore, we demonstrate that a natural transcriptional pause site from *E. coli* can be incorporated into the Timer domain of an AS-RBS riboswitch, resulting in unprecedented sensitivity to theophylline. We then demonstrate that the biosensor's high sensitivity and ligand activation ratio depend on the specific sequence of the Timer domain. Finally, we show that by screening synonymous codon variants of the 5' end of our output gene, we are able to increase the expression levels without impacting the ligand activation ratio of our biosensor.

Methods

AS-RBS switch design

First the conventional bglbrick RBS was combined with its reverse-complement appended 5' in order to create a hairpin expected to dramatically reduce translation initiation rate. A mutational operator was applied so that the AS-RBS itself was not a perfect hairpin in order to increase in silico pool diversity, and to prevent the 5' end of the switch from being identical to the RBS sequence itself, resulting in another site for translation initiation. The RBS calculator was used to ensure that the predicted translation initiation rate for the AS-RBS sequence was much lower than for the RBS without the antisense sequence appended[64]. Screening for B1 barrier heights <= 7.8 kcal/mol was implemented, corresponding to the decreased rate of *E. coli* RNA polymerase nucleotide addition, relative to T7. Screening for B2* barrier heights <= 2.9 kcal/mol was implemented, corresponding to the same increase in barrier height as in B1. Toe-target distances were screened for values <= 8.5 arbitrary units, as for K-As previously. Screening for pathway convergence >= 0.7 was implemented as it was demonstrated to improve performance in kinetically-controlled ribosensor design (Figure S1).

In vivo Timer pool screening

In order to generate a pool of plasmids containing diverse Timer domains, we constructed a destination vector with two outward-facing SapI restriction sites placed into the computationally-designed switch candidate in the location where the Timer domain would be. Golden Gate plasmid assembly was used to insert a pool of short, double-stranded DNA fragments, generated through primer extension PCR. The Timer domain pool contained the *ThiC* transcriptional pause site flanked by variable positions.

The pools of plasmids were transformed into *E. coli* strain DH10B cells and plated onto plates containing the relevant antibiotic and grown at 37C for 16-24 hours. At that point, the brightest green colonies were picked and grown up in MOPS EZ-Rich defined media containing the appropriate antibiotics for 24 hours. The liquid cultures were then diluted 1:1000 into 400 uL of fresh media containing 0 mM or 1 mM theophylline. After an additional 16-24 hours, 150 uL of culture was read in a 96-well plate format in a Synergy HTX plate reader (BioTek) with gain 35.

Titration of modified pause-containing Timers

At least 2 biological replicates were grown up in 400 ul of MOPS EZ-Rich defined media containing the appropriate antibiotics for 24 hours. The liquid cultures were then diluted 1:1000 into 400 uL of fresh media containing a 2-fold dilution series of Theophylline starting at 2.5 mM. Cultures were grown for an additional 24 hours, and then 150 uL of culture was read in a 96-well plate format in a Synergy HTX plate reader (BioTek) with gain 35.

Synonymous codon pool screening

A destination vector containing outward-facing SapI restriction enzyme sites was assembled using standard molecular cloning techniques. A pool of short fragments, flanked by internal-facing SapI restriction enzyme sites was assembled, containing DNA containing the first 11 codons of the sfGFP gene, with positions varied such that synonymous codons, coding for the same 11 N-terminal amino acids of sfGFP could be identified (Figure SX). Partial doping was used to keep the pool to as many synonymous codon replacements as possible. The doped oligo to perform the assembly was ordered from IDT and assembled into a double-stranded fragment using primer extension. The assembled pool was transformed onto LB-Agar plates containing no theophylline. 48 of the brightest green colonies were

picked, and grown in 400 μL of MOPS EZ-Rich defined media containing the appropriate antibiotics for 24 hours. The liquid cultures were then diluted 1:1000 into 400 μL of fresh media containing 0 mM or 1 mM theophylline. After an additional 16-24 hours, 150 μL of culture was read in a 96-well plate format in a Synergy HTX plate reader (BioTek) with gain 35.

Results and Discussion

Identification and characterization of a theophylline-responsive AS-RBS riboswitch

Utilizing our computational approach outlined in Chapter 1, we generated a candidate AS-RBS riboswitch engineered to respond to theophylline. The primary difference was that the barrier heights used for screening were increased corresponding to the slower rate of elongation for *E. coli* RNA polymerase compared to T7 RNA polymerase used in Chapter 1. In order to increase the sensitivity of candidate AS-RBS riboswitch constructs within a bacterial cell, we decided to incorporate a natural transcriptional pause site from *E. coli* into the Timer domain of the AS-RBS riboswitches. Bacterial transcriptional pause sites resemble rho-independent transcriptional terminators, wherein a hairpin is followed by a 3' poly-T stretch. However, unlike transcriptional terminators, this poly-T stretch is not continuous, and is interrupted by other bases. As transcriptional terminators are known to have very rapid and specific co-transcriptional folding trajectories that enable them to function, we reasoned that transcriptional pause sites would as well, and would therefore only function under very specific folding contexts. Without knowing what these precise folding contexts should look like, we opted to screen for functional switches *in vivo*, as opposed to *in silico*. In the future, we hope that folding analysis of functional switch variants will allow us to determine the folding rules to predict pause site function purely computationally.

To perform the *in vivo* screening, we used scarless Golden Gate plasmid assembly to generate a pool of plasmids containing variable sequence within the Timer domain adjacent to the transcriptional pause site from the *ThiC* gene in *E. coli*. Individual colonies were picked into liquid media and grown in the presence and absence of theophylline. Colonies that showed the largest fold change in normalized GFP fluorescence were then isolated and sequenced. Interestingly at least one variant even demonstrated the ability to increase fluorescence in response to theophylline, counter to the intended

mode of action. The isolated sequence displaying the largest fold-change in response to theophylline was named Theo_48. To characterize the sensitivity of Theo_48, we characterized its response at several different theophylline concentrations. Strikingly, it responded at very low concentrations, displaying an EC_{50} of 47μM (Figure 2A). Notably, this EC_{50} is four times lower than has been observed from theophylline-responsive riboswitches in bacteria previously (Figure 2B). In fact, it is more than 10-fold lower than the average measured EC_{50} from similar biosensors, and more than 10-fold lower than the only other engineered riboswitch known to exhibit kinetic behavior[18,70–75]. This suggests that the pause site is indeed having the desired impact, resulting in extended co-transcriptional ligand binding windows, and enabling unprecedented sensitivity of the engineered biosensors.

To validate the role that the transcriptional pause site played in achieving the unprecedented sensitivity and high activation ratio of the Theo_48 switch, we created a series of Timer domain variants designed to reduce, or eliminate the duration of the transcriptional pause (Figure 3). We either introduced point mutations designed to make the *ThiC* pause site look less like a transcriptional terminator, by mutating the poly-T stretch 3' of the hairpin within the pause site, or deleted sections of the Timer domain, or pause site. Introducing 2 point mutations within the poly-T stretch had essentially no impact on switch performance, but introducing 6 point mutations began to shift the EC_{50} to higher values, while still maintaining the ligand activation ratio of the switch. This seems to suggest that the pause duration may be decreasing with increasing number of point mutations, resulting in a shortened ligand binding window, without otherwise impacting switch function. In the case of the deletions, all of the modified Timer domains resulted in reduced ligand activation ratios and significantly increased EC_{50} values. This is largely expected in the cases of the Timer_Only and PolyT_Only constructs, where some, or all, of the pause site has been deleted (Figure S1). It is very interesting, however that the ThiC_Only construct, in which the entire *ThiC* pause site is retained but only the variable sequence upstream of the pause site is removed, shows the lowest sensitivity of all the tested constructs. This suggests that the pause site alone is not sufficient for transcriptional pausing; To cause a significant transcriptional pause, the *ThiC* pause site must be surrounded by an appropriate sequence construct that enables it to fold correctly.

Increasing AS_RBS signal though screening of synonymous N-terminal codons

While the excellent sensitivity and ligand activation ratio of the Theo_48 construct represented a significant accomplishment for applying our kinetic biosensor design pipeline to another output domain, the maximum signal of the Theo_48 construct was substantially lower than the maximum signal observed from a positive control using the same RBS sequence and promoter. In fact, another candidate AS-RBS riboswitch designed to respond to pAF demonstrated a much higher maximum expression level, suggesting that there is no fundamental limitation on gene expression imposed by the architecture. Computationally predicted structures of the theophylline- and pAF responsive biosensor candidates suggested that the RBS in the theophylline construct maintained a larger degree of residual structure, even when the target molecule is absent, and that some of this residual structure was with the 5' end of the output sfGFP gene. In order to increase the maximum expression level of the Theo_48 construct, we wished to decrease the degree of structure between the RBS and the 5' end of sfGFP. Inspired by the observation that rare codons are enriched at the 5' end of bacterial genes, likely to reduce unintended structure with the adjacent RBS, we sought to use a similar strategy[76]. Without changing any part of the Theo_48 biosensor itself, we created a pool of plasmids containing synonymous codons for the first 11 amino acid positions of sfGFP (Figure 4A). Despite possessing different RNA structure, due to their divergent sequence, the pool variants should contain the same amino acid sequence, resulting in no modification to sfGFP when translated. We then picked colonies exhibiting brighter green color when grown on LB-Agar plates. All the selected colonies indeed displayed increased GFP levels when grown in liquid culture (Figure 4B). Interestingly, while each of the variants displayed different levels of fluorescence, the activation ratio in response to theophylline remained nearly identical in all of the isolates (Figure S2). This serves as additional evidence that the designed switches are behaving in a co-transcriptional, kinetic, manner where the behavior and identity of the sequence 3' of the switch don't impact the switch state once the co-transcriptional ligand binding window has closed.

Figures

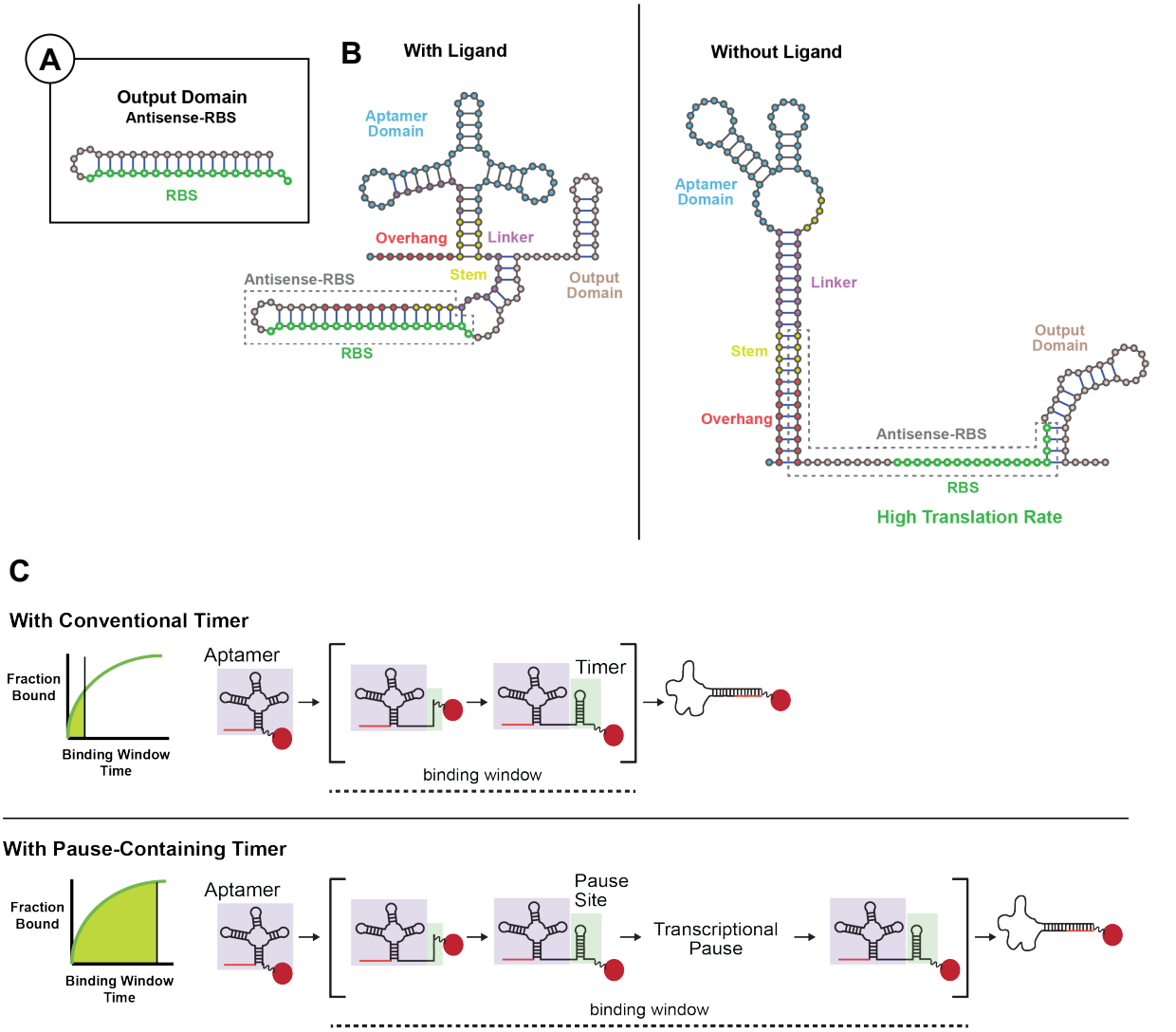


Figure 1. Schematic of an AS-RBS riboswitch. A) In order to make an RBS suitable for implementation within our kinetic biosensor molecular architecture, the RBS is combined with a 5' antisense sequence. B) When incorporated into a candidate AS_RBS riboswitch, it results in high translation in the absence of the target molecule, and low translation when the target molecule is bound. C) Incorporating a transcriptional pause into the Timer domain increases the duration of the binding window dramatically, resulting in increased sensitivity to the target molecule.

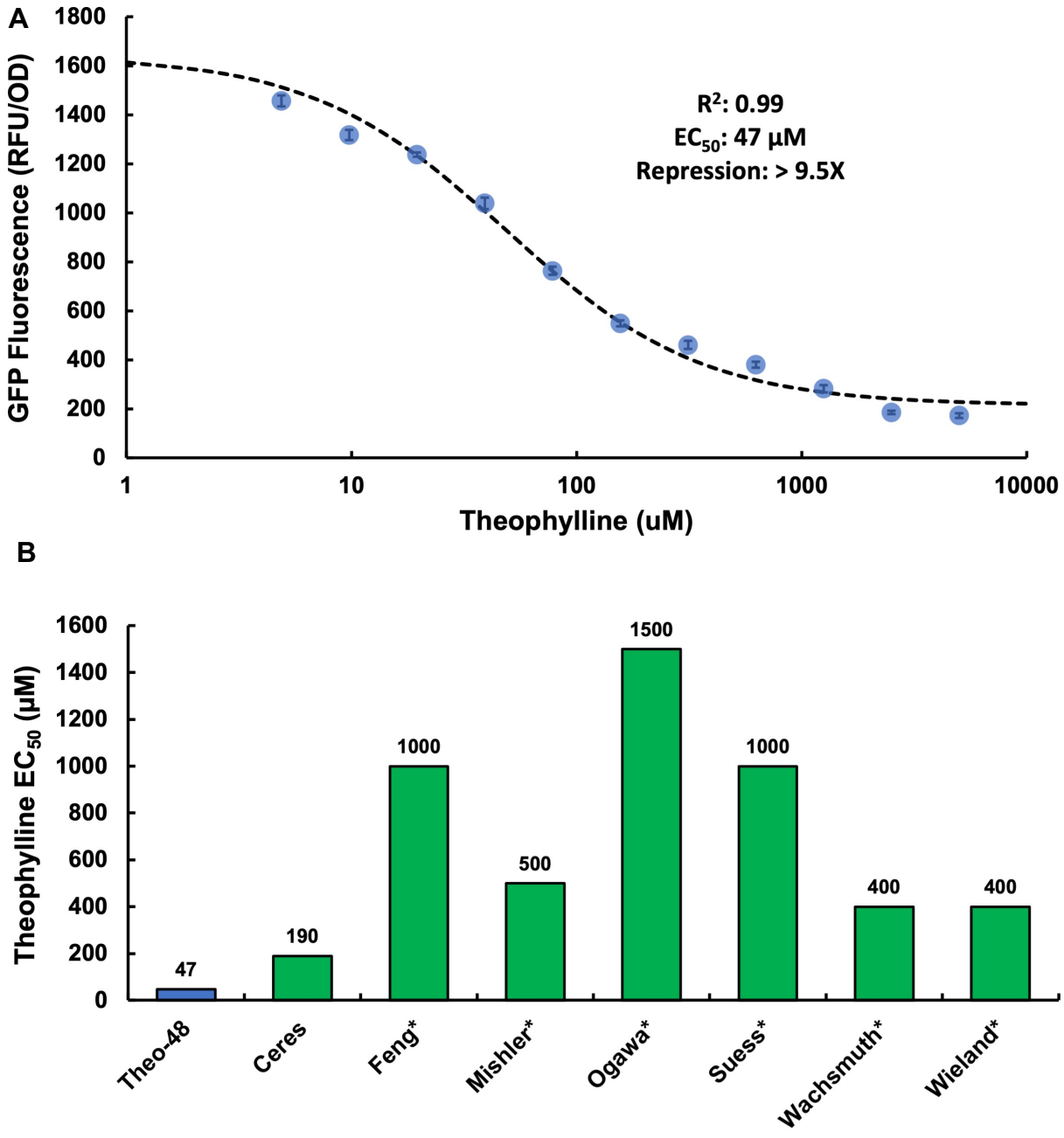


Figure 2. Characterization of a high-sensitivity AS-RBS riboswitch. A) The Theo-48 AS-RBS riboswitch possesses a significant fold-change in response, which means that at least 90% of all transcribed biosensors are able to bind to the target molecule. In addition, it possesses an extremely low EC_{50}. B) The EC_{50} of Theo-48 is substantially lower than all other cis-acting theophylline-responsive riboswitches in bacteria. Asterisk indicates that the EC_{50} was not presented, and was estimated from presented titration data.

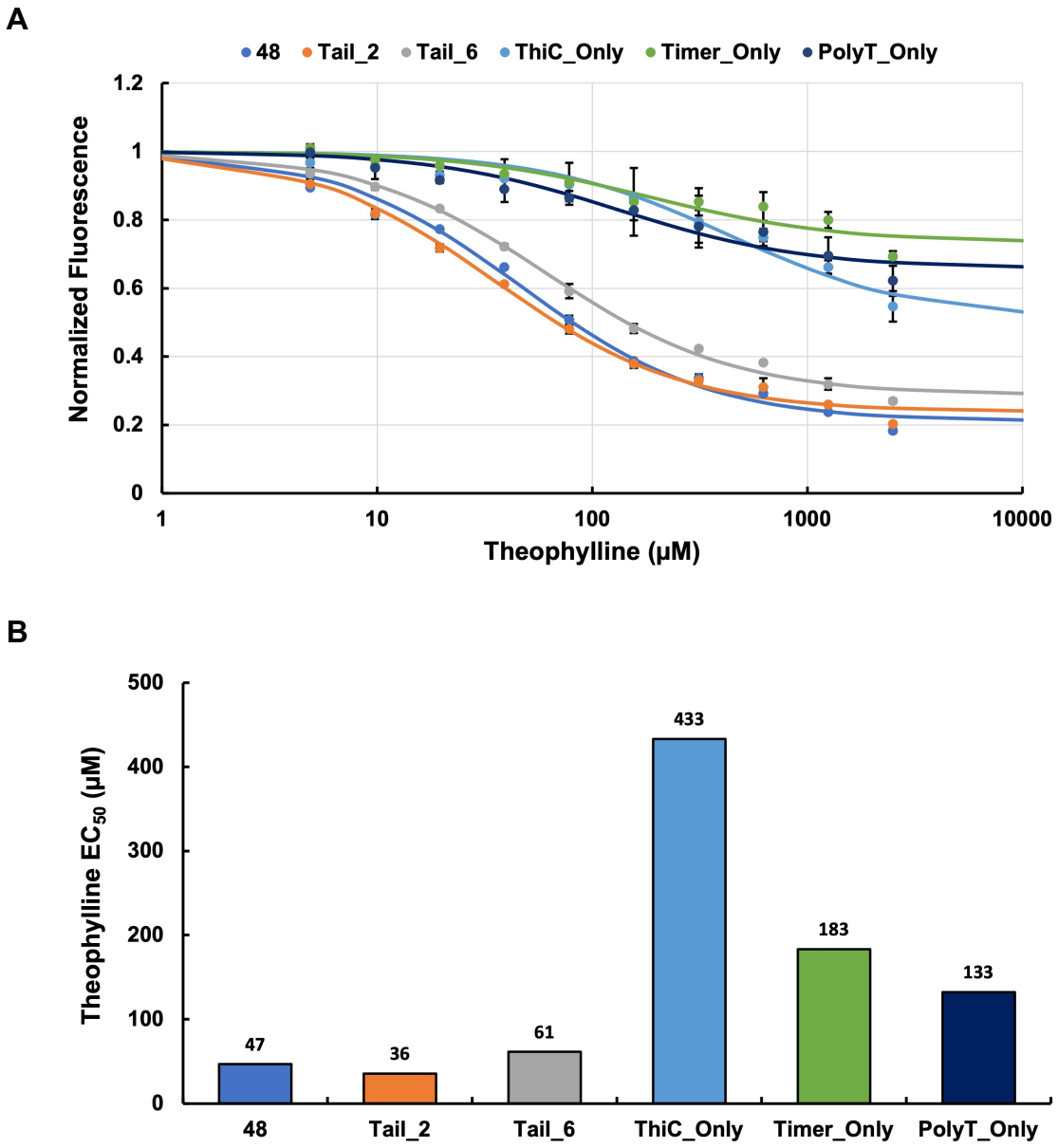


Figure 3. Pause sequence and context are critical for pause site activity. A) Modifying a pause-containing Timer domain results in reductions in activation ratio and sensitivity to the target molecule. B) EC_{50} values increase for all mutations or deletions. Interestingly this is especially true for ThiC_Only, which contains the pause site, but none of the flanking sequence. This suggests that folding context is critical for effective transcriptional pausing.

A

```
M  S  K  G  E  E  L  F  T  G  V   -Amino acid sequence
ATGAGCAAAGGAGAAGAACTTTTCACTGGAGTT -DNA sequence
     T  G  G  G  GT A  T  A  G  A -Synonymous
           C        G     C  C  G
           T        C     G  T  C

ATGAGYAARGGNGARGARYTNTTYACNGGNGTN -Ordered DNA pool
```

Total diversity of pool: 32,768
Diversity of pool with correct AA sequence: 24,576
Fraction of pool elements with correct AA sequence: 0.75

B

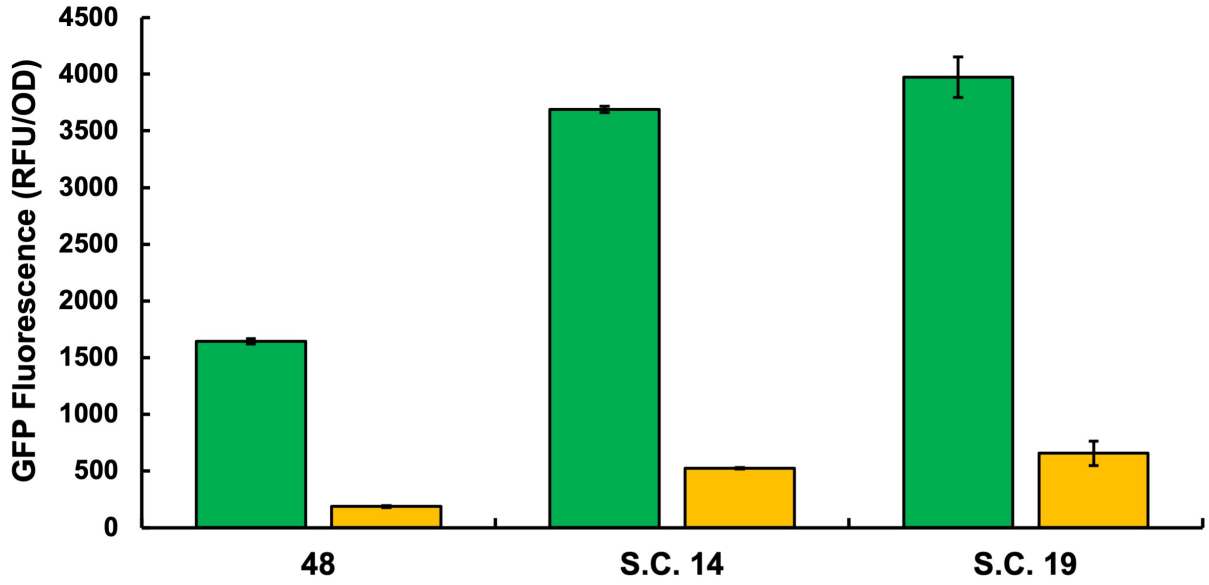


Figure 4. Synonymous codon variants increase AS-RBS riboswitch expression levels. A) Varying specific positions allows the production of a pool where nearly all sequences code for the same amino acid sequence, despite variable RNA sequence. B) Selected synonymous codon variant increase gene expression levels in both the presence and absence of theophylline.

Supplementary Materials

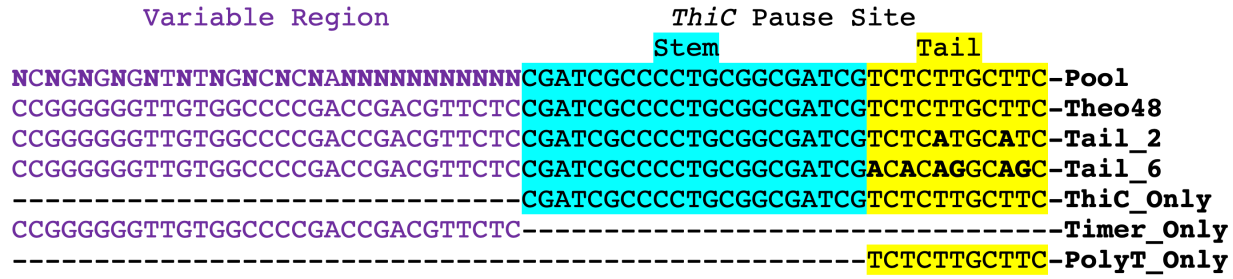


Figure S1. Sequences of modified Timer variants of Theo-48. DNA sequences shown represent the entire Timer domain from the tested constructs. “Stem” refers to the palindromic hairpin within the *ThiC* pause site, while “Tail” refers to the single-stranded poly-T-like sequence within the pause site.

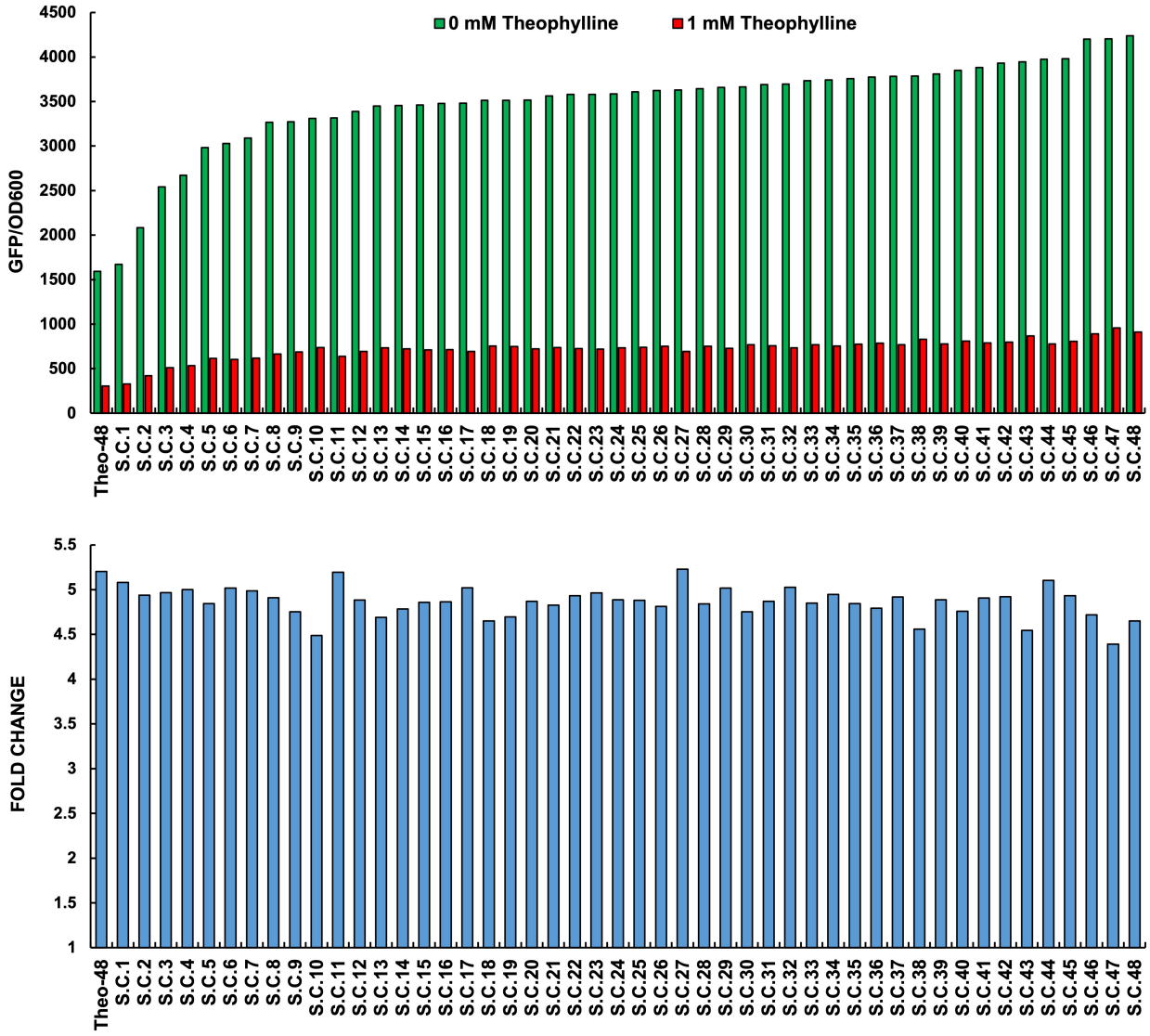


Figure S2. Synonymous codon variants increase the expression levels without changing activation fold change. A) Synonymous codon variants increase the expression levels in both the presence and absence of theophylline. B) Despite differences in expression levels, the fold change between variants and the wild-type sequence are nearly identical.

Chapter 3: Challenges and opportunities with CRISPR activation in bacteria for data-driven metabolic engineering

Jason Fontana[*,1], David Sparkman-Yager[*,1], Jesse G. Zalatan[+,2], and James M. Carothers[+,3]

1: Molecular Engineering & Sciences Institute and Center for Synthetic Biology
2: Department of Chemistry
3: Department of Chemical Engineering
University of Washington
Seattle, WA 98195
United States

*: these authors contributed equally
+: Corresponding authors
zalatan@uw.edu
206-543-1670
jcaroth@uw.edu
206-221-4902
Keywords: CRISPRa, CRISPR-Cas activation, CRISPRa design rules, gRNA design

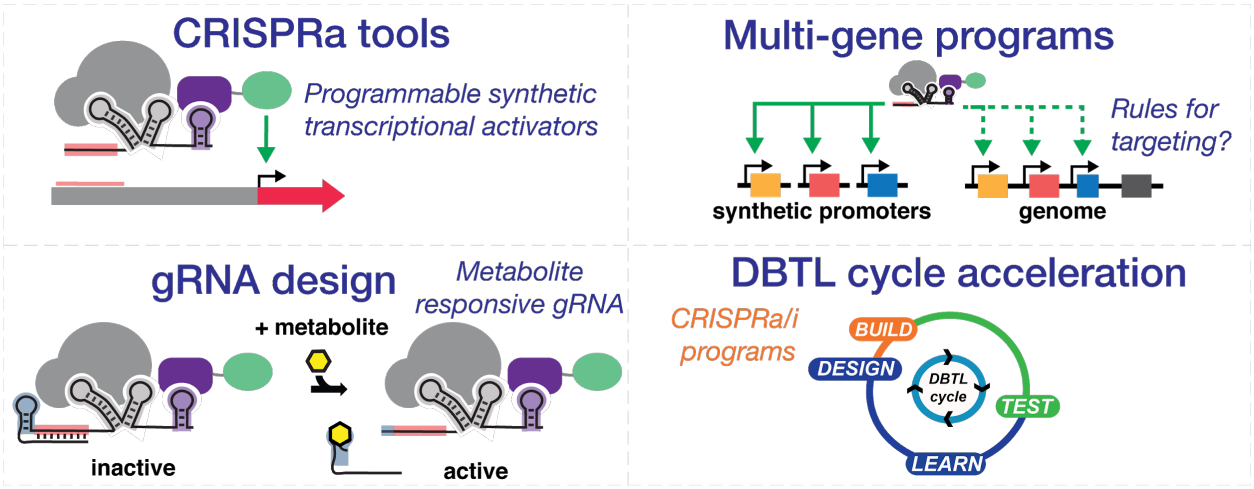
Highlights

- CRISPR activation is an emerging tool for engineering bacterial metabolism.
- The promoter design rules for effective CRISPR activation in bacteria are complex.
- Guide RNAs can be engineered as flexible platforms to program CRISPR activation.
- CRISPR activation may accelerate strain optimization Design-Build-Test-Learn cycles.

Abstract

Creating CRISPR gene activation (CRISPRa) technologies in industrially-promising bacteria could be transformative for accelerating data-driven metabolic engineering and strain design. CRISPRa has been widely used in eukaryotes, but applications in bacterial systems have remained limited. Recent work shows that multiple features of bacterial promoters impose stringent requirements on CRISPRa-mediated gene activation. However, by systematically defining rules for effective bacterial CRISPRa sites and developing new approaches for encoding complex functions in engineered guide RNAs, there are now clear routes to generalize synthetic gene regulation in bacteria. When combined with multi-omics data collection and machine learning, the full development of bacterial CRISPRa will dramatically improve the ability to rapidly engineer bacteria for bioproduction through accelerated design-build-test-learn cycles.

Graphical abstract



Introduction

Bacterial metabolism is made up of complex gene networks that can be engineered to produce medically- and industrially-important chemicals. The complexity of these networks means that sophisticated organism engineering efforts are typically needed to optimize the production of high-value compounds [77,78]. In principle, synthetic multi-gene transcriptional programs could be constructed to engineer metabolic networks for efficient industrial chemical production [79,80]. In practice, however, an incomplete understanding of metabolic networks, combined with a limited ability to predictably control the expression of multiple genes, makes achieving this goal difficult [81]. A recurring challenge when engineering strains for chemical production remains the difficulty of predicting the optimal expression level of pathway and non-pathway genes that will result in optimal yields. To overcome this challenge, there is a need for new technologies for rapidly implementing and analyzing combinatorial multi-gene expression programs. These technologies could be combined with advanced capabilities for multi-omics data collection and machine learning to enable accelerated design-build-test-learn cycles (DBTL) [77,82,83]. CRISPR-Cas tools have changed every aspect of microbial engineering, including the speed with which genomes can be edited and the ability to target specific genes for activation and repression [84,85]. CRISPR-Cas tools for programming gene expression use the catalytically-inactive Cas9 protein (dCas9) along with guide RNAs that recognize DNA targets through predictable Watson-Crick base pairing [85]. A major strength of CRISPR-based synthetic gene regulation is that new combinatorial multi-gene expression programs that include simultaneous transcriptional activation and repression could be rapidly implemented.

There are well-established approaches for repressing genes (CRISPRi) by targeting dCas9 to physically block RNA polymerase and inhibit transcription [86]. In eukaryotic cells, robust transcriptional activation can be applied using CRISPR-Cas to direct activation domains upstream of target genes (CRISPRa) [84,85]. However, the development of CRISPRa in bacteria has been hindered by the lack of effective activation domains. The recent discovery that at least four different bacterial activators can be linked to programmable CRISPR-Cas DNA binding domains has promised to significantly change the outlook for CRISPRa in bacteria [87–90]. Further, new efforts to uncover practical rules for activating transcription with bacterial CRISPRa may make it possible to build complex multi-gene programs that

regulate the expression of both heterologous and endogenous genes. By building on recent efforts, the further development of engineered guide RNAs as flexible platforms for programming CRISPRa may create new capabilities for predictable and metabolite-responsive synthetic gene regulation [91,92]. This review focuses on new advancements in bacterial CRISPRa technologies that promise to significantly accelerate strain optimization through data-driven metabolic engineering.

CRISPRa for regulating bacterial transcription

In bacteria, the implementation of complex multi-gene CRISPR-Cas expression programs has been limited by a lack of effective gene activators. To address this problem, new synthetic transcriptional activators have been developed in *E. coli* that link activation domains to programmable CRISPR-Cas DNA binding domains [87–90]. The resulting CRISPRa tools have proven capable of driving heterologous gene expression at levels suitable for metabolic engineering. Some successes have also been achieved in activating the expression of endogenous genes from genomic loci. The further development of these capabilities may permit the optimization of metabolic production through the construction of multi-gene programs simultaneously targeting heterologous and endogenous genes.

There are two mechanistic approaches that have been employed to link activation domains to CRISPR-Cas DNA binding domains. Activation domains can be (i) directly fused to dCas9, or (ii) recruited to dCas9 using modified sgRNAs (scaffold RNAs or scRNAs) that bind to RNA binding protein-activation domain fusions [93–95]. Using the first approach, CRISPRa has been achieved in *E. coli* by fusing the ω-subunit of RNA polymerase (*rpoZ*) to dCas9 to obtain 23-fold increases in reporter gene expression from synthetic promoters [87]. These fusions have been applied both in *E. coli* and non-model bacteria. In *E. coli*, dCas9-RpoZ was used activate transcription and identify genes that increase tolerance to the monoterpene pinene, as well as new epistatic interactions between antibiotic resistance genes [96,97]. These tools were successfully ported to *B. subtilis* to obtain 3-fold activation of reporter gene expression and applied to systematically improve production of amylase BLA by 260-fold compared to a commonly used strong promoter [98]. In *L. enzymogenes*, dCas9-RpoZ was used to enhance production of anti-MRSA antibiotics up to 9-fold [99]. In *M. xanthus*, dCas9-RpoZ was able to generate 8-fold increases in the expression of the epothilone production gene cluster, leading to a 6.8-fold improvement in epothilone A

production [100]. Finally, a new portable CRISPRa system where the activation domain AsiA was fused to dCas9 was recently introduced [90]. Using this system, reporter gene expression could be activated by 135-fold in *E. coli*, ~3-fold in *S. enterica,* and ~12-fold in *K. oxytoca.*

The second approach for bacterial CRISPRa relies on modified gRNAs (scRNAs) that recruit an RNA binding protein fused to an activation domain to the CRISPRa complex. One successful strategy uses the RNA binding protein MCP fused to the SoxS activation domain (MCP-SoxS) (Figure 1) [88]. Using MCP-SoxS and a corresponding MS2 scRNA, 50-fold CRISPRa activation was demonstrated and applied to drive ethanol production in *E. coli* from a *Z. mobilis* gene cluster. MCP-SoxS can activate expression from genes that use σ factors σ^{70} [88] and, at lower levels, σ^{38}, σ^{32}, and σ^{24} [101]. While together these σ^{70}-family promoters cover the majority of the *E. coli* genome, σ^{54} promoters, which drive nitrogen starvation genes, could not be activated by MCP-SoxS. Recently, an alternative bacterial CRISPRa system that is effective at σ^{54} promoters was introduced based on the PspFΔHTH::λN22plus activator [89]. Therefore, PspFΔHTH::λN22plus and MCP-SoxS can be used in combination to target a different, non-overlapping set of promoters in *E. coli*. Further, PspFΔHTH::λN22plus was reported to activate two promoters in the nitrogen fixation pathway of *K. oxytoca* by up to 6-fold [89]. In the dCas9-RpoZ and PspFΔHTH::λN22plus systems, obtaining the highest levels of activation requires knocking out the native copy of *rpoZ* [87,102] or *pspF* [89], respectively, to remove the competing, endogenous functions. It is possible, however, to obtain significant activation without using knockout strains. In contrast, the MCP-SoxS and dCas9-AsiA systems do not require any host engineering to achieve their highest levels of activation.

The available CRISPRa tools are uniquely positioned to rapidly implement combinatorial multi-gene expression programs targeting synthetic promoters and identify optimal expression conditions for metabolite production [103]. These tools were recently applied in a proof-of-concept experiment to tune the expression of three genes in the pathway responsible for producing violacein, a pigment with antitumoral properties [89]. Further improving our ability to predictably tune CRISPRa at multiple sites independently could provide a technology for the rapid combinatorial optimization of multi-gene pathways. Dynamically-controlled CRISPRi was recently shown to improve production of salicylic acid in engineered *E. coli* through the conditional knock-down of essential genes [104]. Developing dynamically-controlled CRISPRa

could provide additional avenues to control both the timing and expression levels of multiple genes in engineered metabolic pathways and networks [79].

Promoter design rules improve CRISPRa in bacteria

Recent work has identified multiple features of bacterial promoters that impose stringent requirements on CRISPRa-mediated gene activation [101] (Figure 1). These behaviors suggest an explanation for why CRISPRa and other tools for gene activation in bacteria have lagged far behind comparable tools in eukaryotic systems, where such strict target site requirements are absent. For instance, the activity of CRISPRa using MCP-SoxS is influenced by the strength of the target promoter, the sigma factor regulating the promoter and the sequence composition of the promoter [101]. Most strikingly, when activating synthetic promoters in bacteria, CRISPRa is sensitive to the position and periodicity of the scRNA target site relative to the transcription start site (TSS) [89,101]. Activation can only be performed at precisely defined positions in phase with the transcription start site, which are intervened by regions of lower activity or inactivity [89,101]. These requirements are much more stringent than those for activation in eukaryotic cells [93] and constrain CRISPRa to precisely-positioned PAM sites which may not be found on every gene. Engineered Cas9 variants and alternative Cas proteins have been introduced that expand the range of PAM sequences that can be targeted and increase the density of available PAM sites up to 6 times [101,105–107]. One of the variants, dxCas9(3.7), has been used to demonstrate activation of *E. coli* genes previously inaccessible by dCas9 [89,101]. By combining dxCas9(3.7) and newly defined rules for CRISPRa, 3 out of 7 endogenous *E. coli* genes were successfully activated [101]. However, the field still lacks integrated models for predicting effective CRISPRa target sites for arbitrary genes, and explanations for the failure to activate some genes remain elusive. Genome-wide CRISPRa screens of endogenous promoters could more fully elucidate the requirements for CRISPRa targeting. Once predictive rules for targeting endogenous genes are available, combinatorial multi-gene programs for optimizing bioproduction could be extended to endogenous genes, in addition to synthetic promoters.

gRNAs can be engineered to program CRISPRa responses

gRNA engineering has long been understood to provide routes for tuning CRISPR-Cas functions, and more recently, as a mechanism for encoding dynamic responses to molecular targets. While most of the gRNA design work to date has been performed on guides used for DNA cleavage or CRISPRi, the principles, whether controlling the stability of the guide-Cas9 complex or the entire DNA-guide-Cas9 complex, may be readily applicable to CRISPRa efforts. Guide RNAs are comprised of two main components: the twenty nucleotide spacer sequence, which hybridizes to the target DNA, and the Cas9-binding handle, which drives the formation of the gRNA-Cas9 complex. Structurally, the only difference between a gRNA and a scRNA used for CRISPRa is the presence of an additional 3' RNA hairpin. This hairpin enables the co-localization of the activation domain by binding its cognate RBP tag. While this motif adds additional complexity, it also provides more opportunities for design. Several alternative architectures have also been described that insert the RNA hairpin motif at multiple points within the guide [108,109]. To date, three cognate pairs of RNA binding protein (RBP) and RNA hairpin have been utilized to implement CRISPRa in bacteria (Figure 2) [88,89]. Other pairs have been demonstrated in eukaryotes and may be functional in bacteria as well [94,110]. These orthogonal pairs provide the opportunity to simultaneously implement multiple activators in the same cell.

While the relative simplicity of gRNAs allows the rational specification of target sites, the sequence identity of a chosen site has been demonstrated to have a significant impact on its CRISPR activity. Both the sequence identity and secondary-structure (base-pairing interactions) of gRNA elements are critical for function [111]. There have been several attempts to predict CRISPR activity for novel target sequences, and while these models can be used to increase the probability of selecting a functional guide, they primarily utilize sequence elements, rather than structural information, and have not been applied directly to CRISPRa [112–114]. One key feature that has been demonstrated to influence CRISPR activity is the secondary structure that the guide RNA adopts [115,116]. The degree of secondary structure, whether internal to the spacer or between the spacer and the rest of the guide, has been observed to reduce gRNA effectiveness (Figure 2) [115]. Even the transiently-stable structures the guides adopt during transcription have been demonstrated to impact their activity [115]. As the guides utilized for CRISPRa are actively transcribed from heterologous promoters inside the cell, avoiding transient misfolding may prove

important for achieving predictable activity. Thus, developing tools for screening gRNA co-transcriptional folding pathways may aid in the *a priori* selection of highly functional spacer sequences.

To generate differences in the expression levels of multiple genes, it is necessary to develop a general strategy to fine-tune CRISPRa-mediated gene expression at each promoter. To date, two main strategies have been demonstrated to modulate the CRISPR activity for a given target sequence: spacer truncations and 5' extension. In CRISPRi systems, the level of transcriptional repression applied to target genes has been reduced by truncating the sgRNA target sequence from the 5' end [86,117]. Practically, spacers shorter than 12 nucleotides may increase off-target activity as the first 12 nt, and even more so the first 7 nt known as the 'seed region', have an especially large impact on the activity [86,118]. Alternatively, it has been demonstrated that adding a 5' extension onto the guide, which folds back to occlude the spacer, results in monotonic drops in guide activity with increasing stability of the designed interaction; this correspondence even applies to guides from other Cas proteins with different guide architectures [119]. This demonstration provides evidence that computational predictions of gRNA structure may be sufficient to predict guide function. In order to improve the forward engineering of CRISPRa systems and accelerate DBTL cycles, developing quantitatively-accurate predictions of CRISPRa activity based on scRNA structure will be essential.

Towards nucleic acid-responsive gRNAs for CRISPRa

In order to implement complex genetic and metabolic circuits, it becomes necessary to be able to link the intracellular concentration of target molecules to the regulatory circuit being implemented. One such implementation would be to use the levels of cellular RNAs to regulate the activity of CRISPR-based transcriptional programs. To that end, there have been several demonstrations that the activity of a gRNA can be regulated by the presence of target nucleic acid sequences that hybridize to the gRNA [120]. While there have been slightly different implementations, the general principle is that a trans-acting 'trigger' strand is able to bind to a gRNA and either occlude or reveal the spacer sequence, modulating the guide's activity. One of the most common mechanisms is inspired by a previously published 'toehold switch', in which a cis-repressed gRNA is activated upon toehold-mediated hybridization (Figure 2) [121]. Several gRNA switches have even demonstrated the ability to respond to RNA trigger strands within a

cell to control gene expression levels [122–125]. For example, 'toehold-gated' sgRNAs (thgRNAs) were capable of inducing CRISPRi in response to endogenous small RNAs (sRNAs) and mRNAs in *E. coli*, with repression up to 5-fold [122]. However, half of the thgRNAs responsive to endogenous RNAs resulted in low levels of repression (< 2-fold). Unlike short synthetic RNA trigger strands, longer endogenous RNAs may be less effective as trigger strands due to competition for binding from intramolecular RNA structure and cellular proteins. Improving the activity of thgRNAs responsive to cellular RNAs will require advancements in the *a priori* identification of sites within cellular RNAs that can be utilized as highly-active trigger strands. Furthermore, in order to apply these mechanisms to CRISPRa, it will be necessary to ensure that the interaction between a trigger-responsive scRNA and a large cellular RNA does not itself interfere with the mechanism of CRISPRa activation.

Metabolite-responsive gRNAs for CRISPRa

In addition to regulation by cellular RNAs, the ability to regulate CRISPR activity in response to real-time concentrations of cellular metabolites would provide many opportunities for implementing and accelerating DBTL cycles for metabolic engineering. Metabolite responsive gRNAs could be used for both readouts of intracellular metabolite concentrations to inform machine learning models, or for implementing model-suggested regulation such as feedback or feedforward motifs. While efforts to design metabolite-responsive gRNAs are fairly new, metabolite-responsive RNAs have become useful tools in metabolic engineering [126,127]. By combining a metabolite-binding RNA aptamer with a control structure, the binding state of the aptamer can be converted into a conditional genetic output.

There have been several demonstrations that small molecule responsive gRNA activity can be dynamically regulated with an aptamer in *cis* [91,92,128,129]. Some demonstrations involve inserting the aptamer at the 3' end of the guide, where it stabilizes the active gRNA structure in a ligand-responsive manner [128,129]. However, adding both an aptamer and a recruitment hairpin to the 3' end of the RNA could interfere with gRNA folding and function. Other strategies, which utilize 5' extension or Cas9 handle insertion, may therefore be preferable. For example, an aptazyme, or ligand-responsive self-cleaving ribozyme, was used to remove a repressive 5' extension from the guide upon the addition of the target ligand (Figure 2) [92]. This resulted in ligand-responsive control over both Cas9-mediated cleavage and

CRISPRa in mammalian cells. In another example, aptamers were inserted into the Cas9-binding handle, or one of two other gRNA hairpins, generating ligand-responsive CRISPRi in *E. coli* (Figure 2) [91]. Depending on the aptamer insertion site within the guide, the addition of ligand can either activate or deactivate CRISPRi.

The above successes in identifying ligand-responsive gRNAs provide great confidence that it will be possible to engineer small molecule-responsive scRNAs for conditional CRISPRa. However, there are still hurdles to overcome before metabolite-responsive CRISPRa can be used effectively for metabolic engineering applications. First, design rules that allowing reliable integration of aptamers with diverse sequences and secondary structures into scRNAs must be uncovered. Second, mechanisms for tuning the response to match the desired metabolite concentrations must be developed [130]. Aptamer-regulated kinetic control mechanisms, similar to those found in natural bacterial riboswitches, may provide an approach for engineering metabolite-responsive CRISPRa targeted to specific concentrations of metabolites [131,132]. For example, 10-fold variations in switching concentration among a family of *E. coli* thiamine pyrophosphate (TPP) riboswitches are known to be the result of differences in the amount of time the RNA is available to interact with the ligand [132]. Creating kinetically-controlled aptamer-regulated scRNAs may confer the ability to engineer metabolite-responsive CRISPRa systems functional as feedback controllers, or production biosensors useful for optimizing strain performance.

Conclusions

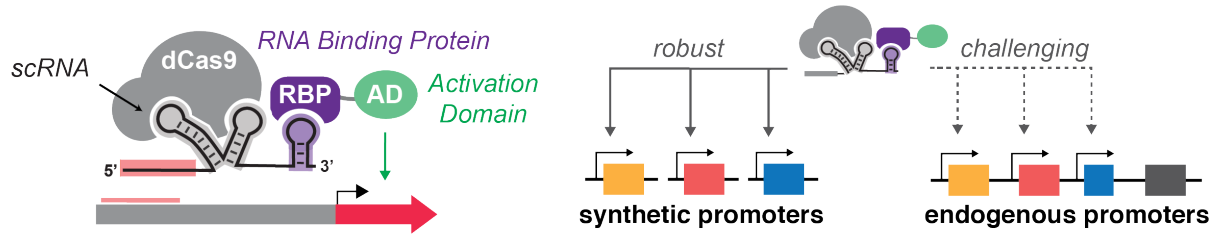
The relationships between the expression levels and reaction kinetics for enzymes in both endogenous and engineered metabolic networks are poorly understood. This incomplete knowledge constitutes a major limitation for the field of metabolic engineering [77]. Because of these gaps, data-driven methods relying on cycles of genetic engineering, high-throughput production screening, multi-omics analysis, and machine learning have become increasingly central to strain optimization [133,134]. To accelerate data-driven metabolic engineering, methods to independently target and predictably manipulate the expression levels of multiple genes are needed. By coupling new tools for CRISPRa with existing approaches for CRISPRi, it should be possible to more efficiently search gene expression spaces and optimize bioproduction in engineered bacteria through accelerated DBTL cycles (Figure 3).

CRISPRa can now be used to selectively activate synthetic promoters with large dynamic ranges and in a way that is relatively straightforward to implement. Recent advances in gRNA design have enabled the identification of small-molecule responsive gRNAs able to dynamically-regulate gene expression in *E. coli*, opening the door for the development of CRISPRa-based metabolite biosensors and circuit controllers. The ability to use CRISPRa to activate endogenous genes remains limited by the sequence constraints of the native genomic loci, where less-than-optimal position of PAM site or the inherent features of the promoters can significantly impact the activation that can be achieved. Predictive models are needed to identify which endogenous genes can be activated and which target sites are the most effective. The refinement of sequence and structure-based rules for constructing synthetic promoters and cognate scRNAs for expressing heterologous genes will improve the ability to precisely tune multi-gene pathways. CRISPRa has been demonstrated in *E. coli* [87–90,96,97,101] and other industrially and medically relevant bacteria including *B. subtilis* [98], *K. oxytoca* [89,90], *L. enzymogenes* [99], *M. xhantus* [100] and *S. enterica* [90]. Porting these tools to other non-model bacteria with diverse substrate utilization, a range of metabolic capabilities, and resistance to harsh bioprocessing conditions could accelerate the development of efficient bioproduction processes. Collectively, these strategies lay the groundwork for more widespread use of bacterial CRISPRa in basic research and advanced applications in data-driven metabolic engineering.

Figures

CRISPRa recruits an activation domain upstream of target genes.

Multi-gene programs can be implemented by expressing multiple scRNAs.



CRISPRa activity is determined by multiple factors, including:

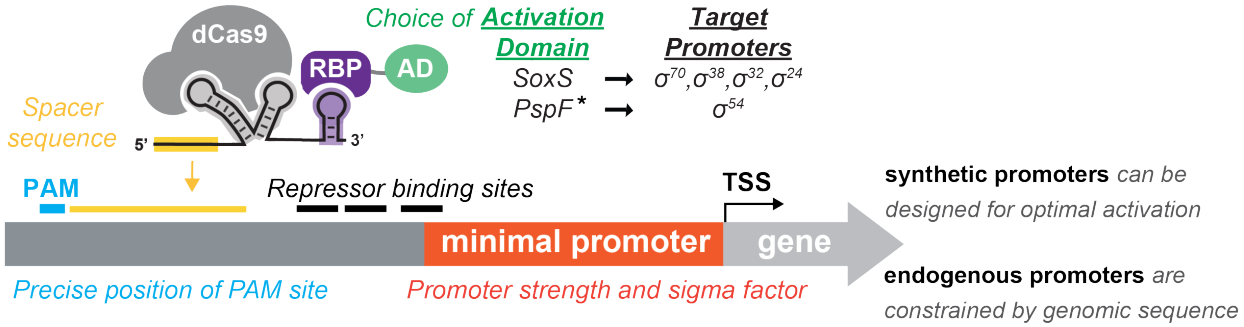


Figure 1. CRISPR activation (CRISPRa) is a powerful tool for programmable activation of genes in bacteria. A CRISPRa system is shown where an activation domain is recruited to dCas9 using a modified guide RNA (scaffold RNA, scRNA) that binds to an RNA binding protein-activation domain fusion (RBP-AD). While CRISPRa can be used to robustly activate synthetic promoters designed for optimal activation, activating endogenous genes is constrained by the genomic sequence. Factors known to determine CRISPRa activity are indicated. * : PspF activation was demonstrated using a different modified sgRNA design where two BoxB aptamers were incorporated into the Cas9 handle.

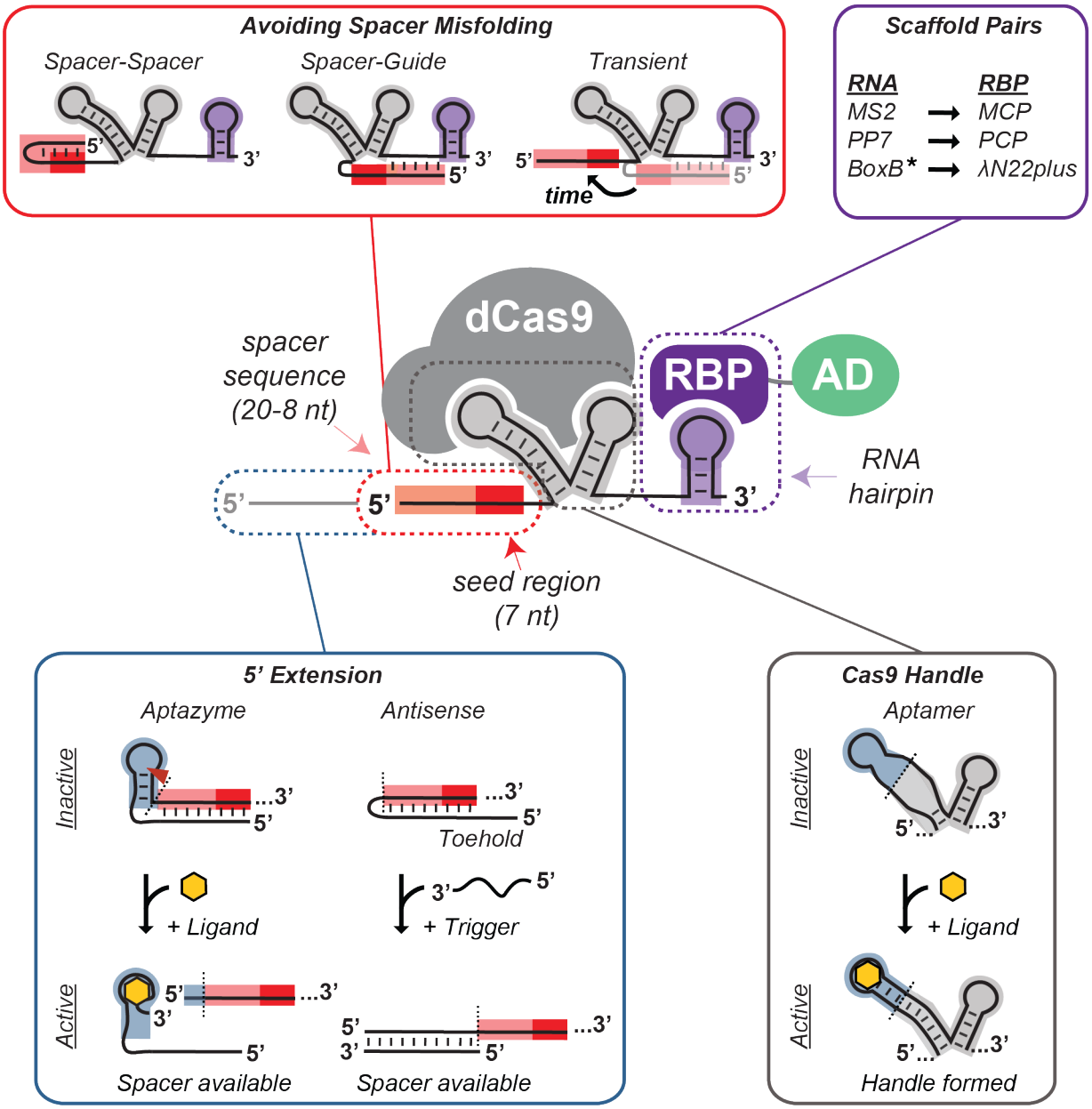


Figure 2. Guide RNA (gRNA) structural determinants of CRISPRa activity. gRNA structure can be deleterious to function, as in the case of spacer misfolding, or can be useful for programming dynamic responses. Example dynamic gRNA engineering strategies that may apply to bacterial CRISPRa include extending the 5' end of the spacer to respond to ligands or RNA trigger strands. Ligand-responsive CRISPRa activities may also be obtainable by inserting ligand-binding aptamers into the Cas9 binding handle. Several cognate pairs of RNA binding protein (RBP) and RNA hairpin have been demonstrated in bacteria, enabling the simultaneous implementation of CRISPRa with different activation domains (AD). * : CRISPRa using the BoxB:λ22plus pair was demonstrated using two BoxB aptamers incorporated into the Cas9 handle.

CRISPR-Cas engineering can accelerate DBTL cycles.

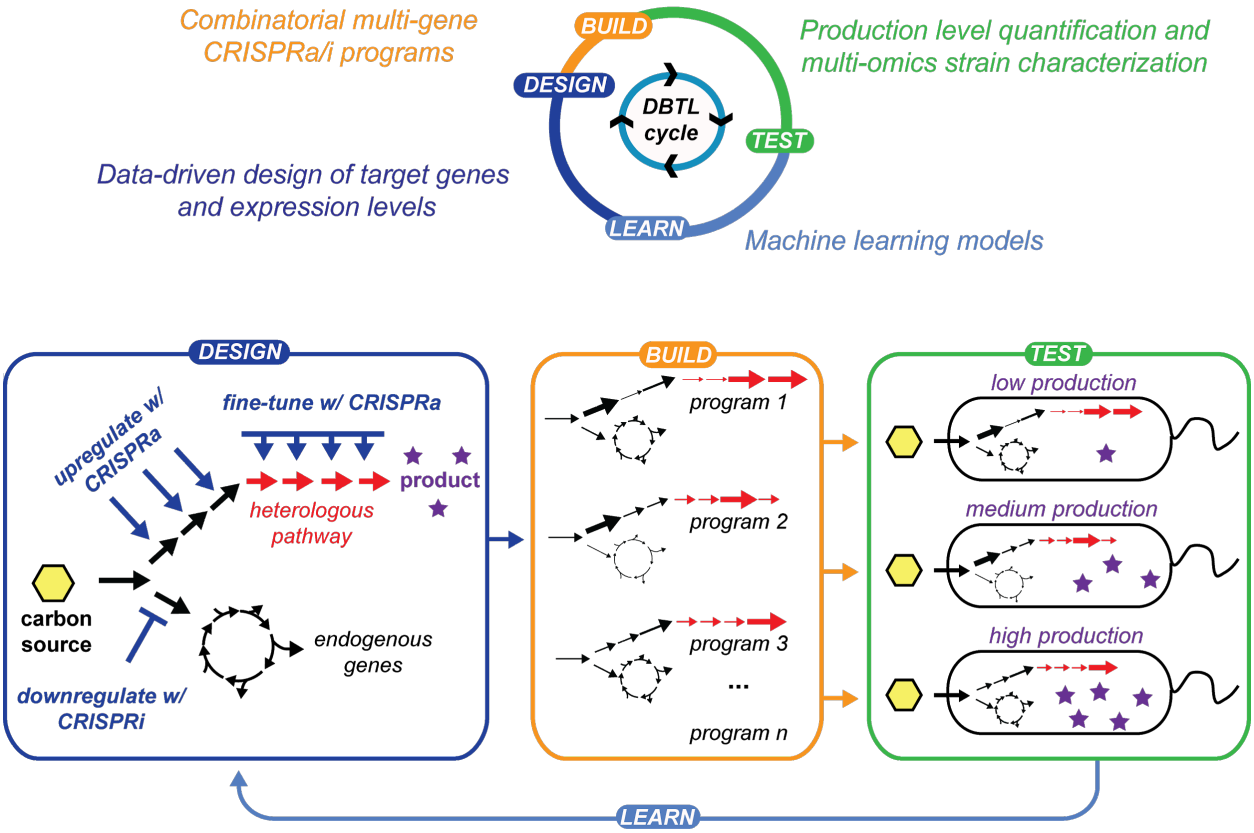


Figure 3. Developing robust workflows to integrate CRISPRa/i engineering into data-driven workflows will create new capabilities for rapidly optimizing chemical production. In this conception, each Design-Build-Test-Learn (DBTL) cycle uses machine learning and data-driven design to engineer multi-gene CRISPRa/i programs. After each build phase, production titers are measured and the strains are characterized using multi-omics analysis. These data are employed to refine the models and drive the design of CRISPRa/i programs for the next DBTL cycle.

Chapter 4: Optimization of CRISPR Activation in Bacteria

Introduction

CRISPRa has emerged as a powerful new tool for the facile re-wiring of cellular metabolism in bacteria[88,135,136]. The ability to simultaneously regulate the independent gene expression levels of multiple heterologous genes within the same cell provides tremendous opportunity for the combinatorial implementation of complex metabolic pathways. The ability to uniquely define the spacer sequence inserted within synthetic CRISPRa promoters, combined with the highly orthogonal nature of scRNA-mediated transcriptional activation, means that this system can theoretically be used to generate arbitrarily large networks of orthogonal transcription factors.

While many of the rules that govern CRISPRa activity have already been derived, there is one glaring exception[137]. scRNAs that vary only in their spacer sequence exhibit wide ranges of activity, yet there is no published data or tools to predict their activity. Guide RNA design tools trained on eukaryotic gene editing data sets have extremely poor predictive power over the spacer-specific variation in bacterial CRISPRa activity. For this reason, we decided to apply our knowledge of RNA folding to determine whether or not CRISPRa activity levels are the result of underlying biophysical relationships.

There are a number of strategies that have been proposed to modulate the activity of guide RNAs for various CRISPR applications, including 5' antisense extensions, spacer-target mismatches, and spacer truncations[117,119,138,139]. Utilizing modified guides with altered activity levels is extremely attractive, as it allows the implementation of different CRISPRa activities at each target gene, ideal for generating combinatorial libraries. In theory, spacer truncations represent the simplest solution to reducing scRNA activity. However, truncations often display non-monotonic behaviors and diverse activities at a given truncation length. In order to predict activity levels across different spacer sequences and truncation lengths, we apply our computational RNA prediction tools to identify a common set of biophysical parameters that correlate with activity.

While the ability to predict the level of CRISPRa activity from the spacer sequence alone would represent a significant step forward for the forward engineering of CRISPRa-based systems, there are a number of applications for which the prediction of scRNA activity, when the non-spacer elements are varied, would be extremely useful. For example, as the number of scRNAs simultaneously expressed

increases, so too does the genetic instability. The re-use of a large number of DNA components increases the likelihood of homologous recombination and therefore loss of desired system behavior. To combat this, creating modified sequence variants of the constant regions within scRNA, while retaining activity, has been a high priority. Additionally, the ability to develop a set of rules for the computational generation of functional Cas9-binding handle sequences would enable our kinetic biosensor design pipeline to be applied to the generation of ligand-responsive scRNAs for dynamic CRISPRa.

Here we present the Wayfinder Algorithm for computationally predicting the activity of an scRNA solely from its nucleotide sequence. Next, we show that Wayfinder can also predict the activity of scRNA truncations, allowing a wide range of expression levels to be achieved without preliminary in vivo validation. This opens the door for the forward engineering of complex systems in which the intermediate CRISPRa levels cannot be readily assayed. Next, we compare the Wayfinder Algorithm to other guide activity prediction tools, and demonstrate that it dramatically outperforms the rest in our system. This suggests that the Wayfinder Algorithm is gaining a biophysical insight into guide activity that these other models are not, and it may bring important value to broader guide RNA design efforts. Finally, we determine the sequence and structure-conservation rules of the Cas9-binding handle in order to enable the engineering of ligand-responsive scRNAs.

Methods

Plasmid Assembly

Plasmids were cloned using standard molecular biology protocols. Plasmids expressing the CRISPRa components (dCas9, the activation domain and one or more scRNAs) were constructed using a p15A vector. *S. pyogenes* dCas9 (*Sp*-dCas9) was expressed using the endogenous *Sp.pCas9* promoter. The MCP-SoxS activation domain containing mutant SoxS was expressed using the BBa_J23107 promoter (http://parts.igem.org). scRNAs used the b2 design, in which where the endogenous tracr terminator hairpin upstream of MS2 is removed[88]. The scRNAs, including the LR-scRNA were expressed using the BBa_J23105 promoter. Plasmids expressing target genes for CRISPRa were constructed using a low-copy pSC101** vector. mRFP1 and metabolic pathway genes were

expressed from the weak BBa_J23117 minimal promoter (http://parts.igem.org) preceded by synthetic DNA sequences containing the CRISPRa target sites.

Wayfinder Algorithm for spacer activity prediction

A set of spacers was generated containing diverse sequence and structural properties. The only consistent rule was that the ms2 aptamer at the 3' end of the construct was predicted to fold correctly. This was done for two reasons. One reason was to eliminate any confounding cases where an scRNA doesn't fold in a way that allows it to readily bind to the MCP-SoxS activator, which could enable the scRNA to occupy the target DNA without the activator present, leading to unpredictable outcomes. The other reason was due to the ms2 hairpins resemblance to a transcriptional terminator, due to it being a hairpin immediately 5' of a poly-T stretch. This would potentially aid in transcriptional termination after transcription of the scRNA, as read-through due to transient misfolding could yield scRNA sequences with 3' tails of variable length, again confounding the results.

The scRNAs were expressed from a strong BBa_J23119 promoter. The scRNA-containing plasmids were transformed into *E. coli* strain MG1655 containing a second plasmid with the corresponding reporter gene. Three colonies for each double transformation were grown for 24 hours in 400 µL of MOPS EZ-Rich defined medium (Teknova) containing the appropriate antibiotic. Cultures were grown in 96 deepwell plates with rapid shaking at 37C. After 24 hours of growth, 200 µL of each culture was measured in a 96-well plate format in a Synergy HTX plate reader (BioTek) with gain 35.

Wayfinder generate predictions were generated using a combination of the MFEpath algorithm for co-transcriptional folding described above (Chapter 1) or various algorithms from the ViennaRNA folding package version 2.3.5. MFEpath was used, with a mfthreshold barrier of 7.8 kcal/mol to predict the structures that the scRNAs would adopt during transcription, as well as estimates for how long the scRNAs would take to transition to their MFE structures. Net binding energy was calculated by calculating the RNA-RNA free energy of the spacer sequence binding to its reverse-complement sequence using RNAduplex (Vienna). Subsequently, the $\Delta\Delta G$ between the MFE of the scRNA and the structure wherein the Cas9-binding handle is correctly folded and the spacer is unstructured was evaluated using RNAfold with constraint folding. This value was then added to the duplex energy in order to estimate the net

energetics of binding to a single-stranded nucleic acid target. The kinetic barrier was calculated by using the Findpath algorithm to predict the barrier height for the direct refolding pathway from the MFE structure to the structure wherein the Cas9-binding handle is correctly folded and the spacer is unstructured.

Handle sequence and structure conservation

Pools containing randomized bases within the Cas9 binding handle were transformed into *E. coli* strain DH10B along with a second plasmid containing the reporter gene. The reddest colonies on the plate were picked and grown up in LB for 24 hours in 14 mL culture tubes. After 24 hours, 200 µL of each culture was measured in a 96-well plate format in a Synergy HTX plate reader (BioTek) with gain 35. All cultures with RFU/OD_{600} values greater than 90% of the wild type handle sequence's RFU/OD_{600} value were subsequently submitted for sequencing.

For analysis of sequence and structure conservation, only the positions that were variable in a given sub-pool were considered. With the exception of the closing G-U (or U-G) bases, only bases represented in more than 10% of the total sequences were considered allowed bases. For structural conservation, only positions in which the pairing status (base-paired or not base-paired) matched the computationally-predicted pairing status at the corresponding position of the MFE structure of the wild-type sequence were considered. Only base-pair types (G-C, A-T, or G-U) that were represented in more than 10% of the total sequences were considered allowed base-pair types.

Using the sequence and structure conservation rules derived above, novel handles were generated and inserted into a common scRNA context with the J306 spacer sequence. The scRNAs were either expressed from a medium-strength BBa_J23105 promoter, or a strong BBa_J23119 promoter. The scRNA-containing plasmids were transformed into E. coli strain DH10B containing a second plasmid containing the reporter gene. Three colonies for each double transformation were grown for 24 hours in 400 µL of MOPS EZ-Rich defined medium (Teknova) containing the appropriate antibiotic. Cultures were grown in 96 deepwell plates with rapid shaking at 37C. After 24 hours of growth, 150 µL of each culture was measured in a 96-well plate format in a Synergy HTX plate reader (BioTek) with gain 35.

Results and Discussion

Wayfinder Algorithm for spacer activity prediction

In order to increase the predictability of the bacterial CRISPRa system we set out to determine whether the kinetics of RNA folding were a significant cause of variation among scRNAs with different spacer sequences. We observed that scRNAs with randomly-selected spacer sequences displayed wide variations in CRISPRa activity despite satisfying all of the known rules for basic CRISPRa activity in bacteria[137]. To determine whether we could use computational RNA folding predictions to quantitatively predict scRNA activity, we built and tested 25 novel scRNA constructs that varied only in the sequence of their 20-base spacer sequence and the corresponding 20-base target DNA sequence within a synthetic CRISPRa promoter driving RFP expression. As expected, the tested scRNAs exhibited dramatic differences in reporter fluorescence, varying by more than 40-fold. We applied the Wayfinder algorithm to predict the barrier height (kinetic barrier) for conversion from the MFE structure, (in which the spacer sequence forms base-pairs with itself, the rest of the scRNAs, or both) to the structure in which the handle is correctly folded, the MS2 hairpin is correctly folded, and in which the spacer is unstructured (Figure 1A). Strikingly, this one parameter explained the majority of the variation we observed amongst the tested scRNAs (Figure 1B). To investigate whether or not reducing the barrier value even further would result in higher activation, we designed an additional 5 scRNAs containing little-to-no undesired structure (Figure 1B, green dots). While all 5 scRNAs did display high activation levels, they were not all as high the best performing scRNAs possessing larger kinetic barrier heights (Figure 1B, orange dots), which were used in the subsequent chapter. The sigmoidal shape of the relationship between kinetic barrier and activation levels may explain some of the lack of increase, as it appears the response saturates at kinetic barrier heights smaller than 10 kcal/mol.

We then compared the ability of the Wayfinder algorithm to predict scRNA activity to the most commonly used guide RNA activity prediction tools from the literature[112–114]. Interestingly, the other tools showed extremely poor correlation with our dataset (Figure 2). The Azimuth model displayed the best Spearman rank correlation with a value of 0.25, however this was much smaller to the value of 0.8 from the Wayfinder algorithm. While the other algorithms were generally multiple linear regression models trained on very large gene editing datasets, the lack of predictive power in our simpler system suggests that these models could be improved by the biophysical predictions provided by the kinetic barrier.

Despite this dramatic success in the prediction of CRISPRa activity a priori from the spacer sequence, there remain a number of areas that merit further investigation. Interestingly, the handle fraction, which is the fraction of the population of scRNAs expected to have the Cas9-binding handle correctly folded, demonstrates no correlation with CRISPRa activity in our system (R^2: 0.05). This suggests that the primary determinant of CRISPRa activity in our system is not how well the scRNA binds to dCas9, but instead how well the scRNA-dCas9 complex is subsequently able to bind to its target DNA. Another interesting result is that the guides designed to have kinetic barrier heights of zero did not perform as well as some guides with higher barriers, and similar net binding energies. Another area for future investigation is how the structure of the spacer domain impacts the RNA half-life of the scRNA, and therefore the effective concentration of scRNA available to bind to dCas9. It is known that cellular RNAs with unstructured 5' ends are more rapidly degraded than those with structured 5' ends[140]. Thus, while the reduction of unwanted structure reduces the kinetic barrier, it may also decrease the scRNA abundance, resulting in suboptimal performance.

In order to implement different levels of transcriptional activation at a target promoter, without changing the DNA sequence of the target promoter, scRNAs with varied degrees of spacer truncation can readily be implemented (Figure 3A). However, while truncating the spacer generally results in the reductions of CRISPRa activity, the response is often nonmonotonic. The length of the spacer sequence alone is a poor predictor of CRISPRa activity for scRNAs with truncated spacer sequences (R^2: 0.66). In order to improve predictions, we applied the Wayfinder algorithm in order to capture the decreased energetic favorability of binding with spacer truncations, we combined the net binding energy and kinetic barrier metrics, yielding a unified metric with good prediction accuracy (Figure 3B). One area for future consideration is variability in transcription start site. Bacterial sigma70 promoters are known to have preference for initiating transcription from a G or A[141]. When T or C are present at the +1 position, transcription can initiate from the -1 or +2 position instead, with additional impacts on the total transcriptional yields for a given promoter sequence. Truncations to bases other than G or C likely provide uncertainty in the actual sequence that is being transcribed, which in turn may have significant adverse effects on our ability to accurately predict their activity. Adding a constant 5' sequence element to

standardize transcription initiation site and rate should enable the more accurate prediction of the activity of scRNAs with truncated spacers.

Handle sequence and structure conservation

In order to be compatible with our kinetic biosensor design pipeline, we need to be able to vary the sequence of the dCas9-binding handle without interfering with any of the conserved sequence, or structure, elements necessary to bind to dCas9 effectively. To do so, we set out to randomize the identity of the positions within the handle, and then screen the resulting pool for specific sequences that are capable of retaining the activity of the wild-type handle (Figure 4A). By collecting enough sequence isolates it should be possible to reconstruct the sequence- and structure-conservation rules necessary to generate highly functional alternative handles *de novo*[142]. Due to the low probability of recovering a base-pair present in the wild-type handle, when a given position is allowed to be any base, a number of smaller pools were used in screening. Each of the smaller pools only varied a subset of the positions, in order to make sure that the odds of recovering function were greater than 1:1000, and would therefore be amenable to plate-based screening. After collecting 43 isolates possessing at least 90% of the wild-type activity, the conservation of base type, and base-pair type, at each position was determined by accepting only elements that occurred in more than 10% of the isolates (Figure 4B). This screening approach differed from a previous iterative semi-rational design and identified similar, though not identical, sequence conservation rules[142].

After identifying the conservation rules, we wished to know how well the rules could predict functional handles *de novo* (Figure 4C). We tested 10 novel handles, alongside another modified handle we had previously identified (20f), which was used to engineer the highly-unstructured scRNAs (Figure 1, green points). All of the engineered handles except for one showed significant CRISPRa activity, while several retained nearly wild-type levels. The one handle with minimal activity appears to have been caused by folding issues other than the handle sequence, however, as when it is tested with a different spacer sequence, the activity increases significantly. In addition to testing the handle constructs under a high strength promoter (119), we also tested them when expressed from a weaker promoter (105), as we suspected that the highest performing scRNAs were saturating the CRISPRa response. As expected, the

difference in performance with the wild-type sequence was exacerbated for nearly all of the handles. As the original screening of the handle variant pools was performed with the scRNAs being expressed from a strong promoter, it is very possible that the responses we observed were compressed on the upper end, and had the scRNAs been expressed from the weaker promoter, we may have developed a more stringent set of sequence- and structure-conservation rules.

Figures

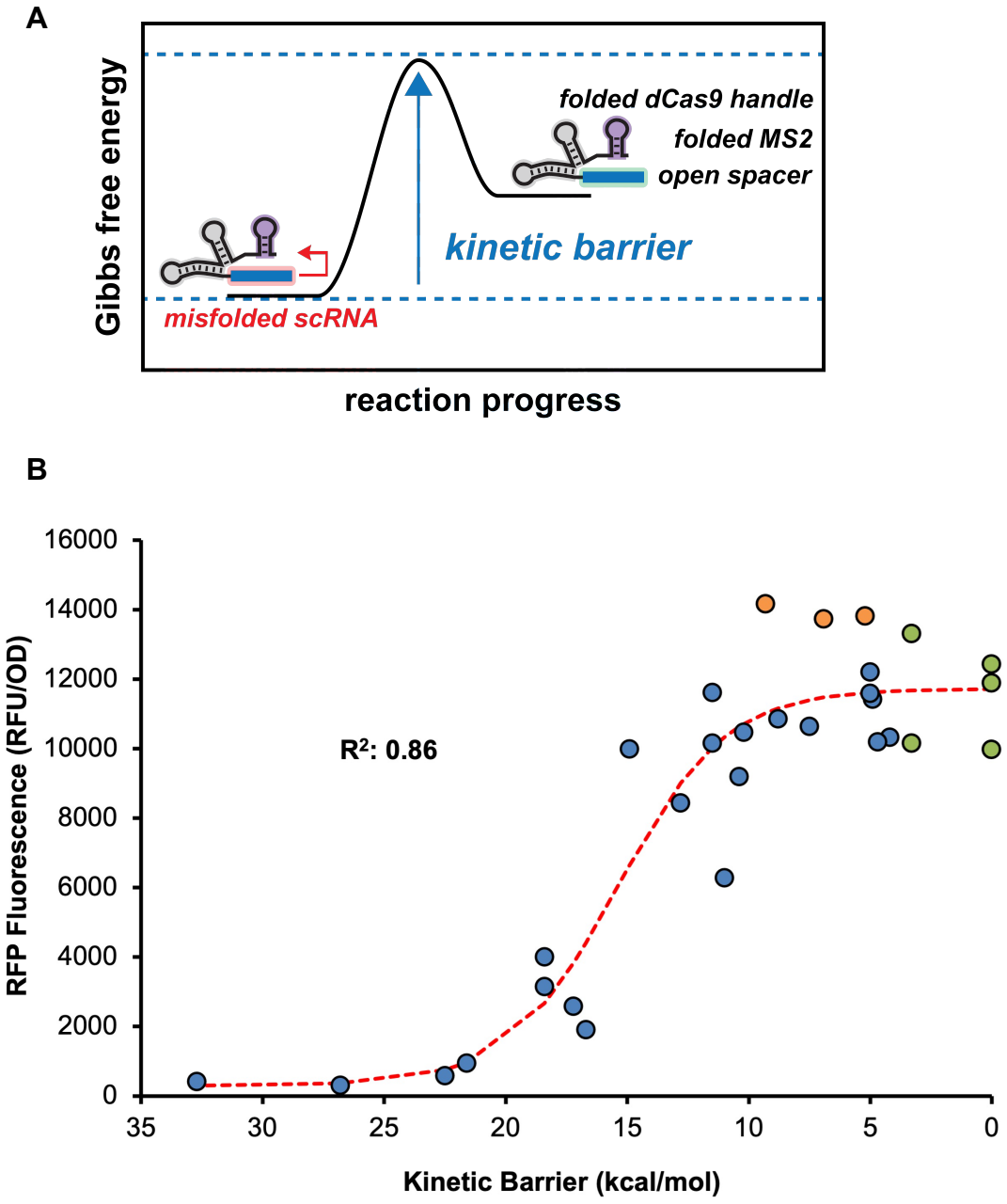


Figure 1. The kinetic barrier predicts impact of spacer sequence on CRISPRa. A) As the kinetic barrier gets smaller, scRNAs can more rapidly fold into the Cas9-competent conformation, where they are able to participate in complex formation. B) We tested 30 scRNAs in which we varied their 20 base spacer sequence. The computationally-predicted kinetic barrier height accurately predicts scRNA activity. Orange dots represent scRNAs that were chosen to engineer synthetic CRISPRa promoters for subsequent applications. Green dots represent scRNAs designed with a more-stable handle that enables highly-unstructured spacer sequence.

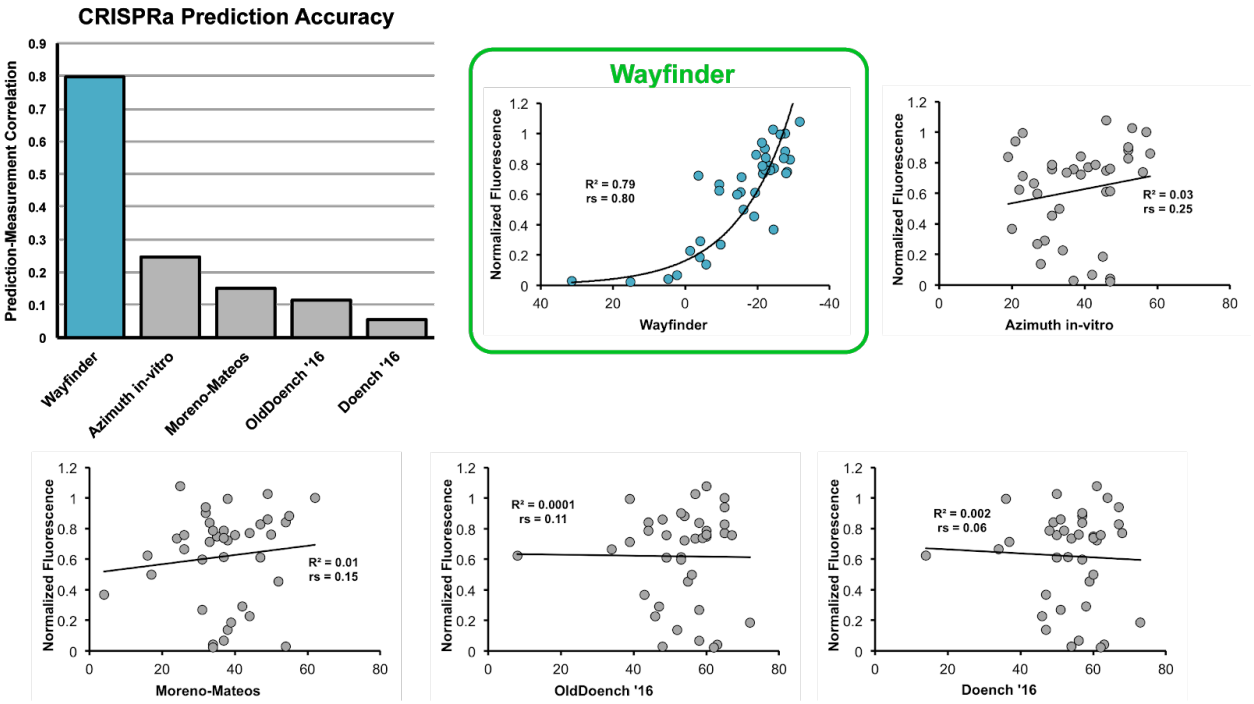


Figure 2. Comparison of Wayfinder algorithm to other common guide prediction tools. Wayfinder significantly outperforms the other common guide RNA activity prediction tools when applied to our CRISPRa dataset. Most other tools are trained on large eukaryotic gene editing datasets. When combined with other types of models, our biophysics-based model may provide new insight into the activity of guide RNA activity in diverse contexts.

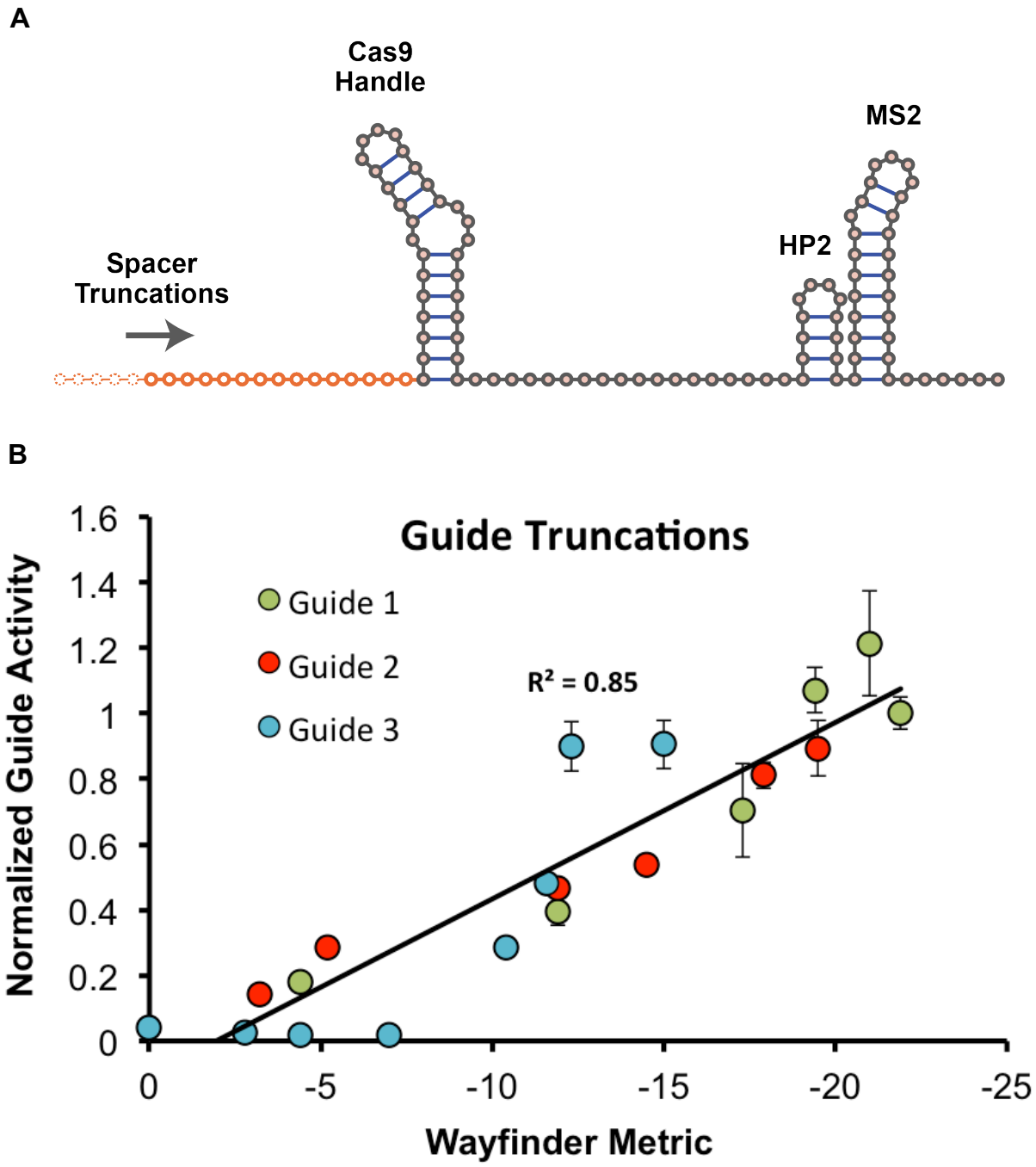
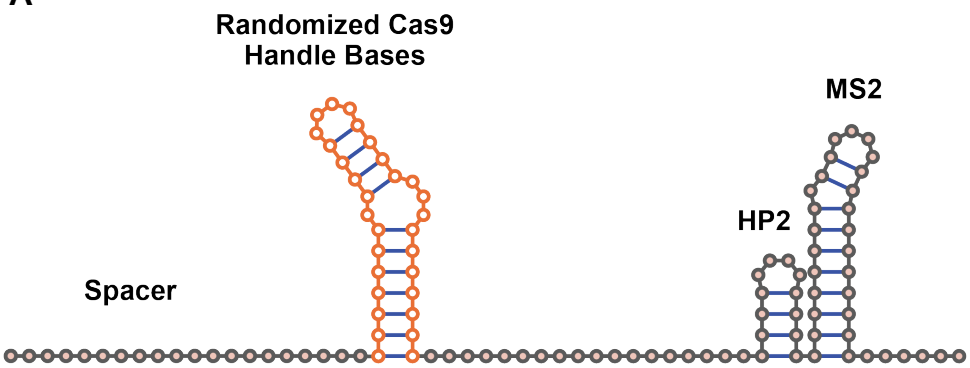


Figure 3. The Wayfinder algorithm can predict the activity of truncated scRNAs. A) Truncations of the 5' end of the spacer sequence result in decreased CRISPRa levels, by decreasing the stability of the CRISPRa complex binding to its target DNA. B) By incorporating the net binding energy of the RNA binding to its target with the kinetic barrier, the Wayfinder algorithm is able to predict the activity of scRNAs with truncated spacer sequences.

A



B

Sequence Constraints

```
101125*24220216423222 1^0661034G *=11 ^=10
13002502120130200221020052110 0C
11815513211435140224520042483 0A
3385660352512412411535 0*551812T *=12
```

Structure Constraints

```
(((((((.((((....)))).)..)))))))
1311 4*0-2421----1242---0*41131X *=10
14*5791-6334----4336---1975*41Y *=16
4001 42*-4031----1304---*241004Z *=11
```

C

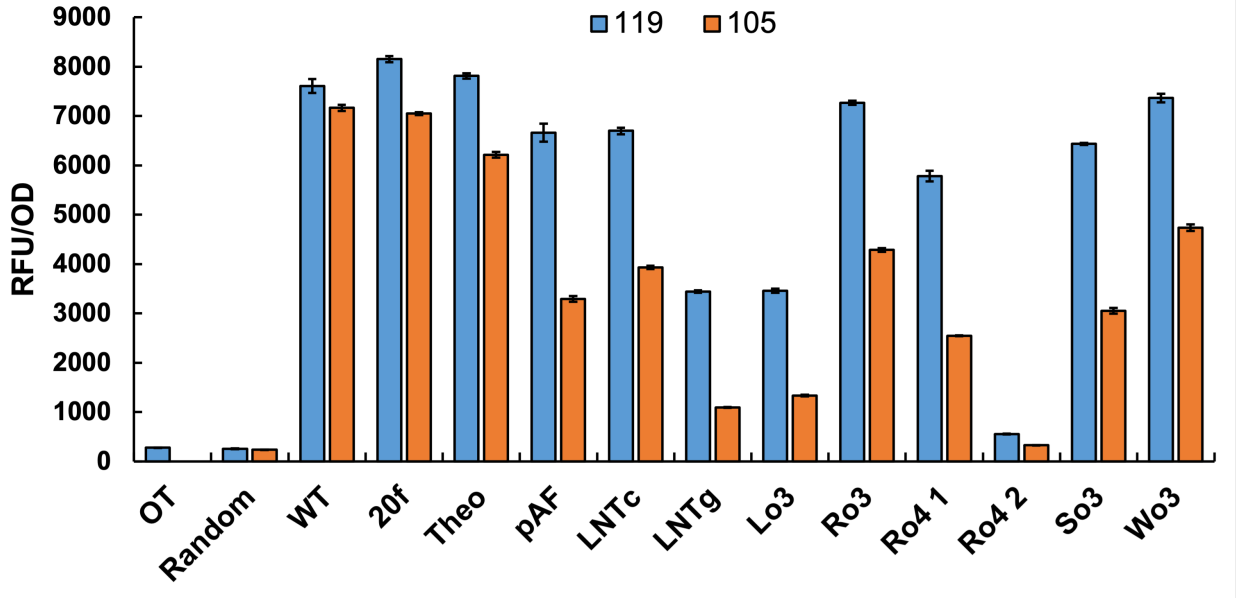


Figure 4. Identification of rules for the modification of scRNA handles. A) In order to determine the conservation rules of the Cas9-binding handle, we varied all positions and identified isolates retaining near wild-type levels of CRISPRa activity. B) The base conservation rules identified from our *in vivo* screen. Counts of isolated sequences possessing the given base or base-pair type are presented. Base-pair types are as follows: X= G-C, Y= A-U, Z = G-U. Bases and base-pair types represented less than 10% of the time are considered forbidden and highlighted in red. Due to the number of isolates possessing no base-pair in the closing stem, a G-U base pair was selected, as this matches the wild-type sequence, and represents the least-stable conventional base-pair. C) Handles designed using the identified conservation rules are almost all functional when expressed from a strong 119 promoter. When expressed from a weaker 105 promoter, differences between the engineered handles and the wild-type handle become exacerbated.

Supplementary Materials

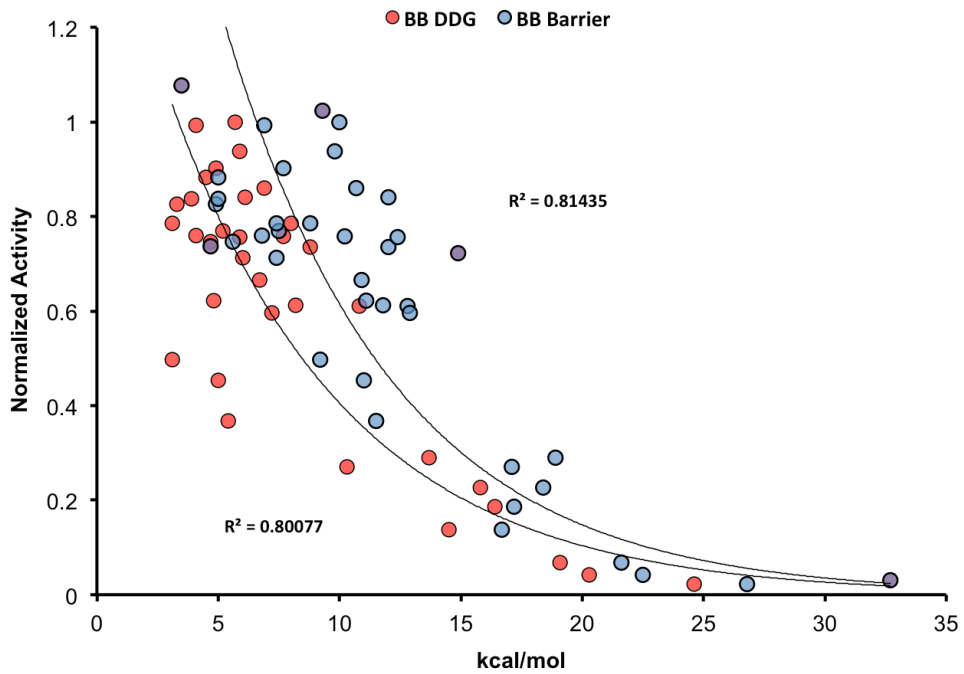


Figure S1. Barrier heights outperform ΔΔG values for identifying high-performing scRNAs. While the barrier height of the transition to the binding-competent state, and the difference in free energies between the two states show similar correlation coefficients, the barrier height helps avoid a number of significantly underperforming scRNAs, which is critical for minimizing design failures.

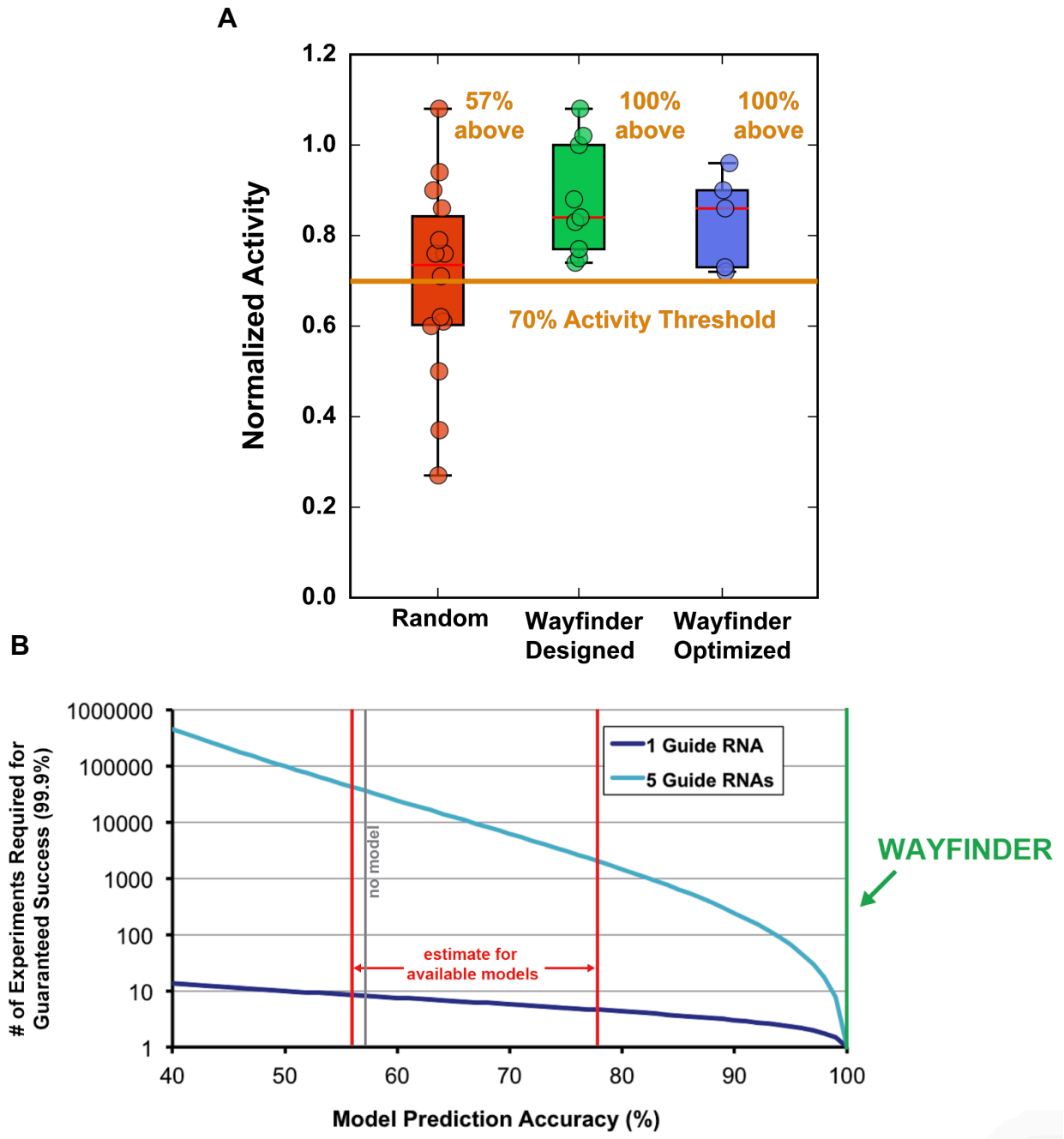


Figure S2. Accurate model predictions enable forward-engineering of complex systems. A) scRNAs screened and optimized using the Wayfinder algorithm show consistently high CRISPRa activity, while random scRNAs frequently show activity falling below a 70% threshold. B) In order to forward-engineer complex systems in which multiple novel scRNAs are required to function without prior validation, highly accurate predictions become necessary.

Name	Isolate Sequence	Pool Sequence
sm_1	CTTTTTGAGATCGAAATTCTAAGTAGAAAG	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
sm_2	GTTTAGGAAATTGAAAGGTTAAGTCTAAAA	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
sm_3	ATTTACGAACAGGAAAGGTAAAGTGTAAAT	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
sm_4	TTTTTCGATCTGGAAAAGGAAAGTGGAAAG	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
sm_5	CTTTATGAGGGTGAAATCCTAAGTGTAAAT	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
sm_6	ATTTAGGATTAAGAAATTTTAAGTCTAAAG	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
sm_7	CTTTATGAGGGTGAAATCCTAAGTGTAAAT	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
sm_8	TTTTTCGATCTGGAAAAGGAAAGTGGAAAG	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
sm_9	ATTTCCGATTTAGAAACAAGAAGTGGAAAG	**N**TTT**NN**GA**NNNN**GAAA**NNNN**AAGT**NN**A AA**N**
m_1	TTTTTAGACCTCGAAAAAGGAAGTTAAAAT	**N**TTTT**N**GA**N**CT**N**GAAA**N**AG**N**AAGT**N**AA AA**N**
m_2	ATTTTTGAACTAGAAAAAGTAAGTAAAAAG	**N**TTTT**N**GA**N**CT**N**GAAA**N**AG**N**AAGT**N**AA AA**N**
m_3	TTTTTGGATCTTGAAATAGAAAGTCAAAAT	**N**TTTT**N**GA**N**CT**N**GAAA**N**AG**N**AAGT**N**AA AA**N**
m_4	ATTTTGGATCTAGAAATAGAAAGTCAAAAC	**N**TTTT**N**GA**N**CT**N**GAAA**N**AG**N**AAGT**N**AA AA**N**
m_5	TTTTTCGACCTCGAAATAGTAAGTGAAAAG	**N**TTTT**N**GA**N**CT**N**GAAA**N**AG**N**AAGT**N**AA AA**N**
m_6	ATTTTTGATCTAGAAATAGTAAGTAAAAAA	**N**TTTT**N**GA**N**CT**N**GAAA**N**AG**N**AAGT**N**AA AA**N**
m_7	GTTTTTGAGCTCGAAAGAGCAAGTTAAAAT	**N**TTTT**N**GA**N**CT**N**GAAA**N**AG**N**AAGT**N**AA AA**N**
b_1	GTTTTAGAGCTAGAAATAGCAAGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT
b_2	GTTTTAGTGCTAGAAATAGCTCGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT
b_3	GTTTTAGTGCTAGAAATAGCTCGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT
b_4	GTTTTAGGGCTAGAAATAGCGTGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT
b_5	GTTTTAGCGCTAGAAATAGCATGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT
b_6	GTTTTAGAGCTAGAAATAGCGTGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT
b_7	GTTTTAGTGCTAGAAATAGCAAGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT
b_8	GTTTTAGGGCTAGAAATAGCATGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT
b_9	GTTTTAGCGCTAGAAATAGCATGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAA AAT

b_10	GTTTTAGAGCTAGAAATAGCTGGTTAAAAT	GTTTTAG**N**GCTAGAAATAGC**NNN**TTAAAAT
r_1	GCGTTATAGCTATTCTTAGCAAGTTAACGT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_2	GGATTAGAGCTAGTCATAGCAAGTTAATGT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_3	GCATTAGAGCTACAGATAGCAAGTTAATAT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_4	GAATTAGAGCTAGAGATAGCAAGTTAATCT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_5	GCATTAGAGCTACAGGTAGCAAGTTAATAT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_6	GAATTATAGCTATAGTTAGCAAGTTAATTT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_7	GCATTAGAGCTACTAGTAGCAAGTTAATTT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_8	GGTTTACAGCTAATTGTAGCAAGTTAAAGT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_9	GGTTTAGAGCTAAAGATAGCAAGTTAAAAT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
r_10	GAATTATAGCTAAGGGTAGCAAGTTAATAT	G**NN**TTA**N**AGCTA**NNNN**TAGCAAG**NTAANN**T
f_1	GCATGAGAGCTAGAAATAGCAAGTTCGTGT	G**NNNNNN**AGCTAGAAATAGCAAG**NNNNNN**T
f_2	GGTATAGAGCTAGAAATAGCAAGTTTTACT	G**NNNNNN**AGCTAGAAATAGCAAG**NNNNNN**T
f_3	GTTTTAGAGCTAGAAATAGCAAGTTAAAAT	G**NNNNNN**AGCTAGAAATAGCAAG**NNNNNN**T
f_4	GTTTTAGAGCTAGAAATAGCAAGTTAAAAT	G**NNNNNN**AGCTAGAAATAGCAAG**NNNNNN**T
f_5	GCTTCGGAGCTAGAAATAGCAAGTCGAAAT	G**NNNNNN**AGCTAGAAATAGCAAG**NNNNNN**T
f_6	GCTTTTAAGCTAGAAATAGCAAGTAGAAGT	G**NNNNNN**AGCTAGAAATAGCAAG**NNNNNN**T
20f_1	GTTGGTGAGTTAGAAATAACAAGTACCAAT	G**NNNN**TGAG**NNN**GAAA**NNN**CAAGTA**NNNN**T

Table S1. Sequences of isolated Cas9-binding handle variants. All sequences that displayed at least 90% of the CRISPRa activity of the WT sequence when paired with the J3 spacer sequence. The pool from which the isolate was isolated is shown, with variable positions bold

Chapter 5. Engineering ligand-responsive scRNAs (LR-scRNAs)

Introduction

While we have previously demonstrated the applicability of our computational biosensor design pipeline to the regulation of gene expression in *E. coli*, through the engineering of AS-RBS riboswitches (Chapter 2), we have not yet demonstrated our ability to engineer genetically-encoded biosensors able to respond to the production of a target molecule being synthesized within the same cell. In order to develop a robust system for the engineering and screening of diverse metabolic pathway variants, we chose to synthesize the lessons we had learned about the engineering of kinetic RNA biosensors with the lessons we had learned about optimizing scRNAs for CRISPRa. To do so, we developed a class of kinetic RNA biosensors known as ligand-responsive scRNAs (LR-scRNAs). These LR-scRNAs utilize the Cas9-binding handle of an scRNA as the output domain of our kinetic biosensor molecular architecture. As the handle is critical for the formation of the scRNA-dCas9 complex we hypothesize that the selective deformation of the handle will result in selective formation of the CRISPRa complex, and therefore ligand-responsive CRISPRa activity[111].

One significant benefit of LR-scRNAs as an alternative to AS-RBS Riboswitches is the direction of the response. As metabolic pathways express large amounts of burdensome enzymes, cellular fitness can be dramatically impacted by their production, often leading to genetic instability and suppression of the expression levels of heterologous genes[143]. In turn, many metabolic pathway variants will express significantly lower levels of a reporter protein in response to excess burden. In the case of an AS-RBS riboswitch this result could be falsely interpreted as a reduction signal due to the biosensor's regulation of the expression level of the reporter gene. While AS-RBS Riboswitches decrease translation levels upon detection of the target molecule, LR- scRNAs increase transcription levels in response. Thus, any increase in output gene expression is very unlikely to occur spontaneously, and furthermore gives an indication that the underlying genetics necessary for gene expression remain intact. In addition, an aptamer-regulated scRNA would not only enable us to regulate a fluorescent protein for extracellular quantification of intracellular metabolite concentration, but would also allow us to implement complex

genetic networks in response to those metabolite levels. For example, recent efforts have demonstrated that incoherent feed forward network motifs can be realized using CRISPRa and CRISPRi components.

In order to determine whether or not our molecular architecture applies to the engineering of LR-scRNAs able to respond to the *in vivo* production of biosynthetic products, we first decided to validate the biosensors using a well-studied aptamer that would allow extracellular addition of the target molecule. To do so we chose to engineer theophylline responsive scRNAs. By implementing the bacterial transcriptional pause sites that resulted in highly-sensitive AS-RBS riboswitches (Chapter 2), we hoped to achieve the same sensitivity, despite an entirely different mode of action.

In order to demonstrate a practical application, we chose to engineer a biosensor for a valuable industrial product. We chose to target a human milk oligosaccharide (HMO) composed of lactose combined with additional sugar monomers. HMOs are a class of oligosaccharide found almost exclusively in human breast milk, and are the third most abundant solid component in human breast milk behind lactose and lipids (Figure 1A). There exist a number of hypothesized roles that HMOs play in infant development, such as prebiotics, antimicrobials, and immune modulators[144,145]. Critically, HMO's are nearly completely absent in conventional cow's milk. As most infant formula is derived primarily from cow's milk, this discrepancy may lead to suboptimal nutrition. Thus, in order to produce a new generation of infant formulas with properties more similar to human breast milk, the identification of a cheap, renewable, and abundant source of HMOs will be essential. In fact, historically, the primary source of HMO's added to infant formula has been those purified from human breast milk. The development of a genetic biosensor able to detect HMO production in microbes would enable the high-throughput screening of extremely large numbers of microbial variants, leading to increased biosynthetic yields and a new generation of superior infant formula.

Here we demonstrate the production of theophylline-responsive scRNAs able to modulate CRISPRa activity in a highly-sensitive and dose-dependent fashion. Subsequently, we describe the development of a CRISPRa-based metabolic pathway for the biosynthesis of our HMO product (HMO-p) in *E. coli*, wherein different combinations of gene activation result in significantly varied amounts of HMO-p production. We then describe the production of completely novel HMO-responsive scRNAs. We start by performing systematic evolution of ligands by exponential enrichment (SELEX) to identify an RNA

aptamer able to bind HMO-p[146]. Finally, we incorporate our selected aptamer into LR-scRNA switch candidates and validate that the switch candidates are able to detect HMO-p production from our engineered pathway when they are genetically encoded within the same bacterial cells.

Methods

Computational screening for ligand-responsive scRNA switch candidates

Candidate LR-scRNAs were screened using the computational methods established with K-As and AS-RBS riboswitches. Elongation barrier heights of 7.8 kcal/mol were used for MFEpath predictions of co-transcriptional folding, corresponding to the rate of E. coli RNA polymerase elongation. Screening for B1 barrier heights <= 7.8 kcal/mol was implemented. Screening for B2 barrier heights <= 2.9 kcal/mol was implemented. Toe-target distance, calculated as ln(linear distance^3), were screened for values <= 10.5 arbitrary units. Screening for pathway convergence >= 0.7 was implemented. The Cas9 binding handle of the scRNAs was treated as the output domain in order to define the Overhang, and Stem sequences within the molecular architecture. In order to generate enough diversity to find satisfactory computational solutions, the Cas9 binding handle was varied, using the sequence and structure conservation rules from Chapter 4. In order to optimize the likelihood of identifying sequences that were simultaneously good switches, and high-performing scRNAs, the Linker sequence was considered to be the 3' end of the spacer (Figure 1B). The remaining 5' bases of the spacer were subsequently considered to be part of the Timer domain, and the entire switch (containing a full 20-base spacer) was re-screened using the same screening metrics.

In addition to the conventional switch screening metrics outlined in Chapter 1, the candidate switches were also screened for their ability to act as high-functioning scRNAs when the target molecule is bound to the aptamer domain. To do so, the aptamer domain was constrained and the candidate LR-scRNAs were screen using the previously established computational thresholds for highly-functional scRNA spacers. The following screening thresholds were applied: Net binding energy <= -25.0 kcal/mol, handle fraction >= 0.5, bind barrier < =10.0 kcal/mol, and bind barrier >=20.0 kcal/mol when evaluated without constraining the aptamer domain.

Once computational solutions were identified, Timer pools containing the *ThiC* transcriptional pause site were inserted 5' of the spacer sequence, and subsequently screened for performance. Colonies with low leak, corresponding to those without red coloring when plated on LB-Agar plates lacking the target molecule, were grown for 24 hours in 400 µL of MOPS EZ-Rich defined medium (Teknova) containing the appropriate antibiotic. Cultures were grown in 96 deepwell plates with rapid shaking at 37C. After 24 hours of growth, the cultures were diluted 1:100 into fresh media. The media contained varied concentrations of theophylline. After 24 hours of growth, 150 µL of each culture was measured in a 96-well plate format in a Synergy HTX plate reader (BioTek) with gain 35.

In vitro transcription and preparation of SELEX RNA

DNA pools were ordered as two separate oligonucleotides from IDT and were PAGE purified to ensure that only the correct length DNA was present. One oligo contained the variable positions of the pool, and was synthesized using equimolar base-ratios produce through hand-mixing. The second primer contained the sequence for the T7 promoter. Primer extension was performed using DreamTaq PCR Master Mix (2X) (Thermo) to produce the full-length transcription template. The resulting PCR product was treated with Exonuclease I (Thermo), and subsequently purified using a QIAquick PCR clean up kit (Qiagen).

Transcriptions, prepared using T7 RNA polymerase (NEB), were run overnight (10-12 hours) at 37˚C. For round 1, a 2 mL transcription reaction was prepared to maintain pool diversity. For subsequent rounds 200 µL transcription reactions were prepared. Transcriptions were terminated by adding 3 µL DNase I (Thermo) per 200 µL transcription and incubating for 30 minutes at 37˚C. The resulting RNA was run on a 1.5 mm PAGE gel at 20 W for 60-90 minutes, and the correct bands were identified using UV shadowing by comparison with xylene cyanol and bromophenol blue dye bands and then cut from the gel. RNA was recovered from the cut gel slices with 0.5X TBE using a Whatman Elutrap; for recovery, the Elutrap was operated at 200 volts for 2 hours. Eluted RNA was extracted from the Elutrap, ethanol precipitated with KCl, and then resuspended in water. Alternatively, RNA was extracted from the gel slices using a Small-RNA PAGE Recovery Kit (Zymo Research) used with IICR columns from an RNA Clean & Concentrator Kit (Zymo). Resuspended RNA totaled between 125-350 µg of RNA.

In vitro selection (SELEX)

Purified RNA was resuspended or eluted in 200 µL water, melted for 1 min at 80˚C and reannealed at room temperature, and combined with selection buffer containing 2.5 mM magnesium. The affinity column was packed with 200 µL gravity-settled HMO-p-gel beads (IsoSep) and equilibrated with 20 column volumes of water, followed by 20 CVs of selection buffer, followed by a final wash with 100 µg/mL BSA in selection buffer. The purified RNA in selection buffer was then added to the column and incubated for 30 minutes before flowing through. After incubation, the column was washed with 20 CVs of selection buffer. Finally, the RNA was eluted by adding 2 CVs at a time of 30 mM HMO-p in selection buffer, four times (8 total CVs), incubating the last for 1 hour, then collecting the flowthrough. All column work was performed at room temperature.

DNA pool regeneration

After elution from the column, the dilute RNA samples were concentrated using an RNA Clean & Concentrator Kit (Zymo). After elution into a small volume of elution buffer, RNA recovery was quantified using a NanoDrop spectrophotometer (Thermo). RNA samples were then reverse-transcribed using a standard reverse primer and RevertAid reverse transcriptase (Thermo). A maximum of 5 ng or RNA was used per µL of RT reaction. RT reactions were run using the manufacturer's suggested protocol. After the RT reaction, the resulting cDNA was PCR purified using a QIAquick PCR clean up kit (Qiagen). The cDNA was then amplified to return the template concentration to a constant level between selection rounds. In order to prevent a loss of diversity, due to exhaustion of amplification primers and subsequent self-priming, the correct number of PCR cycles was estimated using a predicted doubling-efficiency of 1.4, and checked by running on agarose gels as the predicted cycle number was approached. The resulting PCR product was treated with Exonuclease I (Thermo), and subsequently purified using a QIAquick PCR clean up kit (Qiagen).

Cloning and sequencing

DNA pools from the indicated selection rounds were cloned into plasmid vectors using the TOPO TA Cloning Kit and chemically transformed into competent DH10B *E. coli* cells. Transformants were plated on LB agar plates with 50 µg/mL carbenicillin, 40 µg/mL X-gal, 100 µM IPTG, and 15 white colonies per round were grown overnight in LB, miniprepped with a Qiagen vacuum kit, and submitted for Sanger sequencing provided by Genewiz.

In-line probing of aptamer candidates

Prior to analysis, aptamer sequences were structurally optimized to be suitable for subsequent use. Single-stranded tails were truncated, and the closing stem of the suggested structure were lengthened where necessary (Figure SX). Aptamer candidates were transcribed using the standard procedure above. RNA samples were purified as above. RNA samples were then incubated at room temperature for 2 days in 1X SB at either 0 mM or 10 mM of HMO-p. After incubation, reactions were quenched by mixing the samples with a formamide-EDTA loading buffer. 0.5 µL of each sample was run on an extra-long 0.75 mm denaturing (7.5 M) Urea-PAGE gel (10% monomer). Gel was stained for 15 minutes using SYBR Gold RNA stain, and imaged using a PharosFX Molecular Imager (Bio-Rad) with a SYBR Gold emission filter and high sample intensity.

Assembly of a CRISPRa-regulated biosynthetic pathway for HMO-p in *E. coli*

Plasmids were cloned using standard molecular biology protocols. Plasmids expressing the CRISPRa components (dCas9, the activation domain and one or more scRNAs) were constructed using a p15A vector. *S. pyogenes* dCas9 (*Sp*-dCas9) was expressed using the endogenous *Sp.pCas9* promoter. The MCP-SoxS activation domain containing mutant SoxS was expressed using the BBa_J23107 promoter (http://parts.igem.org). The scRNAs, including the LR-scRNA were expressed using the BBa_J23105 promoter. scRNAs used the b2 design, in which where the endogenous tracr terminator hairpin upstream of MS2 is removed[88]. Plasmids expressing target genes for CRISPRa were constructed using a low-copy pSC101** vector. mRFP1 and metabolic pathway genes were placed expressed from the weak BBa_J23117 minimal promoter (http://parts.igem.org) preceded by synthetic DNA sequences containing the CRISPRa target sites.

HMO-p pathway production quantification

Single colonies from LB-agar plates were inoculated in 2 mL MOPS EZ-Rich defined medium (Teknova) with 10 g/L glucose, 2 g/L lactose and supplemented with appropriate antibiotics. Cultures were grown in 14 mL plastic culture tubes at 37 °C and shaking for 48 h. 500 µL of supernatant from each culture were loaded onto 10 kDa microcentrifuge filters (Millipore) and spun for 20 min at 14000 rcf. 1 µL of filtered supernatants were assayed with a Shimadzu HPLC using UV-vis detection at 210 nm. HMO-p was separated using a Rezex ROA-Organic Acid H+ column (Phenomenex) and a 20 mM H_2SO_4 isocratic mobile phase. A standard curve was prepared by spiking known amounts of HMO-p into supernatants derived from cultures of *E. coli* strain JM109 transformed with empty vectors.

Validation of HMO-p pathway production sensing

Plasmids containing the HMO-responsive scRNA candidates, along with Cas9 and MCP-SoxS, were co-transformed into *E. coli* strain JM109 with either the plasmid containing the 3 pathway genes, or a plasmid with the same origin of replication and resistance marker but expressing blue fluorescent protein instead. 3 individual colonies were picked for each co-transformation, and were grown for in 400 µL of MOPS EZ-Rich defined medium (Teknova) supplemented with 8 g/L glucose and 2 g/L of lactose, as well as the appropriate antibiotics. Cultures were grown in 96 deepwell plates with rapid shaking at 37C. After 24 hours of growth, the cultures were diluted 1:1000 into fresh media. The media contained all combinations of (0 or 2 g/L lactose) and (0, 5, 10, or 20 nM aTc). After 24 hours of growth, 150 µL of each culture was measured in a 96-well plate format in a Synergy HTX plate reader (BioTek) with gain 35.

Results and Discussion

Development of Theophylline-Responsive scRNAs

In order to demonstrate that our molecular architecture could be used to measure the concentration of biosynthetic products, we first wished to validate that we could regulate CRISPRa activity in response to a membrane-permeable small molecule added to the cell culture media. To do so, we computationally designed scRNAs to be controlled by the binding state of the theophylline aptamer. In the

presence of theophylline, the scRNA should fold correctly, giving rise to an increase in CRISPRa activity (Figure 2).

As in the case of AS-RBS riboswitches (Chapter 2), we first designed candidate switches *in silico*, and then subsequently inserted a Timer pool containing the *ThiC* transcriptional pause site. The subsequent plasmid pool was screened using the plate-based method described previously. Our computational screening yielded 2 initial candidate switches, Theo-1 and Theo-2, that were expected to produce high ligand activation ratios, and large maximum signals, coupled with the characteristic high sensitivity observed in AS-RBS riboswitches containing a pause site. In initial screening, both switches produced at least 2-fold increases in RFP levels when theophylline was added to the media (Figure 3A). Interestingly, Theo1 D5 showed a significant increase in RFP levels at sub-millimolar theophylline levels, unlike the rest of the isolated switches. When characterized at lower theophylline concentrations, we determined that Theo1 D5 had an EC_{50} of 42 μM, which is essentially the same as the 47 μM EC_{50} we observed for the Theo-48 AS-RBS riboswitch (Figure 3B). Considering the unprecedented sensitivity to theophylline in a bacterial host, as well as the shared transcriptional pause sequence, this result corroborates our hypothesis that the *ThiC* transcriptional pause site is enabling our kinetic biosensors to access increased sensitivities through increases in the co-transcriptional ligand binding window.

While the overall activation ratio remained low in the theophylline-responsive isolates, this demonstration was sufficient to confirm that our molecular architecture could be applied to the regulation of scRNA activity at low concentrations of target molecule. Further investigation will be required to determine whether it is possible to couple these highly-sensitive responses with large changes in CRISPRa activity. Based on preliminary results, we believe that the absence of other scRNAs competing for binding with dCas9 drives the equilibrium towards the formation of the scRNA-Cas9 complex, even when only a small portion of the available pool of LR-scRNAs is ever in the Cas9-binding competent state. Due to the extremely long dissociation kinetics of Cas9, it is possible that this infrequent, spontaneous structural sampling is sufficient to form an effectively-irreversible complex that delivers high background activation levels. We believe that the introduction of additional guides able to readily bind to unoccupied dCas9 molecules will result in significantly lower leak levels, while hopefully still providing access to large, visible increases in CRISPRa activity.

In vitro selection of a novel aptamer for HMO-p

In order to develop an RNA biosensor able to detect HMO-p production in *E. coli*, it was first necessary to generate an aptamer able to selectively bind to HMO-p, as none existed in the literature. In fact, only a small number of aptamers targeting oligosaccharides exist in the literature, and those that do exist tend to have large, complicated structures, providing an interesting test to the molecular architecture that had previously only been validated with the smaller theophylline and pAF aptamers. To generate the aptamer, we used a classic technique known as SELEX (Figure 4A). In SELEX the target molecule is immobilized on a solid matrix, and then a highly diverse (10^{12}-10^{15}) pool of unique RNA sequences is flowed past the matrix, and allowed to interact. Those RNA sequences able to bind to the target molecule are retained, while those that don't are eliminated. After extensive washing, excess uncoupled target molecule is flowed past the solid matrix in order to competitively bind to the RNA aptamers and cause them to elute. Subsequently, the eluted population of RNA molecules is reverse-transcribed, amplified and the process in repeated until the pool of RNA is reduced to only a handful of enriched sequences. Generally, this process takes between 5 and 16 rounds of selection to achieve significant enrichment. This enrichment usually results in an increase in the mass of RNA recovered from the column, as the total number of aptamers, able to bind and be selectively eluted, within the DNA pool increases.

The HMO-p aptamer selection was run for 12 cycles, at which point the pool was submitted for sequencing, despite the lack of a significant bump in RNA recovery (Figure S1). Interestingly, the pool contained only one sequence represented multiple times in the 13 sequenced clones. Subsequent sequencing of earlier rounds revealed a total of 7 unique enriched sequences, that first began appearing in round 5 (Figure 4B). All of the enriched sequences possessed the complicated branched secondary structures associated with previous carbohydrate-binding aptamers (Figure S2).

While encouraging that sequences had enriched, it was notable that the sequences either disappeared, or reduced in abundance in the subsequent round after they became enriched. Somewhat puzzled by this result, we investigated a number of hypotheses as to why this might have occurred. We believe that the most likely cause was that the very high concentration of largely-palindromic RNA sequences resulted in homo-dimerization prior to being loaded on the selection column. We hypothesized

that these dimerized sequences would therefore be unable to interact with the matrix, and would subsequently be removed from the enriched pool. Notably, as dimerization is a second order reaction, the degree of dimerization would occur non-linearly as the concentration of specific RNA molecules increased, due to enrichment. In support of this hypothesis, HMOg, the only aptamer candidate that was able to enrich to the majority of the pool, was predicted to form the least stable homodimer. We ran samples of RNA on a native PAGE gel in order to determine whether or not higher-order species were present. Indeed, we observed bands corresponding to larger species than aptamer monomer alone (Figure S5).

To validate that the sequences that had enriched were indeed binding to HMO-p, and not enriching for other reasons such as PCR bias, we utilized another classic technique known as in-line probing. In in-line probing, an RNA aptamer is incubated in either the presence or absence of the target molecule for an extended period of time. Spontaneous hydrolysis of the RNA backbone occurs at different rates for each position based on the conformation of the molecule, and gives a characteristic banding pattern. By comparing the banding pattern in the presence and absence of the target molecule, any significant differences in intensity imply an interaction between the aptamer and its target. We performed in-line probing on a number of aptamer candidates, observing changes in banding intensity in all of the candidates (Figure S4). To confirm that the changes in banding were specific to aptamers expected to bind to HMO-p, we simultaneously incubated the pAF aptamer with HMO-p and observed no obvious change in banding pattern suggesting that all of the enriched sequences bind specifically to HMO-p.

Assembly of a CRISPRa-regulated biosynthetic pathway for HMO-p in *E. coli*

In order to generate a metabolic pathway able to produce HMO-p, we drew inspiration from the literature. Previously, another lab was able to synthesize HMO-p from lactose by introducing the HMOg1, HMOg2, and HMOg3 genes into *E. coli* strain JM109. The HMOg1 gene facilitates the uptake of lactose into the cell, while HMOg2 converts lactose into an intermediate compound (HMO-i). Finally, HMOg3 converts HMO-i into HMO-p, our desired product. In order to facilitate the ease of engineering strains with diverse combinations of gene expression levels, we placed each of the 3 heterologous genes under the control of a previously validated CRISPRa-responsive promoter. In doing so, we simply needed to

introduce scRNAs with various degrees of spacer truncations, and therefore CRISPRa activities, controlling each gene to implement a different expression program.

In order to readily screen biosensor variants for the ability to respond specifically to HMO-p production, we placed the transcriptional activator MCP-SoxS under the control of aTc-inducible Tet promoter. Thus, upon induction, the already-transcribed scRNA would bind the activator and activate expression of the target genes. While the various guides could theoretically each be placed under the control of their own inducible promoter, most such promoters contain a promoter "scar" that would be transcribed along with the scRNA, resulting in unpredictable folding, and activity, of the engineered scRNA.

Using scRNA truncation variants with diverse, and previously characterized, CRISPRa activation levels, we tested how different activation combinations impacted the cellular production of HMO-p, and its precursor HMO-i (Figure 5A). Combinations of high, low, or off-target activation at the three pathway genes resulted in HMO-p production levels varying from nearly 0 up to 200 μM. Interestingly it was the 33 variant, containing low activation of HMOg1, and then high activation of HMOg2 and HMOg3 that gave the highest production levels of HMO-p (Figure 5B). This seems to suggest that excess expression of HMOg1 is either unhelpful, or adversely impacts cellular fitness. Unsurprisingly, when the HMOg3 gene is not activated, there is no measurable HMO-p production, but its precursor, HMO-i accumulates to the highest observed levels (Figure 5C). For subsequent analysis of candidate HMO-responsive scRNAs, the 33 pathway variant was used.

Engineering an HMO-responsive scRNA biosensor

We applied the same rules as with the Theo-1 and Theo-2 LR-scRNAs to the engineering of the seven candidate HMO-binding aptamers. Of the seven, two of them (HMOc and HMOg) readily produced computational solutions, satisfying all our design constraints (Figure S6). For each switch, we identified 2 spacer variants that satisfied the biosensor screening metrics, as well as the scRNA screening metrics. As such, we proceeded with four candidate biosensors named HMOc1, HMOc2, HMOg1, and HMOg2. Interestingly, the output promoter for HMOc2, containing the target site for its spacer, upstream of a standard promoter, produced approximately 10-fold increases in uninduced expression levels compared

to the other 3 constructs. This was quite surprising as the spacer sequence in HMOc2 differed from the spacer sequence in HMOc1 by only a handful of bases. While likely not ideal for maximizing the activation ratio for any potential biosensor, this did have the added benefit of increasing the RFP expression level to the point that the redness of colonies on LB-Agar plates could be readily evaluated by eye.

Timer pools containing a *ThiC* transcriptional pause site were inserted into the 4 candidates, and the resulting pool of plasmids were evaluated for the lack of leaky CRISPRa levels in the absence of the pathway genes. In all cases the colonies were not red enough to differentiate based on color, and thus only positive selection was applied to the pool. The pool was then subsequently screened for increased redness on plates containing aTc when co-transformed with a plasmid containing the pathway enzymes (Figure 6A). Again, for all except the HMOc2 variants, it was impossible to distinguish the redness of colonies by eye.

To assess whether the candidate HMO-responsive scRNAs were able to detect the biosynthesis of HMO within an *E. coli* cell, we transformed them along with the plasmid containing the pathways genes. We also transformed the candidates along with an unrelated CRISPRa-controlled gene as a negative control. In addition, we tested all of the constructs in media containing, or lacking, lactose. Lactose is the substrate for the production of HMO-p and should be necessary to produce measurable quantities of HMO-p. Thus, only when an HMO-responsive scRNA construct is transformed with the pathway genes, grown in lactose-containing media, and gene expression is induced, should you observe HMO-specific CRISPRa. Indeed, only when the LR-scRNA candidates were combined with the pathway enzymes and were grown in culture with supplemental lactose did the signal increase (Figure 6B, C). Although it appears that the signal saturates at induction with 10 nM aTc, the reduction in signal from the constitutively active J5 scRNA indicates that burden is a factor in expression levels. It is possible that the signal increase would be even higher in a lower metabolic burden expression regime.

While the identification of a completely novel genetically-encoded biosensor to an industrially-relevant product represents a significant achievement, there are a several areas in which the system needs to be optimized in order to represent a robust platform for the high-throughput screening of biosynthetic bacteria. The first is to increase the overall signal levels. When tested in liquid culture, the HMOg variants showed no CRISPRa activity, and HMOc1 variants showed a very small amount. The

candidate HMO-responsive scRNAs were expressed from a low strength promoter, which has been observed to decrease maximal CRISPRa levels. These observations are consistent with the reduced CRISPRa levels observed from scRNAs containing these specific modified handle sequences, when weakly expressed, as in Chapter 4. Simply increasing the expression level of the scRNA candidates may be sufficient to achieve levels of HMO-induced activation able to be screened by eye, resulting in higher performance LR-scRNAs, sufficient for high-throughput screening of pathway variants.

Figures

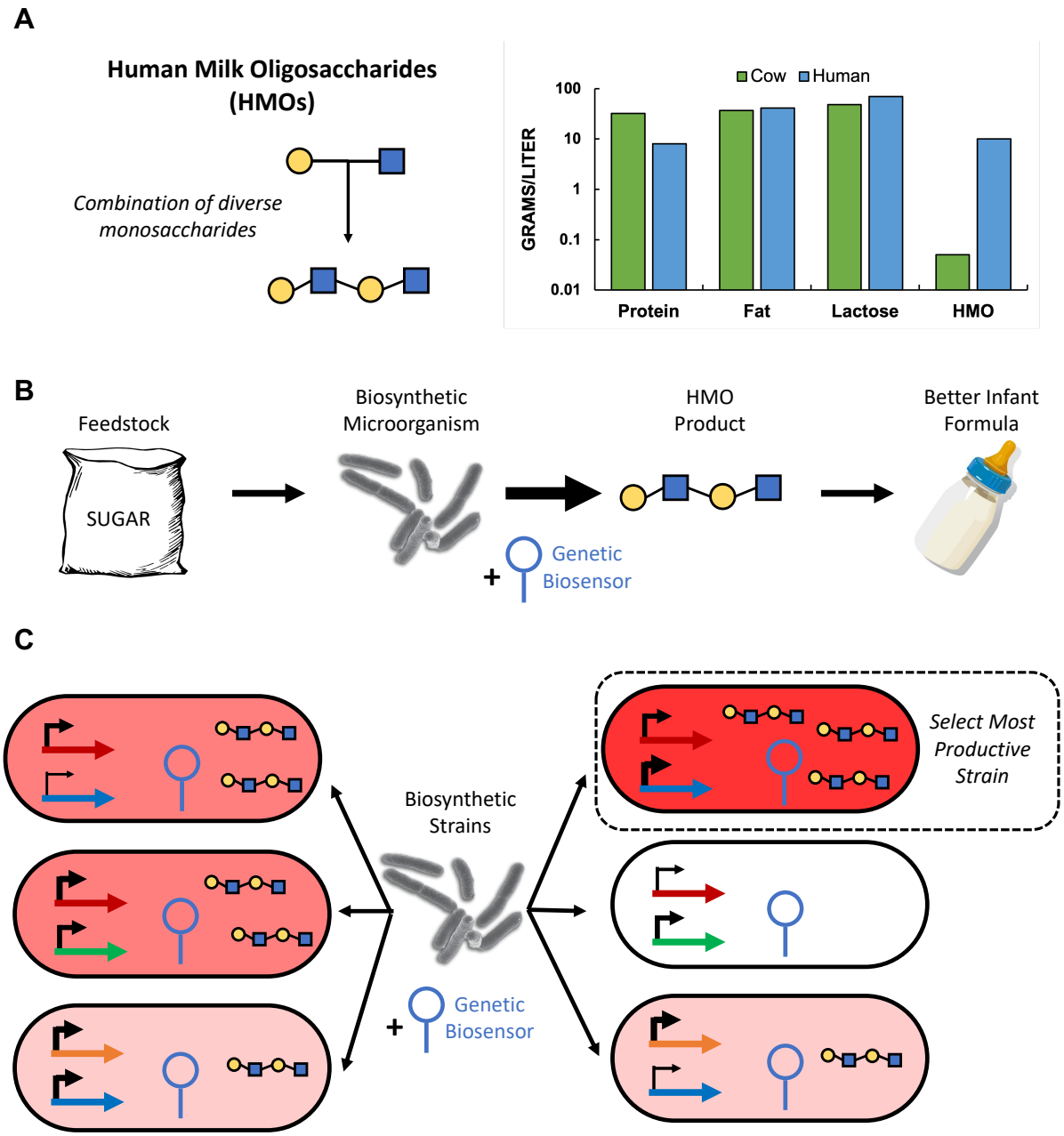


Figure 1. Human milk oligosaccharides represent an exciting target for microbial biosynthesis. A) Human milk oligosaccharides are a critical component of human breast milk and are largely absent from current cow's milk-based infant formula. B) One attractive route for the synthesis of HMOs is through biosynthesis using engineered microorganisms. C) By utilizing a genetically-encoded biosensor for our desired HMO it should be possible to rapidly identify the most productive biosynthetic pathway variants on the basis of cell color.

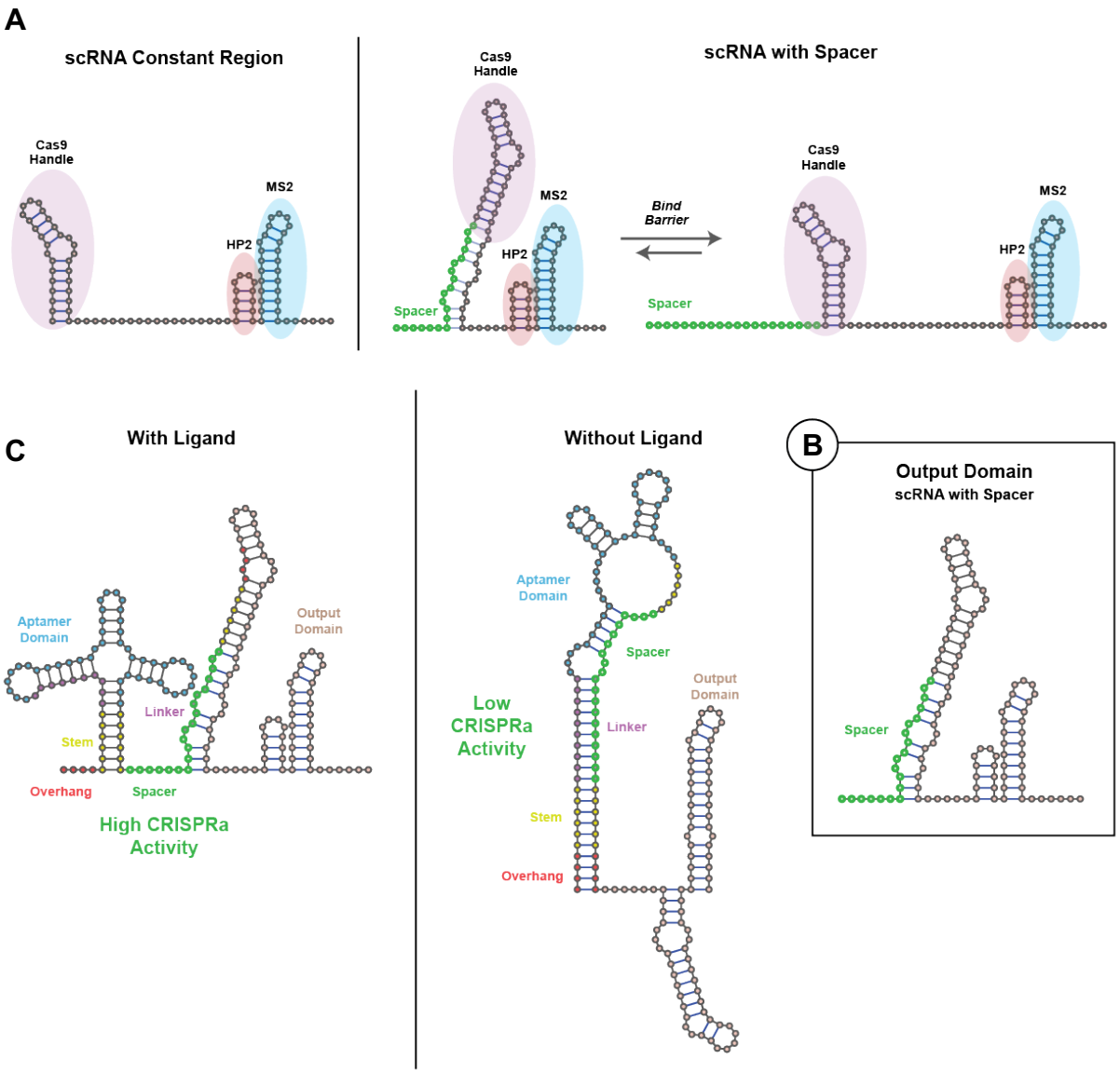


Figure 2. Engineering ligand-responsive scRNAs to regulate bacterial CRISPRa. A) scRNA activity depends on the ability of the scRNA to rapidly adopt a binding-competent conformation. B) The scRNA conformation corresponding to the absence of 5' sequence is used as an objective for structure-based switch screening. C) LR-scRNA candidates are designed such that in the absence of the target molecule the scRNA adopts a conformation incapable of binding Cas9. In the presence of the target molecule, the scRNA component is able to fold as it would in isolation.

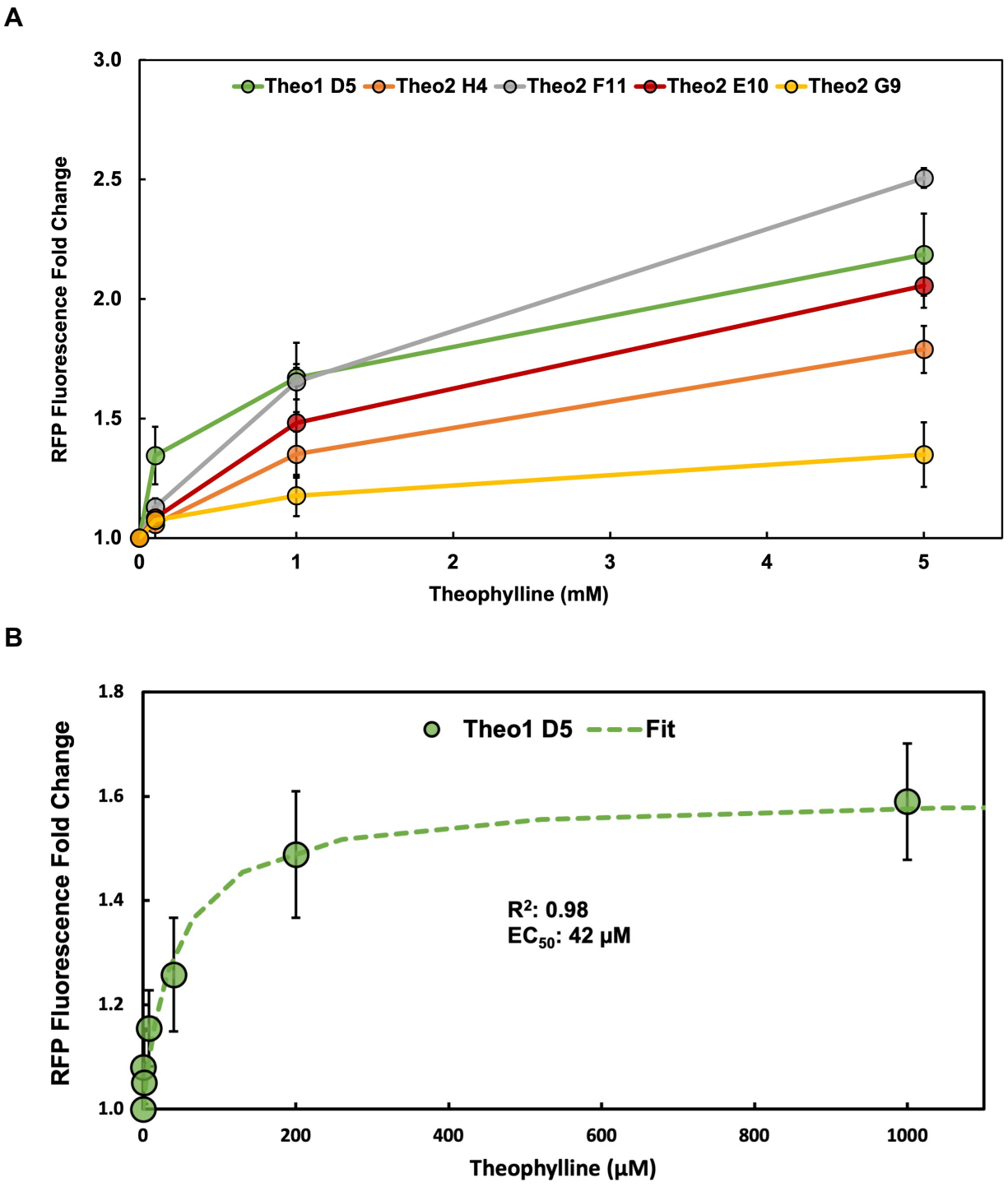


Figure 3. Engineered ligand-responsive scRNAs are able to respond to theophylline. A) Theophylline-responsive LR-scRNAs containing transcriptional pause sites show varied response to theophylline. B) Theo1 D5 shows displays the same low EC_{50} as previously engineered AS-RBS riboswitch constructs.

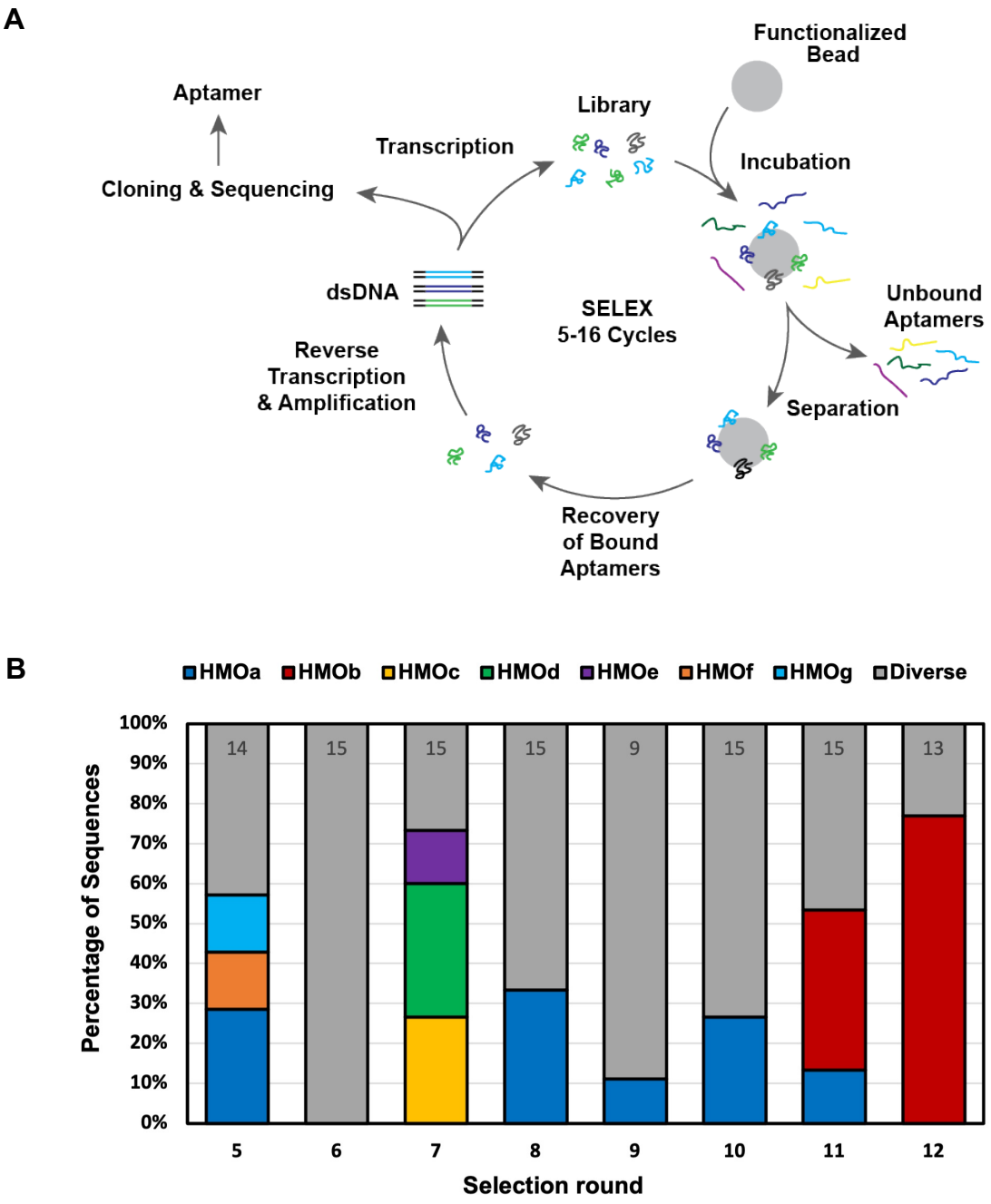


Figure 4. SELEX to identify a novel HMO-binding RNA aptamer. A) Overview of SELEX process for identifying novel RNA aptamers. B) Enrichment of candidate aptamer sequences by round. While HMOb enriches to become the majority of the DNA pool by round 12, other aptamer candidates enrich as early as round 5, before subsequently disappearing from the pool.

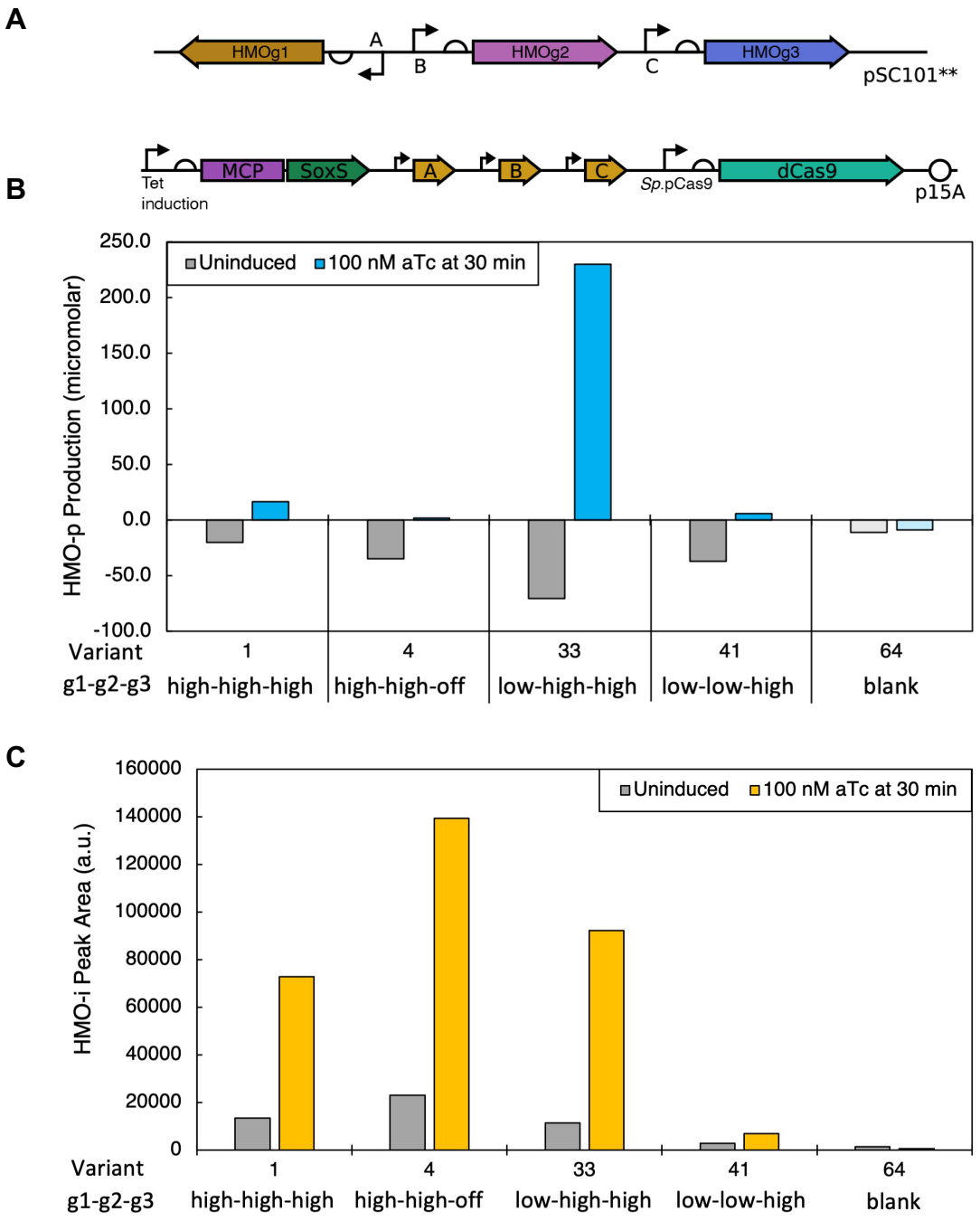


Figure 5. Engineered CRISPRa-regulated metabolic pathway produces our HMO product (HMO-p). A) Three different scRNAs are implemented to activate the three genes necessary to produce HMO-p, expressed from engineered CRISPRa promoters. B) Different promoter specific CRISPRa levels, implemented through truncations of the scRNA spacer sequence, give rise to different levels of HMO-p production. C) Levels of production of the intermediate HMO-i vary independently of HMO-p levels. When HMOg3 is not activated, HMO-i is not converted to HMO-p, and builds up.

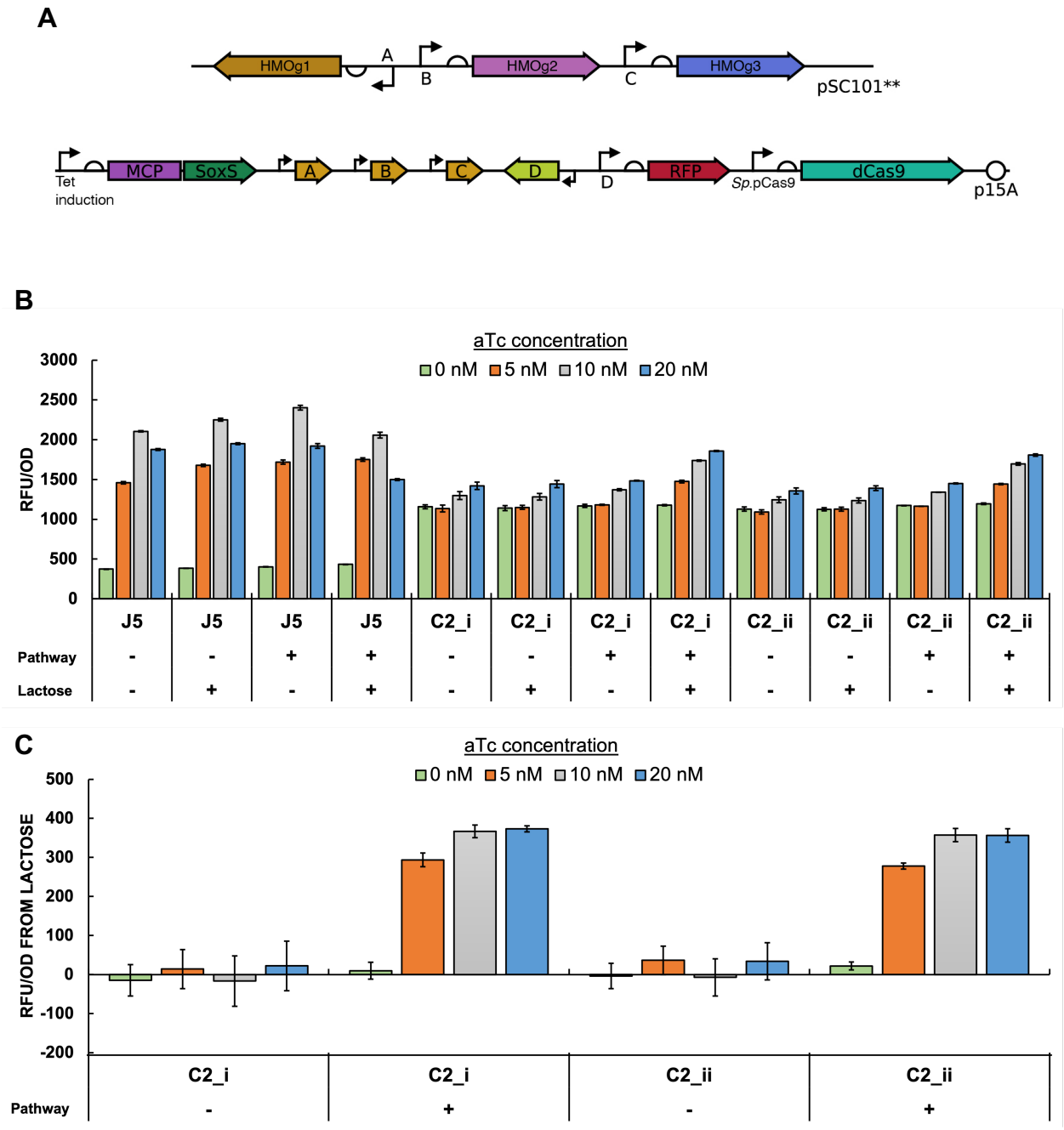


Figure 6. Using an HMO-responsive scRNA to sense production from an engineered metabolic pathway. A) An engineered HMO-responsive scRNA "D" is designed to activate transcription of mRFP1 in response to HMO production from CRISPRa-regulated metabolic pathway. B) Unlike a constitutively active scRNA (J5), candidate HMO-responsive scRNAs show increases in fluorescence when the pathway genes are present, and the cell is grown in media containing the substrate lactose. C) When subtracting the signal from the no-lactose constructs it becomes clear that the HMO-scRNA constructs result in increasing fluorescence as the pathway is induced at higher levels.

Supplementary Materials

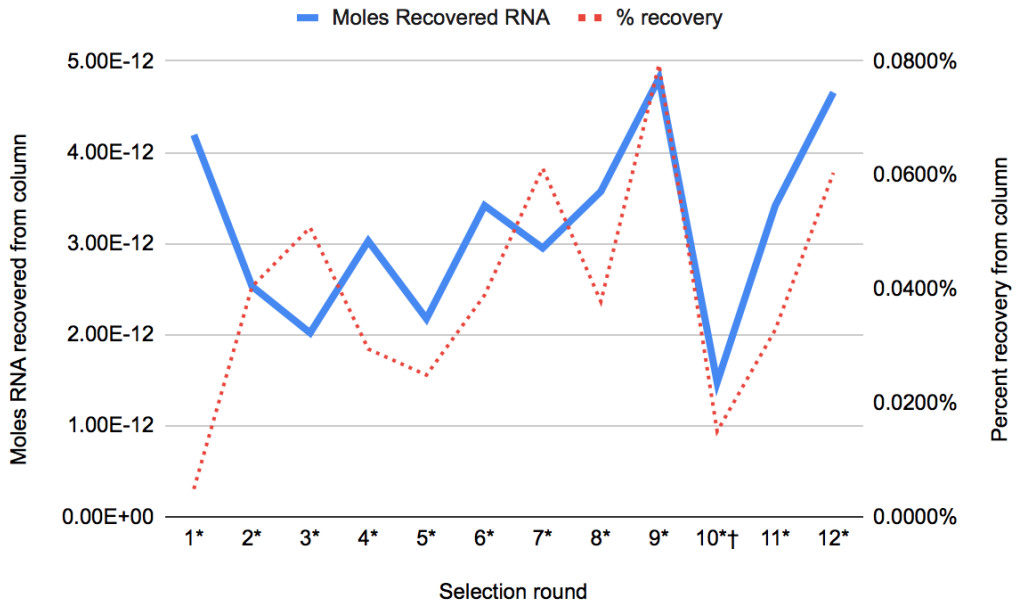


Figure S1. RNA recovery during the HMO aptamer selection. Neither total moles of RNA recovered after each round of selection, nor percent of RNA put on the column increased dramatically over the course of the selection, despite enrichment of specific sequences.

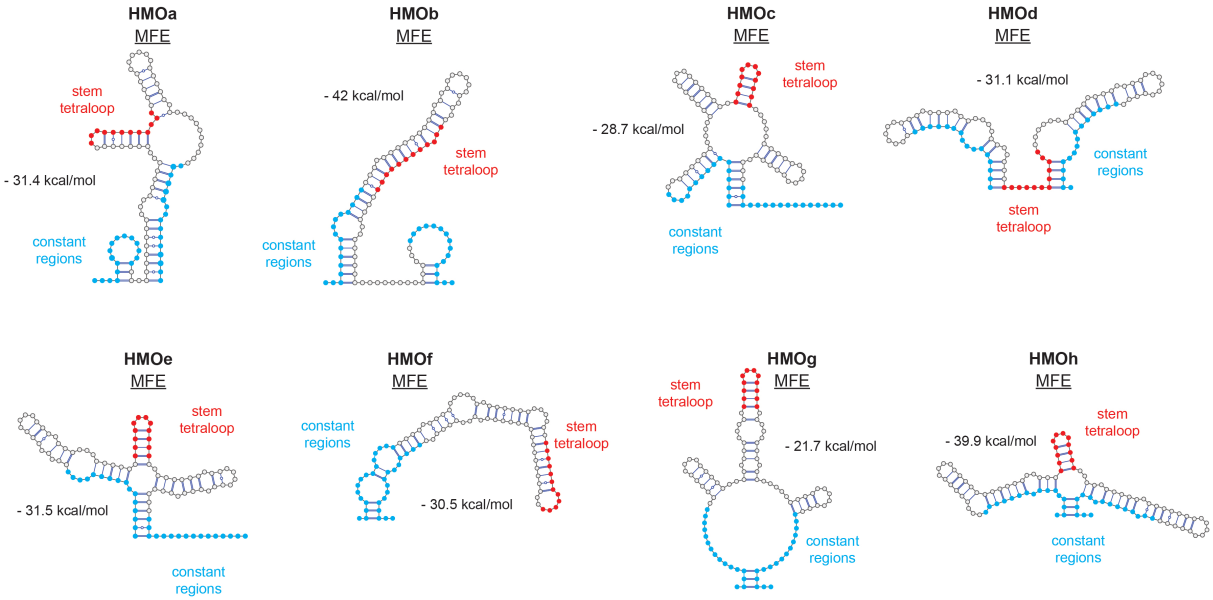


Figure S2. Structures of the candidate HMO-binding aptamers. HMOa-g represent the 7 unique sequences that enriched between rounds 5 and 12 during the selection. HMOh represents a sequence that transiently enriched during a previous failed attempt at the HMO aptamer selection.

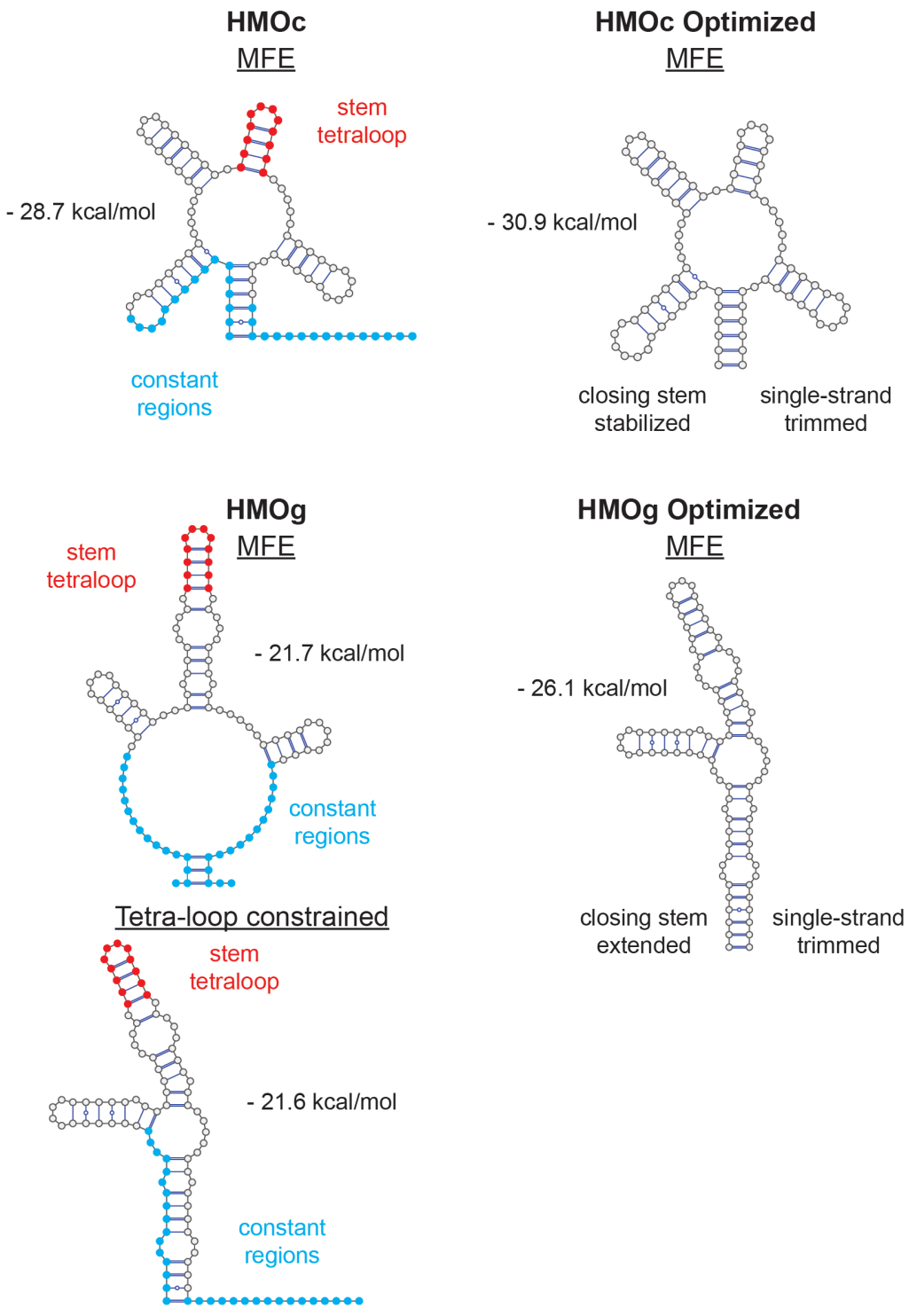


Figure S3. Structural optimization of aptamer candidates. Aptamer candidates were optimized to stabilize their predicted binding structures, and to prepare for input into the kinetic biosensor design pipeline.

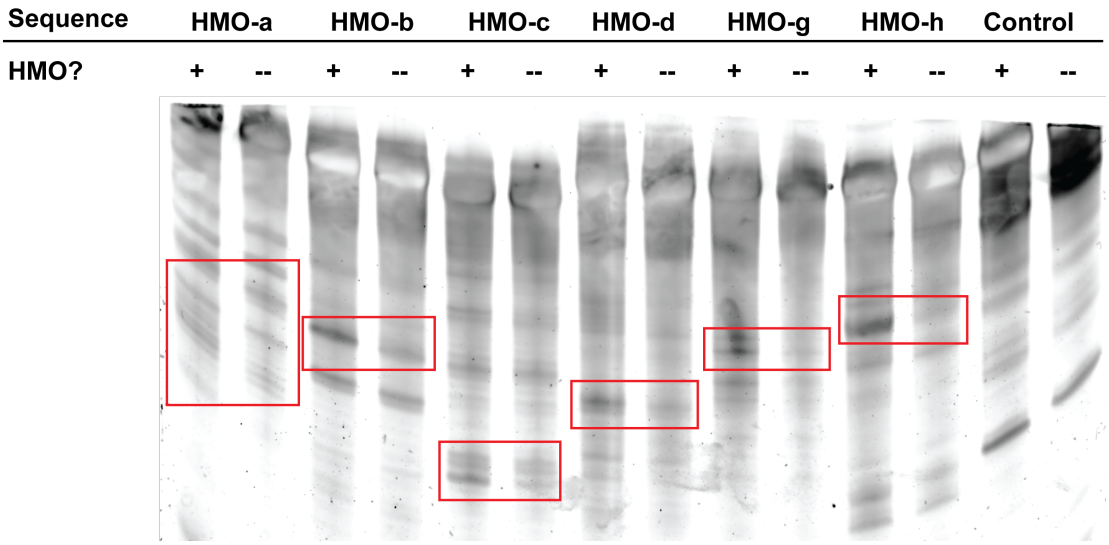


Figure S4. In-line probing to confirm HMO-binding of aptamer candidates. Changes in banding pattern due to structure-specific hydrolysis pattern indicate binding interaction with HMO-p. The lack of changes in band intensity of the control (pAF aptamer) suggests that the binding is a specific interaction.

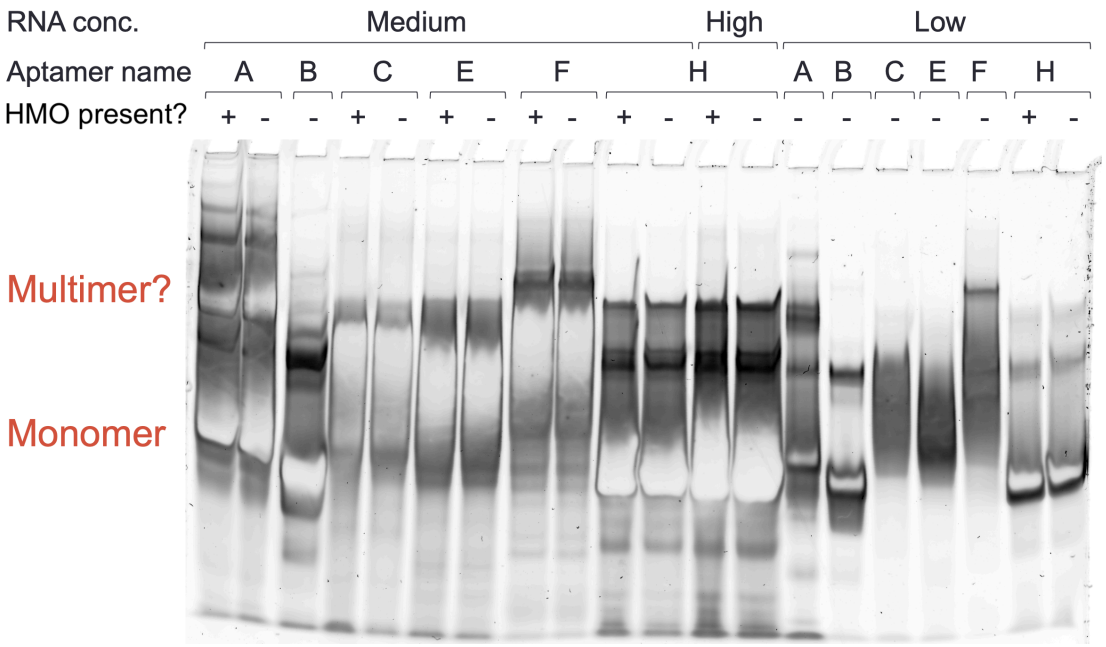


Figure S5. Analysis of aptamer candidate dimerization. To determine whether homo-dimerization was the cause for aptamer candidates to be removed from the pool after initial enrichment, purified aptamer candidates were analyzed on a native PAGE gel. Multiple species, likely resulting to higher-order assemblies, are present supporting this explanation.

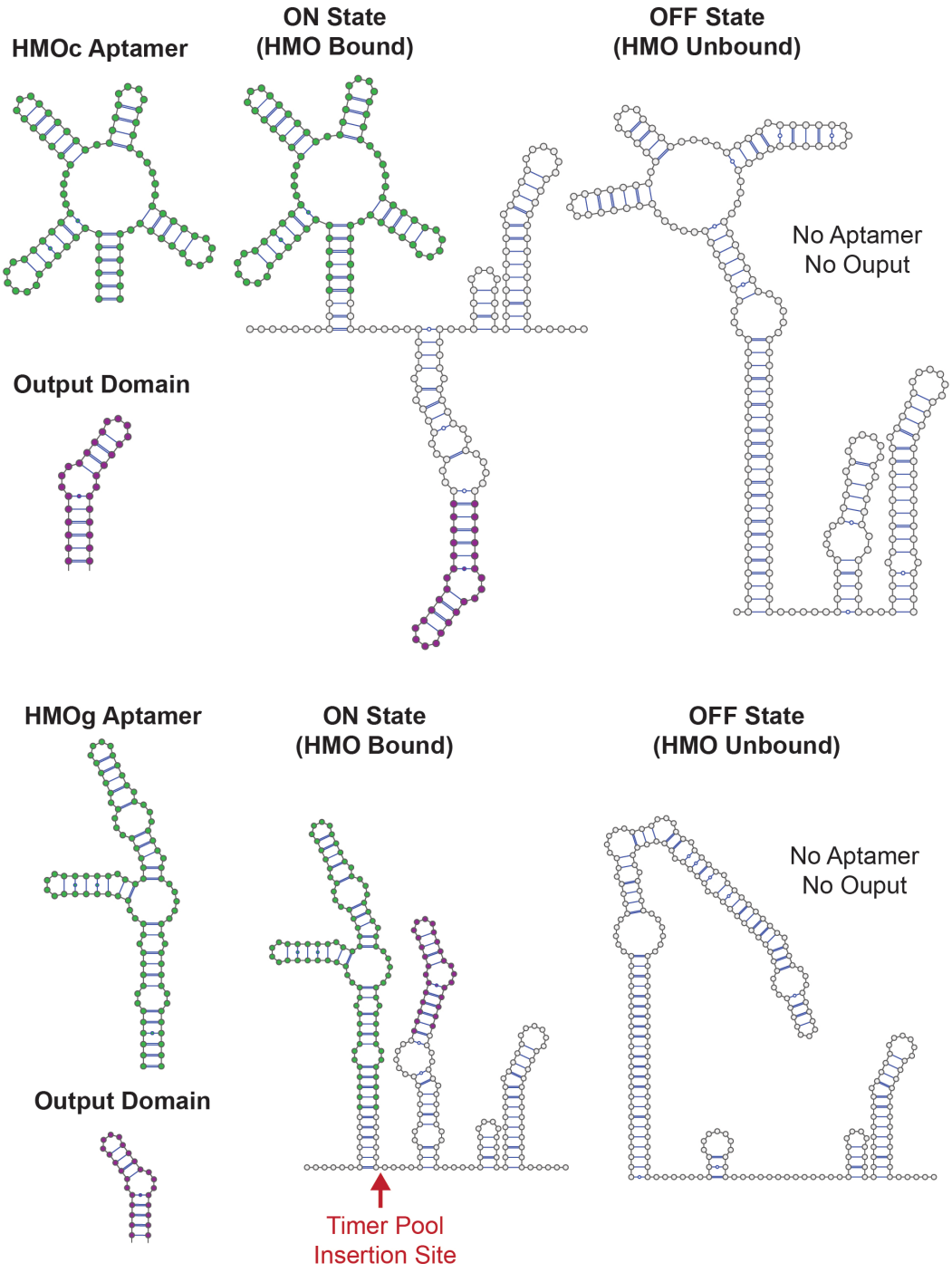


Figure S6. Structures of the HMO-responsive scRNA candidates. Aptamer candidates were input into the kinetic biosensor pipeline in order to design HMO-responsive scRNAs. In the presence of HMO-p the biosensors are designed to allow the rapid binding to dCas9. In the absence of HMO-p the biosensors are designed to rearrange the scRNA to a conformation that should inhibit binding to dCas9.

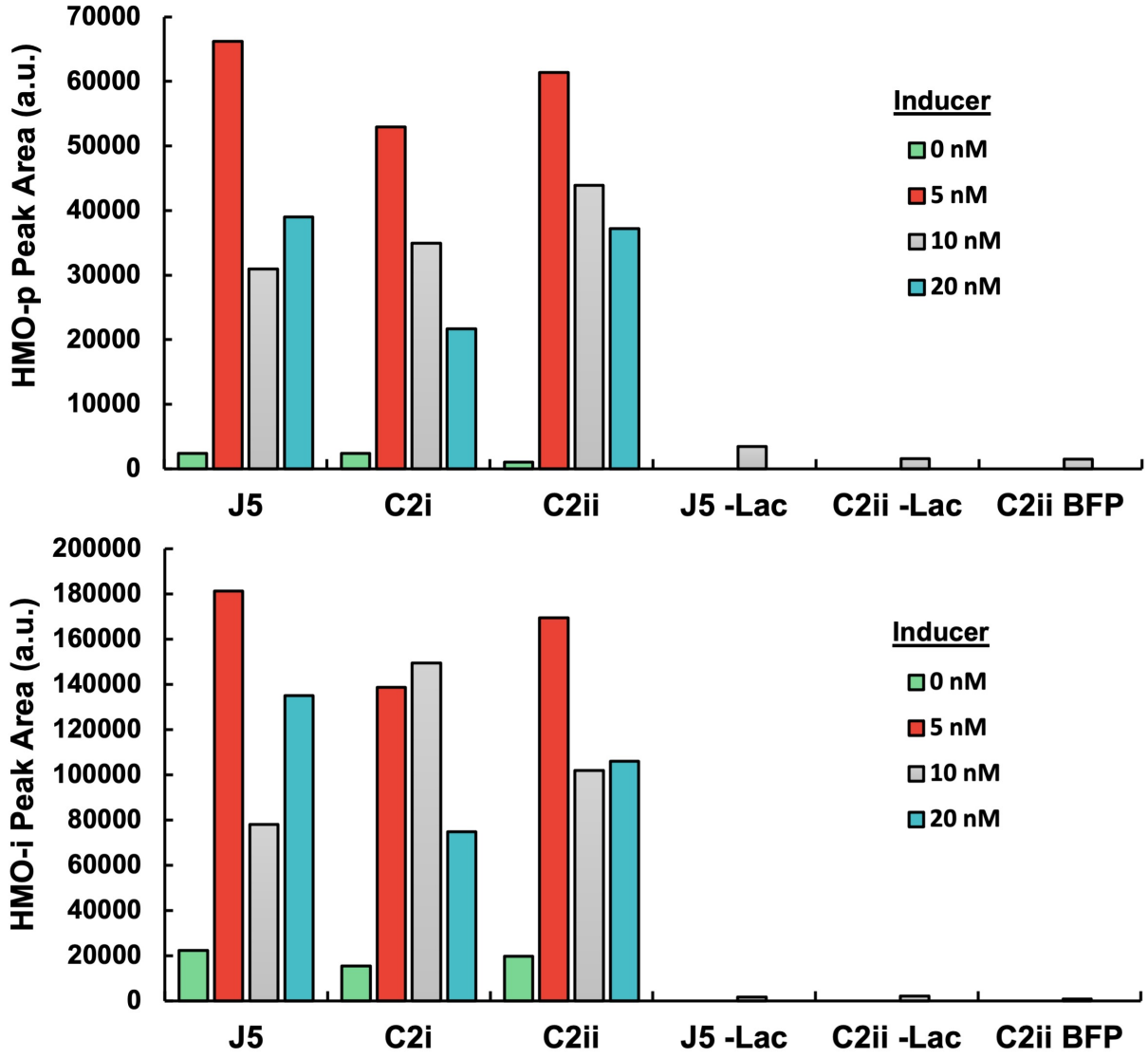


Figure S7. Experimental biosensor constructs still produce the HMO product. Experimental constructs containing HMO-responsive scRNA biosensors show inducer-dependent production of HMO-I and HMO-p. Control constructs lacking lactose in the media, or expressing BFP instead of the pathway enzymes display negligible production of either HMO-i or HMO-p. Values shown are average values from two biological replicates.

Acknowledgements

Attaining a Ph.D. is never an easy feat, however, the untimely passing of my mother and sister, during the first two years of my graduate studies, pushed me to my limits in a variety of ways, academic and beyond. It is not hyperbole to say that without the love, kindness, and support of a great number of friends, family, and colleagues I simply would not be writing this thesis at present. Although this is far from an exhaustive list of those who have contributed to the success of my graduate career, I would be remiss if I did not, at a minimum, thank the following people:

- Thank you, James Carothers, for giving me the opportunity to join your lab as a volunteer nearly 8 years ago, and for providing me with guidance, funding, and the flexibility to pursue difficult and exciting problems, even when exciting solutions were not always easy to come by.
- Thank you, Maya Bragg, for an endless supply of delicious food, design advice, emotional support, and open world video games to explore together. I couldn't ask for a better partner in grad school or in life.
- Thank you, Paul Yager, for all of your support throughout the last 31 years, as well as providing the spark that ignited my love for biomolecular engineering.
- Thank you, Alan Boyd, for pushing me to become the first freshly-minted doctor in the neighborhood.
- Thank you, Rachel Boccamazzo, for your friendship and guidance during the darkest time in my life. I truly would not have made it through graduate school without your support.
- Thank you, Teresa Sparkman, for being the most caring person I have ever met, and for showing me that community success is more important than personal accolades.
- Thank you, Sofie Sparkman-Yager, for teaching me to enjoy the little things and to not be ashamed to embrace what makes me unique.
- Thank you, Young Adult Group at The Healing Center, for providing an outlet to process my grief, and a community that made a very difficult time feel less isolating.
- Thank you, Rodrigo Correa, for showing me the ropes of molecular biology, and demonstrating that it's possible to love science while maintaining a healthy work-life balance.

- Thank you, Cassandra Burke, for saving me about a year of *in vitro* RNA headaches, and for showing me what real hard work looks like.
- Thank you, Jason Fontana, for being my scientific and entrepreneurial counterpoint, and for setting a standard of organization and professionalism that I aspire to match one day.
- Thank you, Willy Voje, for always asking me the hard questions and helping me to grow in confidence as a scientist and communicator.
- Thank you, Chuhern Hwang, for your even-keeled presence, and for experiencing the inevitable highs and lows that can come from an aptamer selection with me (as well as the highs and lows of sports fanaticism).
- Thank you, Ian Faulkner, for being a hardworking and ever-optimistic aptamer selection partner.
- Thank you, committee members, for your guidance through my graduate studies and throughout my thesis defense.
- Thank you, everyone not mentioned by name that has been a classmate, lab mate, professor, or friend, for allowing me to talk your ear off about RNA design and/or NBA basketball.

References

1. Lim, H. G., Jang, S., Jang, S., Seo, S. W. & Jung, G. Y. Design and optimization of genetically encoded biosensors for high-throughput screening of chemicals. *Curr. Opin. Biotechnol.* **54**, 18–25 (2018).
2. Henkin, T. M. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev.* **22**, 3383–3390 (2008).
3. Roth, A., Welz, R. & Breaker, R. R. Riboswitches: Natural Metabolite-binding RNAs Controlling Gene Expression. in *The Aptamer Handbook* (ed. Klussmann, S.) 191–207 (Wiley-VCH Verlag GmbH & Co. KGaA, 2006).
4. Pavlova, N., Kaloudas, D. & Penchovsky, R. Riboswitch distribution, structure, and function in bacteria. *Gene* **708**, 38–48 (2019).
5. Pardee, K. *et al.* Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell* **165**, 1255–1266 (2016).
6. 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).
7. Win, M. N. & Smolke, C. D. Higher-Order Cellular Information Processing with Synthetic RNA Devices. *Science* **322**, 456–460 (2008).
8. Townshend, B., Kennedy, A. B., Xiang, J. S. & Smolke, C. D. High-throughput cellular RNA device engineering. *Nat. Methods* **12**, 989–994 (2015).
9. Goler, J. A., Carothers, J. M. & Keasling, J. D. Dual-Selection for Evolution of In Vivo Functional Aptazymes as Riboswitch Parts. in *Artificial Riboswitches* 221–235 (Humana Press, Totowa, NJ, 2014). doi:10.1007/978-1-62703-755-6_16.
10. Hopfield, J. J. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. U. S. A.* **71**, 4135–4139 (1974).
11. James, L. C. & Tawfik, D. S. Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 12730–12735 (2005).

12. Hartich, D., Barato, A. C. & Seifert, U. Nonequilibrium sensing and its analogy to kinetic proofreading. *New J. Phys.* **17**, 055026 (2015).
13. Guedich, S. *et al.* Quantitative and predictive model of kinetic regulation by E. coli TPP riboswitches. *RNA Biol.* **13**, 373–390 (2016).
14. Wickiser, J. K., Winkler, W. C., Breaker, R. R. & Crothers, D. M. The Speed of RNA Transcription and Metabolite Binding Kinetics Operate an FMN Riboswitch. *Mol. Cell* **18**, 49–60 (2005).
15. Espah Borujeni, A., Mishler, D. M., Wang, J., Huso, W. & Salis, H. M. Automated physics-based design of synthetic riboswitches from diverse RNA aptamers. *Nucleic Acids Res.* **44**, 1–13 (2016).
16. Wolfe, B. R., Porubsky, N. J., Zadeh, J. N., Dirks, R. M. & Pierce, N. A. Constrained Multistate Sequence Design for Nucleic Acid Reaction Pathway Engineering. *J. Am. Chem. Soc.* **139**, 3134–3144 (2017).
17. Rodrigo, G., Landrain, T. E., Majer, E., Daròs, J.-A. & Jaramillo, A. Full Design Automation of Multi-State RNA Devices to Program Gene Expression Using Energy-Based Optimization. *PLoS Comput. Biol.* **9**, (2013).
18. Mishler, D. M. & Gallivan, J. P. A family of synthetic riboswitches adopts a kinetic trapping mechanism. *Nucleic Acids Res.* **42**, 6753–6761 (2014).
19. Klauser, B., Atanasov, J., Siewert, L. K. & Hartig, J. S. Ribozyme-Based Aminoglycoside Switches of Gene Expression Engineered by Genetic Selection in S. cerevisiae. *ACS Synth. Biol.* **4**, 516–525 (2015).
20. Ketzer, P. *et al.* Artificial riboswitches for gene expression and replication control of DNA and RNA viruses. *Proc. Natl. Acad. Sci.* **111**, E554–E562 (2014).
21. Nomura, Y., Zhou, L., Miu, A. & Yokobayashi, Y. Controlling Mammalian Gene Expression by Allosteric Hepatitis Delta Virus Ribozymes. *ACS Synth. Biol.* **2**, 684–689 (2013).
22. Sparkman-Yager, D., Correa-Rojas, R. A. & Carothers, J. M. Chapter Sixteen - Kinetic Folding Design of Aptazyme-Regulated Expression Devices as Riboswitches for Metabolic Engineering. in *Methods in Enzymology* (ed. Burke-Aguero, D. H.) vol. 550 321–340 (Academic Press, 2015).
23. Breaker, R. R. *et al.* A common speed limit for RNA-cleaving ribozymes and deoxyribozymes. *RNA* **9**, 949–957 (2003).

24. Emilsson, G. M., Nakamura, S., Roth, A. & Breaker, R. R. Ribozyme speed limits. *RNA* **9**, 907–918 (2003).
25. Chen, X. & Ellington, A. D. Design Principles for Ligand-Sensing, Conformation-Switching Ribozymes. *PLOS Comput. Biol.* **5**, e1000620 (2009).
26. Beisel, C. L. & Smolke, C. D. Design Principles for Riboswitch Function. *PLOS Comput. Biol.* **5**, e1000363 (2009).
27. Khvorova, A., Lescoute, A., Westhof, E. & Jayasena, S. D. Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. *Nat. Struct. Mol. Biol.* **10**, 708–712 (2003).
28. Penedo, J. C., Wilson, T. J., Jayasena, S. D., Khvorova, A. & Lilley, D. M. J. Folding of the natural hammerhead ribozyme is enhanced by interaction of auxiliary elements. *RNA* **10**, 880–888 (2004).
29. Nomura, Y., Chien, H.-C. & Yokobayashi, Y. Direct screening for ribozyme activity in mammalian cells. *Chem. Commun.* **53**, 12540–12543 (2017).
30. Penchovsky, R. Computational Design and Biosensor Applications of Small Molecule-Sensing Allosteric Ribozymes. *Biomacromolecules* **14**, 1240–1249 (2013).
31. Zhong, G., Wang, H., Bailey, C. C., Gao, G. & Farzan, M. Rational design of aptazyme riboswitches for efficient control of gene expression in mammalian cells. *eLife* **5**,.
32. Link, K. H. *et al.* Engineering high-speed allosteric hammerhead ribozymes. *Biol. Chem.* **388**, 779–786 (2007).
33. Lai, D., Proctor, J. R. & Meyer, I. M. On the importance of cotranscriptional RNA structure formation. *RNA* **19**, 1461–1473 (2013).
34. Watters, K. E., Strobel, E. J., Yu, A. M., Lis, J. T. & Lucks, J. B. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nat. Struct. Mol. Biol.* **23**, 1124–1131 (2016).
35. Herschlag, D. RNA Chaperones and the RNA Folding Problem. *J. Biol. Chem.* **270**, 20871–20874 (1995).
36. Treiber, D. K. & Williamson, J. R. Exposing the kinetic traps in RNA folding. *Curr. Opin. Struct. Biol.* **9**, 339–345 (1999).

37. Xayaphoummine, A., Bucher, T. & Isambert, H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.* **33**, W605–W610 (2005).
38. Proctor, J. R. & Meyer, I. M. C o F old : an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Res.* **41**, e102–e102 (2013).
39. Geis, M. *et al.* Folding Kinetics of Large RNAs. *J. Mol. Biol.* **379**, 160–173 (2008).
40. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* https://www.nature.com/articles/nmeth.2089 (2012) doi:10.1038/nmeth.2089.
41. Zimmermann, G. R., Wick, C. L., Shields, T. P., Jenison, R. D. & Pardi, A. Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer. *RNA* **6**, 659–667 (2000).
42. Carothers, J. M., Goler, J. A., Kapoor, Y., Lara, L. & Keasling, J. D. Selecting RNA aptamers for synthetic biology: investigating magnesium dependence and predicting binding affinity. *Nucleic Acids Res.* **38**, 2736–2747 (2010).
43. Long, D. M. & Uhlenbeck, O. C. Kinetic characterization of intramolecular and intermolecular hammerhead RNAs with stem II deletions. *Proc. Natl. Acad. Sci.* **91**, 6977–6981 (1994).
44. Brion, P. & Westhof, and E. Hierarchy and Dynamics of Rna Folding. *Annu. Rev. Biophys. Biomol. Struct.* **26**, 113–137 (1997).
45. Greenleaf, W. J., Frieda, K. L., Foster, D. A. N., Woodside, M. T. & Block, S. M. Direct Observation of Hierarchical Folding in Single Riboswitch Aptamers. *Science* **319**, 630–633 (2008).
46. Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L. & Stadler, P. F. Efficient computation of RNA folding dynamics. *J. Phys. Math. Gen.* **37**, 4731 (2004).
47. Danilova, L. V., Pervouchine, D. D., Favorov, A. V. & Mironov, A. A. Rnakinetics: a web server that models secondary structure kinetics of an elongating rna. *J. Bioinform. Comput. Biol.* **04**, 589–596 (2006).
48. Golomb, M. & Chamberlin, M. Characterization of T7-specific Ribonucleic Acid Polymerase IV. RESOLUTION OF THE MAJOR IN VITRO TRANSCRIPTS BY GEL ELECTROPHORESIS. *J. Biol. Chem.* **249**, 2858–2863 (1974).
49. Silva, C. de & Walter, N. G. Leakage and slow allostery limit performance of single drug-sensing aptazyme molecules based on the hammerhead ribozyme. *RNA* **15**, 76–84 (2009).

50. Zhang, D. Y. & Seelig, G. Dynamic DNA nanotechnology using strand-displacement reactions. *Nat. Chem.* **3**, 103 (2011).
51. Green, A. A., Silver, P. A., Collins, J. J. & Yin, P. Toehold Switches: De-Novo-Designed Regulators of Gene Expression. *Cell* **159**, 925–939 (2014).
52. Pardee, K. *et al.* Paper-Based Synthetic Gene Networks. *Cell* **159**, 940–954 (2014).
53. Green, A. A. *et al.* Complex cellular logic computation using ribocomputing devices. *Nature* **548**, 117–121 (2017).
54. Srinivas, N. *et al.* On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Res.* **41**, 10641–10658 (2013).
55. Šulc, P., Ouldridge, T. E., Romano, F., Doye, J. P. K. & Louis, A. A. Modelling Toehold-Mediated RNA Strand Displacement. *Biophys. J.* **108**, 1238–1247 (2015).
56. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
57. Purzycka, K. J. *et al.* Chapter One - Automated 3D RNA Structure Prediction Using the RNAComposer Method for Riboswitches1. in *Methods in Enzymology* (eds. Chen, S.-J. & Burke-Aguero, D. H.) vol. 553 3–34 (Academic Press, 2015).
58. Dotu, I., Lorenz, W. A., Van Hentenryck, P. & Clote, P. Computing folding pathways between RNA secondary structures. *Nucleic Acids Res.* **38**, 1711–1722 (2010).
59. Li, Y. & Zhang, S. Predicting folding pathways between RNA conformational structures guided by RNA stacks. *BMC Bioinformatics* **13**, S5 (2012).
60. Kim, P. B., Nelson, J. W. & Breaker, R. R. An Ancient Riboswitch Class in Bacteria Regulates Purine Biosynthesis and One-Carbon Metabolism. *Mol. Cell* **57**, 317–328 (2015).
61. Deana, A., Celesnik, H. & Belasco, J. G. The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. *Nature* **451**, 355–358 (2008).
62. Andreasson, J. O. L., Savinov, A., Block, S. M. & Greenleaf, W. J. Comprehensive sequence-to-function mapping of cofactor-dependent RNA catalysis in the glmS ribozyme. *Nat. Commun.* **11**, 1663 (2020).
63. Breaker, R. R. Riboswitches and Translation Control. *Cold Spring Harb. Perspect. Biol.* **10**, a032797 (2018).

64. Salis, H. M. Chapter two - The Ribosome Binding Site Calculator. in *Methods in Enzymology* (ed. Voigt, C.) vol. 498 19–42 (Academic Press, 2011).
65. Wrist, A., Sun, W. & Summers, R. M. The Theophylline Aptamer: 25 Years as an Important Tool in Cellular Engineering Research. *ACS Synth. Biol.* (2020) doi:10.1021/acssynbio.9b00475.
66. Larson, M. H. *et al.* A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **344**, 1042–1047 (2014).
67. Kang, J. Y., Mishanina, T. V., Landick, R. & Darst, S. A. Mechanisms of Transcriptional Pausing in Bacteria. *J. Mol. Biol.* (2019) doi:10.1016/j.jmb.2019.07.017.
68. Chauvier, A. *et al.* Transcriptional pausing at the translation start site operates as a critical checkpoint for riboswitch regulation. *Nat. Commun.* **8**, 13892 (2017).
69. Chauvier, A., Nadon, J.-F., Grondin, J. P., Lamontagne, A.-M. & Lafontaine, D. A. Role of a hairpin-stabilized pause in the Escherichia coli thiC riboswitch function. *RNA Biol.* **16**, 1066–1073 (2019).
70. Ceres, P., Garst, A. D., Marcano-Velázquez, J. G. & Batey, R. T. Modularity of Select Riboswitch Expression Platforms Enables Facile Engineering of Novel Genetic Regulatory Devices. https://pubs.acs.org/doi/pdf/10.1021/sb4000096 (2013) doi:10.1021/sb4000096.
71. Wieland, M. & Hartig, J. S. Improved Aptazyme Design and In Vivo Screening Enable Riboswitching in Bacteria. *Angew. Chem. Int. Ed.* **47**, 2604–2607 (2008).
72. Feng, X., Liu, L., Duan, X. & Wang, S. An engineered riboswitch as a potential gene-regulatory platform for reducing antibacterial drug resistance. *Chem. Commun.* **47**, 173–175 (2011).
73. Ogawa, A. & Maeda, M. An Artificial Aptazyme-Based Riboswitch and its Cascading System in E. coli. *ChemBioChem* **9**, 206–209 (2008).
74. Suess, B., Fink, B., Berens, C., Stentz, R. & Hillen, W. A theophylline responsive riboswitch based on helix slipping controls gene expression in vivo. *Nucleic Acids Res.* **32**, 1610–1614 (2004).
75. Wachsmuth, M., Findeiß, S., Weissheimer, N., Stadler, P. F. & Mörl, M. De novo design of a synthetic riboswitch that regulates transcription termination. *Nucleic Acids Res.* **41**, 2541–2551 (2013).
76. Goodman, D. B., Church, G. M. & Kosuri, S. Causes and Effects of N-Terminal Codon Bias in Bacterial Genes. *Science* **342**, 475–479 (2013).
77. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164**, 1185–1197 (2016).

78. Lee, S. Y. *et al.* A comprehensive metabolic map for production of bio-based chemicals. *Nat. Catal.* **2**, 18–33 (2019).
79. Fontana, J., Voje, W. E., Zalatan, J. G. & Carothers, J. M. Prospects for engineering dynamic CRISPR–Cas transcriptional circuits to improve bioproduction. *J. Ind. Microbiol. Biotechnol.* **45**, 481–490 (2018).
80. Choi, K. R. *et al.* Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering. *Trends Biotechnol.* **37**, 817–837 (2019).
81. Liu, R., Bassalo, M. C., Zeitoun, R. I. & Gill, R. T. Genome scale engineering techniques for metabolic engineering. *Metab. Eng.* **32**, 143–154 (2015).
82. Carbonell, P. *et al.* An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Commun. Biol.* **1**, 1–10 (2018).
83. Brunk, E. *et al.* Characterizing Strain Variation in Engineered E. coli Using a Multi-Omics-Based Workflow. *Cell Syst.* **2**, 335–346 (2016).
84. Wang, H., La Russa, M. & Qi, L. S. CRISPR/Cas9 in Genome Editing and Beyond. *Annu. Rev. Biochem.* **85**, 227–264 (2016).
85. Dominguez, A. A., Lim, W. A. & Qi, L. S. Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. *Nat. Rev. Mol. Cell Biol.* **17**, 5–15 (2016).
86. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152**, 1173–1183 (2013).
87. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–7437 (2013).
88. Dong, C., Fontana, J., Patel, A., Carothers, J. M. & Zalatan, J. G. Synthetic CRISPR-Cas gene activators for transcriptional reprogramming in bacteria. *Nat. Commun.* **9**, 2489 (2018).
89. Liu, Y., Wan, X. & Wang, B. Engineered CRISPRa enables programmable eukaryote-like gene activation in bacteria. *Nat. Commun.* **10**, 3693 (2019).
90. Ho, H.-I., Fang, J., Cheung, J. & Wang, H. H. Programmable and portable CRISPR-Cas transcriptional activation in bacteria. *bioRxiv* 2020.01.03.882431 (2020) doi:10.1101/2020.01.03.882431.

91. Kundert, K. *et al.* Controlling CRISPR-Cas9 with ligand-activated and ligand-deactivated sgRNAs. *Nat. Commun.* **10**, 1–11 (2019).

92. Tang, W., Hu, J. H. & Liu, D. R. Aptazyme-embedded guide RNAs enable ligand-responsive genome editing and transcriptional activation. *Nat. Commun.* **8**, 15939 (2017).

93. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

94. Zalatan, J. G. *et al.* Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds. *Cell* **160**, 339–350 (2015).

95. Mali, P., Esvelt, K. M. & Church, G. M. Cas9 as a versatile tool for engineering biology. *Nat. Methods* **10**, 957–963 (2013).

96. Niu, F.-X., Huang, Y.-B., Ji, L.-N. & Liu, J.-Z. Genomic and transcriptional changes in response to pinene tolerance and overproduction in evolved Escherichia coli. *Synth. Syst. Biotechnol.* **4**, 113–119 (2019).

97. Otoupal, P. B., Erickson, K. E., Escalas-Bordoy, A. & Chatterjee, A. CRISPR Perturbation of Gene Expression Alters Bacterial Fitness under Stress and Reveals Underlying Epistatic Constraints. *ACS Synth. Biol.* **6**, 94–107 (2017).

98. Lu, Z. *et al.* CRISPR-assisted multi-dimensional regulation for fine-tuning gene expression in Bacillus subtilis. *Nucleic Acids Res.* **47**, e40–e40 (2019).

99. Yu, L., Su, W., Fey, P. D., Liu, F. & Du, L. Yield Improvement of the Anti-MRSA Antibiotics WAP-8294A by CRISPR/dCas9 Combined with Refactoring Self-Protection Genes in Lysobacter enzymogenes OH11. *ACS Synth. Biol.* **7**, 258–266 (2018).

100. Peng, R. *et al.* CRISPR/dCas9-mediated transcriptional improvement of the biosynthetic gene cluster for the epothilone production in Myxococcus xanthus. *Microb. Cell Factories* **17**, 15 (2018).

101. Fontana, J. *et al.* Effective CRISPRa-Mediated Control of Gene Expression in Bacteria Must Overcome Strict Target Site Requirements. *bioRxiv* 770891 (2019) doi:10.1101/770891.

102. Dove, S. L. & Hochschild, A. Conversion of the ω subunit of Escherichia coli RNA polymerase into a transcriptional activator or an activation target. *Genes Dev.* **12**, 745–754 (1998).

103. Tian, T., Kang, J. W., Kang, A. & Lee, T. S. Redirecting Metabolic Flux via Combinatorial Multiplex CRISPRi-Mediated Repression for Isopentenol Production in Escherichia coli. *ACS Synth. Biol.* **8**, 391–402 (2019).

104. Dinh, C. V. & Prather, K. L. J. Development of an autonomous and bifunctional quorum-sensing circuit for metabolic flux control in engineered *Escherichia coli*. *Proc. Natl. Acad. Sci.* **116**, 25562–25568 (2019).

105. Johnny H. Hu, D. R. L., Shannon M. Miller, Maarten H. Geurts, Weixin Tang, Liwei Chen, Ning Sun, Christina M. Zeina, Xue Gao, Holly A. Rees, Zhi Lin. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).

106. Chatterjee, P., Jakimo, N. & Jacobson, J. M. Robust Genome Editing of Single-Base PAM Targets with Engineered ScCas9 Variants. *bioRxiv* 620351 (2019) doi:10.1101/620351.

107. Martella, A. *et al.* Systematic Evaluation of CRISPRa and CRISPRi Modalities Enables Development of a Multiplexed, Orthogonal Gene Activation and Repression System. *ACS Synth. Biol.* **8**, 1998–2006 (2019).

108. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

109. Shechner, D. M., Hacisuleyman, E., Younger, S. T. & Rinn, J. L. Multiplexable, locus-specific targeting of long RNAs with CRISPR-Display. *Nat. Methods* **12**, 664–670 (2015).

110. Cheng, A. W. *et al.* Casilio: a versatile CRISPR-Cas9-Pumilio hybrid for gene regulation and genomic labeling. *Cell Res.* **26**, 254–257 (2016).

111. Briner, A. E. *et al.* Guide RNA Functional Modules Direct Cas9 Activity and Orthogonality. *Mol. Cell* **56**, 333–339 (2014).

112. Moreno-Mateos, M. A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).

113. Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 148 (2016).

114. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

115. Thyme, S. B., Akhmetova, L., Montague, T. G., Valen, E. & Schier, A. F. Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.* **7**, 1–7 (2016).

116. Smith, J. D. *et al.* A method for high-throughput production of sequence-verified DNA libraries and strain collections. *Mol. Syst. Biol.* **13**, 913 (2017).

117. Vigouroux, A., Oldewurtel, E., Cui, L., Teeffelen, S. van & Bikard, D. Engineered CRISPR-Cas9 system enables noiseless, fine-tuned and multiplexed repression of bacterial genes. *bioRxiv* 164384 (2017) doi:10.1101/164384.

118. Cui, L. *et al.* A CRISPRi screen in E. coli reveals sequence-specific toxicity of dCas9. *Nat. Commun.* **9**, 1–10 (2018).

119. Kocak, D. D. *et al.* Increasing the specificity of CRISPR systems with engineered RNA secondary structures. *Nat. Biotechnol.* **37**, 657–666 (2019).

120. Jin, M., Garreau de Loubresse, N., Kim, Y., Kim, J. & Yin, P. Programmable CRISPR-Cas Repression, Activation, and Computation with Sequence-Independent Targets and Triggers. *ACS Synth. Biol.* **8**, 1583–1589 (2019).

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

122. Siu, K.-H. & Chen, W. Riboregulated toehold-gated gRNA for programmable CRISPR–Cas9 function. *Nat. Chem. Biol.* **15**, 217–220 (2019).

123. Li, Y., Teng, X., Zhang, K., Deng, R. & Li, J. RNA Strand Displacement Responsive CRISPR/Cas9 System for mRNA Sensing. *Anal. Chem.* **91**, 3989–3996 (2019).

124. Oesinghaus, L. & Simmel, F. C. Switching the activity of Cas12a using guide RNA strand displacement circuits. *Nat. Commun.* **10**, 2092 (2019).

125. Hanewich-Hollatz, M. H., Chen, Z., Hochrein, L. M., Huang, J. & Pierce, N. A. Conditional Guide RNAs: Programmable Conditional Regulation of CRISPR/Cas Function in Bacterial and Mammalian Cells via Dynamic RNA Nanotechnology. *ACS Cent. Sci.* **5**, 1241–1249 (2019).

126. Abatemarco, J. *et al.* RNA-aptamers-in-droplets (RAPID) high-throughput screening for secretory phenotypes. *Nat. Commun.* **8**, 1–9 (2017).

127. Porter, E. B., Polaski, J. T., Morck, M. M. & Batey, R. T. Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors. *Nat. Chem. Biol.* **13**, 295–301 (2017).

128. Liu, Y. *et al.* Directing cellular information flow via CRISPR signal conductors. *Nat. Methods* **13**, 938–944 (2016).

129. Lin, B. *et al.* Control of CRISPR-Cas9 with small molecule-activated allosteric aptamer regulating sgRNAs. *Chem. Commun.* **55**, 12223–12226 (2019).

130. Ricci, F., Vallée-Bélisle, A., Simon, A. J., Porchetta, A. & Plaxco, K. W. Using Nature's "Tricks" To Rationally Tune the Binding Properties of Biomolecular Receptors. *Acc. Chem. Res.* **49**, 1884–1892 (2016).

131. Chen, X. & Ellington, A. D. Design Principles for Ligand-Sensing, Conformation-Switching Ribozymes. *PLOS Comput. Biol.* **5**, e1000620 (2009).

132. Guedich, S. *et al.* Quantitative and predictive model of kinetic regulation by E. coli TPP riboswitches. *RNA Biol.* **13**, 373–390 (2016).

133. Leavell, M. D., Singh, A. H. & Kaufmann-Malaga, B. B. High-throughput screening for improved microbial cell factories, perspective and promise. *Curr. Opin. Biotechnol.* **62**, 22–28 (2020).

134. Costello, Z. & Martin, H. G. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *Npj Syst. Biol. Appl.* **4**, 1–14 (2018).

135. Ho, H.-I., Fang, J. R., Cheung, J. & Wang, H. H. Programmable CRISPR-Cas transcriptional activation in bacteria. *Mol. Syst. Biol.* **16**, e9427 (2020).

136. Fontana, J., Sparkman-Yager, D., Zalatan, J. G. & Carothers, J. M. Challenges and opportunities with CRISPR activation in bacteria for data-driven metabolic engineering. *Curr. Opin. Biotechnol.* **64**, 190–198 (2020).

137. Fontana, J. *et al.* Effective CRISPRa-mediated control of gene expression in bacteria must overcome strict target site requirements. *Nat. Commun.* **11**, 1618 (2020).

138. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).

139. Hawkins, J. S. *et al.* Mismatch-CRISPRi Reveals the Co-varying Expression-Fitness Relationships of Essential Genes in Escherichia coli and Bacillus subtilis. *Cell Syst.* **11**, 523-535.e9 (2020).

140. 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).

141. Vvedenskaya, I. O. *et al.* Massively Systematic Transcript End Readout, "MASTER": Transcription Start Site Selection, Transcriptional Slippage, and Transcript Yields. *Mol. Cell* **60**, 953–965 (2015).

142. Reis, A. C. *et al.* Simultaneous repression of multiple bacterial genes using nonrepetitive extra-long sgRNA arrays. *Nat. Biotechnol.* **37**, 1294–1301 (2019).

143. Wu, G. *et al.* Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. *Trends Biotechnol.* **34**, 652–664 (2016).

144. Bode, L. Human milk oligosaccharides: Every baby needs a sugar mama. *Glycobiology* **22**, 1147–1162 (2012).

145. Triantis, V., Bode, L. & van Neerven, R. J. J. Immunological Effects of Human Milk Oligosaccharides. *Front. Pediatr.* **6**, (2018).

146. Hwang, C. & Carothers, J. M. Label-free selection of RNA aptamers for metabolic engineering. *Methods* **106**, 37–41 (2016).