Engineering conditional guide RNAs for cell-selective regulation of CRISPR/Cas9

Thesis by
Zhewei Chen

In Partial Fulfillment of the Requirements for the
degree of
Doctor of Philosophy in Bioengineering

Caltech

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2023
Defended December 15, 2022

ACKNOWLEDGEMENTS

In this section, I would like to give thanks to everyone who have helped me along my PhD journey both professionally and personally. These next few paragraphs or pages are probably the least critical for completing my PhD, which is why I am writing them last. Yet, they are probably the most important personally because I can finally memorialize my favorite people and experiences from Caltech in a repository that will be archived until the end of time or civilization as we know it. Apologies in advance colleagues, lab mates, canim, ailem, veya arkadaslar for this long and overly detailed narrative. Without further ado, here it is.

First, I would like to thank my advisor Niles Pierce for his support and guidance all these years. I joined the Pierce lab because the lab's papers, figures, and vision were simply exquisite. After talking to Niles in person, I was completely sold on a PhD in conditional regulation and barely rotated in any other lab. It was to be the next big thing like the Nobel Prize for CRISPR. Looking back, I did not realize how hard it was to implement RNA nanotechnology inside living cells. Throughout this time, I really appreciated the patience and intellectual freedom Niles alotted us to work through this challenging project.

Second, I want to thank my committee members Richard Murray, Bruce Hay, and Paul Rothemund for their helpful suggestions and support in finishing this PhD. When I become a group leader someday, I hope my students can speak of me as highly as Richard Murray. Next, I would like to thank Bruce Hay who taught me most of what I know about gene expression, promoters, and recombination in his introduction to genetics class. Thank you for dealing with many of my annoying questions as a first year graduate student and helping entertain or provide advice on many bold genetic engineering ideas. Bruce also deserves credit for pushing me towards thinking about how to build high throughput screens for cgRNAs. Ultimately, this was the way to go for this PhD project. Next, I want to thank Paul Rothemund for his many helpful suggestions throughout my PhD. If I listened to Paul Rothemund sooner, I could have gotten a few more papers out under my name and finished my PhD a few years sooner.

Third, I want to thank my lab mates in the Pierce lab both past and present. We did not bond much outside of work, but there was always an environment of solidarity and community for seeking advice and gaining better perspective on our graduate school

hours and social activities, feeding us with their generous leftovers. Thank you Lucy Chong, Mike Flynn, Alex Ng, James Linton, Heidi Klumpe, Christina Su, Jan Gregrowicz, Yitong Ma, Ron Zhu, Lea Santat, and others. They remain our friends, and we are sad to see them move to the Chen building next door.

Fourth, I would like to thank my friends and former bosses who were influential in helping me come to Caltech. Pete Galie, Chris Voigt, Dr. Limin Wang, and Dr. Jan Stegemann at CMITE lab at University of Michigan and Ibrahim Gur, Isil Tulum, and Dr. Halil Kavakli at Koc University were fantastic mentors and hosts during my undergrad who all inspired me to eventually pursue a career as an academic. Laura DeForge, Henry Chiu, Guoying Jiang, Ingrid Lesaca, and Linda Chan were some of my bosses and mentors in my first job at Genentech. My skills in bioassays and experiment debugging were learned from Guoying, Ingrid, and Linda. Henry was the best boss I ever worked with. Anne and Ryan often joked he was like my work dad. One of my motivations for doing a PhD was because it was the only way to change the industry hierarchy that prevented people like Henry, Anne, Pam, Nga, Ryan, Kiran, Kwame, Jean-Michelle, Wilman, Daniel, and other research associates at BCP and BT from advancing beyond their current positions. Perhaps I will come back to Genentech in the next few months. Looking forward to reconnecting with old friends!

Fifth, I would like to thank my friends at Caltech who filled graduate life full of adventure and laughter around the campfire. Thank you Elle, Jono, Bradley, Bassam, Mingfeng, Linhan, Corwin, Jen, Marja, Li, Stephane, Tung, Ashish, Aaron, Justin, Katja, Eitam, Nick, Sujung, Ezgi, and others for all the seven years of adventure. I will remember all the mountains we climbed, the times we got benighted on the summit shivering together in the cold, the spectacular sunrises over the high sierras, and our salvation by beautiful valkyries from Yosemite Search and Rescue. I also want to give thanks to Branislav, who taught and kept alive the Caltech surfing and windsurfing club. Outside of climbing, I want to give thanks to Greg Ti, Nadia Blago, Rebecca Saive, Heidi Klumpe, and other students and post-docs who organized the salsa, west coast swing, and GSC social events. They were well needed outlets during times when I was too injured to climb. Thank you also new found Turkish friends Sinan, Arda, Utku, Atakan, Seda, and others for the gastronomic adventures in Glendale and our local area. I want to also thank Erin Burkett, Donna Wrublewski, and other members of the writing center and caltech library, who were immensely helpful in the preparation, review, and formatting of this thesis. Erin's

class on scientific writing was especially enlightening on how to read, structure, and publish scientific discourse. I want to also thank Justin Bois who taught us how to keep data tidy and be master snake charmers of python and statistical inference. Professionally, I also want to thank my fellow cohorts and organizers of the BLP program: Frances Arnold, Kim Mayer, Julie Kelly, Ngozi Eze, Bryan Gerber, and Aiden Aceves who helped organize many events related to industry networking and entrepreneurship. The BLP program led to my internship with Gates Foundation in Seattle during the COVID-19 pandemic. Thank you Luis, Ben, Sam, David, Kevin, and others for the wonderful experience in Seattle. I hope to come visit or work there again in better times. From the BLP program, Bryan was also instrumental in organizing the Nucleate Bio Activator program where myself, Kiran, Arnold, Mikhail, and Michael formulated a startup plan for the cgRNA technology created during my PhD. I want to give thanks to my teammates for winning the Alynlam award for best scientific plan. You guys are the best, and I hope we can spin out some sort of company together someday.

Sixth, I want to thank my wife Bilgenur who has been patiently waiting for me to finish my PhD. Many apologies canim for the instances where I misjudged how long it takes to setup PCR at midnight and/or get to the summit of Mt Whitney or Tenaya Peak. Thank you for introducing me to NGS, reviewing my manuscripts, working through the night with me to polish up my presentations, and generally suffering through everything with me including long hikes up the mountain. After many years of long distance romance from Singapore to Canada and the isolation of the pandemic, we really should be living the relaxed and beautiful fancy life we dreamed of. I promise life after PhD will be much better, and we will buy that dream house you have been yearning for.

Finally, I would like to thank my parents who are true visionaries, risk takers, and heroes in my life. Despite spending much of their childhood in the poverty and violence of the cultural revolution, they studied hard and were somehow lucky enough to be among the first generation of students admitted into chinese universities after those turbulent times. My mother studied at Shanghai Jiaotong University (MIT of China) and my father at Nanjing University (like Yale or Harvard of China). After China opened up to the western world shortly after I was born, they embarked on their own PhD journeys. We left behind most of our extended family, the familiarity, and the comforts of our home to see if we could make a better life here. Like most immigrant families, the first few years were especially hard. My grandfather passed

while my dad was still a graduate student. We were not able to travel back due to visa issues or cost. Long distance phone calls back then were also infrequent due to expense. I cannot imagine how a family of three was able to survive on one graduate student stipend. Nevertheless, they persisted. They mastered out of their programs and found jobs so they can earn enough to support the family. So no PhD. As of today, their proudest achievements would be sending me to university and paying off the house. Though they are not particularly wealthy, famous, or successful (they are still working day and night to save up enough for retirement), they seem content with their choices because of the opportunities it offered me.

Most of what I have accomplished or where I am today were built on the sacrifices and struggles of the previous generation, which is why I want to dedicate the completion of this PhD to everyone here, especially my parents. Let this PhD represent a completion of their initial PhD journey, so that all our paths can now come full circle as we await the beginning of the next one. Love you all. Thank you everyone!

ABSTRACT

CRISPR/Cas9 is a versatile platform for implementing diverse modes of genetic perturbation such as gene silencing, induction, deletion, or replacement. This technology is popularly used in developmental biology to probe genetic circuitry via constitutive gene knockdown. Global gene silencing could introduce artifacts in the study of developmental regulatory pathways, and this motivates the development of cell-selective gene editing. Our lab has recently created conditional guide RNAs (cgRNA) that enable CRISPR/Cas9 systems to silence a desired gene Y conditioned on the detection of an RNA transcript X inside of a cell. cgRNA systems were discovered via insertion and deletion mutations that systematically explored the structure function of the guide RNA. Nucleic acid engineering software (NUPACK) was used to generate orthogonal libraries of cgRNA molecules that executed both ON $\to$ OFF logic (conditional inactivation by an RNA trigger) and OFF $\to$ ON logic (conditional activation by an RNA trigger). A dCas9-based RFP silencing assay in bacteria was developed and used to show these cgRNA sequences were functional and could detect short exogenous trigger sequences in an orthogonal and dose-responsive manner. Subsequent studies on cgRNA structure and function enabled us to engineer next-generation systems that have fewer constraints on the trigger sequence or structure. These next-generation cgRNAs were tested against short synthetic mRNA transcripts, truncated sub-sequences of endogenous mRNAs, and full-length endogenous mRNAs. Synthetic mRNA transcripts were used to study the effect of protein translation on trigger RNA binding. cgRNAs were capable of detecting synthetic sequences embedded in the 3′ UTR of fluorescent protein mRNAs. cgRNAs could also detect short synthetic mRNAs or truncated sub-sequences from endogenous mRNAs. However, the detection of native full-length endogenous mRNAs remained challenging because we cannot reliably predict the local structure of sub-sequences within a long RNA transcript. High-throughput cgRNA screening may prove necessary for finding accessible binding sites on mRNA transcripts. Nevertheless, cgRNA functionalities could be useful in developmental biology by enabling precision perturbation of regulatory events, linking guide RNA activity to an RNA marker X correlated to a specific cell type or temporal expression pattern. This work opens the possibility for future applications such as cell-selective gene therapies.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. 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. **5.** PMCID: PMC6661866, 1241–1249. doi:10.1021/acscentsci.9b00340 (2019).

   ZC participated in the invention, design, analysis, and acquisition of data for the toehold switch and terminator switch cgRNA mechanisms in bacteria. The mechanistic studies on guide RNA structure and function which led to the discovery of the toehold switch and terminator switch systems were done by ZC. The assays used to measure cgRNA activity were developed by ZC and MHHH. Final data collection for the paper was done by MHHH. The paper was written by MHHH and NAP. Supplementary information written by MHHH, LMH and NAP.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION AND MOTIVATION

Conditional guide RNAs (cgRNAs) are CRISPR/Cas9 guide RNAs designed to refold into active or inactive RNA structures after binding to a trigger RNA sequence. The ability of these guide RNAs to toggle Cas9 endonuclease activity in response to the expression of endogenous RNA transcripts is conceptually powerful because it enables programmable control over when and where CRISPR/Cas9 regulation is active in a whole organism. cgRNAs have the potential to be important in the fields of developmental biology and medicine by helping us dissect how individual genes and gene networks drive cancer, aging, neuro-degeneration, and heritable diseases. It is hoped that programmable RNA molecules such as cgRNAs can someday be used to create smart drugs that can target cancer and/or disease-causing cells with accuracy and precision via the logic of RNA base pairing. The goal of this thesis is to outline a framework for engineering cgRNAs that can detect endogenous RNA sequences. The following subsections of this chapter provide a brief background on how RNA structures are predicted, how RNA strand displacement reaction pathways are engineered *in silico*, how *in vivo* RNA nanotechnology have been implemented so far, and why cgRNAs represent a breakthrough in the implementation of *in vivo* RNA nanotechnology and next generation gene editing therapies.

1.1 Prior work on RNA structure prediction

Rational design is a bioengineering methodology that emphasizes the use of computing power to engineer biomolecular structures that could perform novel biological functions such RNA or protein sensing and enzyme catalysis. In contrast to directed evolution, where the search for interesting sequences occurs primarily on the lab bench, rational design seeks to reduce experimental effort by performing most of the sequence space search *in silico*. This methodology hinges on the hope that computational predictions of protein or RNA structure are accurate enough such that most of the non-functional sequences are eliminated *in silico* before the search commences in an experimental setting.

At the time of writing this thesis, rational design is experiencing a new renaissance in biology because of the advent of more powerful computers, more accurate bio-physical models, and more efficient search algorithms. Today, many bioinformatic

tools for nucleic acid structure prediction exist (see Table 1.1). Some of these tools can be compiled and run on generic laptops. Others can be accessed through web interfaces that help leverage the power of computing clusters at universities. Each tool is targeted towards a unique application ranging from RNA secondary structure prediction to RNA-protein docking in 3D. Although each of these computational tools have some relevance to the study of Cas9 and cgRNAs, for the majority of this thesis, I focus on the use of NUPACK in the design of RNA strand displacement reaction pathways that can work *in vivo* in bacteria and eukaryotic cells.

Package	Target Application
UNAFold[1]	Nucleic acid secondary structure prediction
NUPACK[2,3]	Analysis and design of interacting nucleic acid strands in solution
Kinefold[4]	RNA co-transcriptional folding simulation
Rosetta FARFAR2[5,6]	3D coarse grained prediction of protein-RNA folding and docking
OxView[7–11]	Design and modeling of nucleic acid nanostructures using 3D coarse grained models

Table 1.1: A non-comprehensive list of nucleic acid prediction and design software

In NUPACK, RNA structure prediction is modeled as an ensemble of RNA secondary structures in thermodynamic equilibrium, meaning each strand of RNA in a test tube can fold into different RNA secondary structures and multi-stranded complexes. In the nearest neighborhood model of RNA structure, the propensity of an RNA strand adopting a particular structure is determined by the energies of nucleic acid base pairing[12]. These base pairing energies are derived from melt curves of RNA and DNA[13–16]. By recursively enumerating all the possible base pairs, loops, and dangles a particular sequence can adopt, and Boltzmann weighting these features with their thermodynamic energies, a partition function of all the possible structures for a particular sequence can be computed. The probability an RNA sequence will adopt a particular structure is simply the size of the sub-partition of this folded state divide by size of the partition function[12,17]. RNA structure design is thus the iterative optimization and selection of a set of RNA sequences that have a high likelihood of folding into a desired secondary structure versus all other possible states.

RNA strand displacement reaction pathway design is represented as a set of multi-stranded RNA complexes in solution at different stages of a reaction. Each RNA strand added to a reaction mix alters the distribution of possible multi-stranded

complexes. Reaction pathway design is the optimization and selection of a set of RNA sequences that minimize the ensemble defect over a set of target test tubes such that the desired structures at each reaction step are likely to form (See Figure 2.11 and 2.12).[18] This test tube design formula allows NUPACK to evaluate how well a set of guide RNA and trigger RNA pairs can switch from one conformation to another in response to the presence or absence of a trigger RNA strand. Efficient evaluation of the partition function allows NUPACK to quickly search through vast swaths of sequence space and find RNA sequences which can execute a certain RNA stand displacement reaction scheme.

Alternative rational design tools such as Rosetta[5,6] or OxView[7–11] are not tailored for engineering multi-state reaction pathways. Simulating all the possible states of RNA base pairing for multi-stranded complexes interacting in a solution in 3D is computationally intensive because 3D models of RNA have many degrees of freedom associated with bond rotation and atomistic location. Tertiary structure models are primarily useful for crystal structure refinement and coarse grained physics simulations of long nucleic acid scaffolds such as DNA origami. Here, tertiary interactions such as pseudoknot formation and bond curvature are important in determining the final shape of these nucleic acid nanostructures.[19–21] Luckily, these tertiary interactions are not critical in cgRNA engineering because cgRNAs are short (less than 150nt long).

A major drawback of nucleic acid secondary structure models is that they currently do not account for the stabilizing energies associated with RNA-protein docking. This phenomena could be important in systems such as CRISPR/Cas9 because the energy contributions from protein binding could confound the partition function in unforeseen ways. Conceptual gaps such as these may be responsible for the inconsistencies seen between computational predictions and experimental results (see Figure 2.4 and 2.5). Predictions from tools such as NUPACK are thus used with the acceptance that the results of the model are informative, but may only capture a small portion of the underlying physical system.

1.2 Prior work on nucleic acid nanotechnology in living cells

The inaccuracies inherent in computational models of biomolecules have not dissuaded researchers from leveraging computational tools to engineer dynamic RNA nanotechnology. In the last two decades, many breakthroughs have been achieved in this field. Some notable examples are the non-enzymatic amplification and de-

tection of endogenous mRNAs in fixed tissues with metastable DNA hairpins[22–26], the creation of DNA walkers and cargo sorters[27,28], neural network computation with DNA strands in a test tube[29], and the assembly of nucleic acid nanostructures from single stranded DNA or RNA[30–33]. Many of these advances were achieved in the *in vitro* or *in situ* setting where the composition and/or reaction temperature of biomolecules are carefully controlled. However, implementation of these system inside living cells is still challenging.

Implementing RNA nanotechnology inside living cells is challenging because the cellular environment is complex and unpredictable. The composition and concentration of cross interacting RNA molecules also co-expressed inside a cell at any given time is not fully known or under fine control. Native cellular processes such as the stochastic nature of RNA synthesis, co-transcriptional folding, post-translational processing, and degradation by native proteins and RNAses can further confound experimental results because these processes could be significantly perturbed by the choice of promoters used to express an RNA, the spatial arrangement of cross interacting genes on a plasmid, the growth conditions used, and the choice of fluorescent reporter used as the assay read out. These are problems most familiar to researchers developing synthetic gene circuits and cellular bio-assays that measure the behavior of some interesting biological processes such as dynamic gene repression, oscillation, and degradation[34–37].

Despite these challenges, notable examples of dynamic nucleic acid nanotechnology inside cells do exist. Many of these of these systems rely on harnessing the native pathways of RNA transcription, degradation, and translation. The STAR terminator system is an example where the mechanism of RNA transcript termination is manipulated to conditionally produce or not produce an RNA transcript.[38] This system works via premature termination of an RNA transcript via a synthetic terminator sequence that senses a trigger RNA. Conditional RNAi is an example where the native process of RNA degradation is manipulated to conditionally silence genes via the RNAi Dicer pathway.[39,40] Here, small conditional RNAs (scRNAs) are engineered to assemble an RNA-induced silencing complex when a trigger RNA is co-expressed. The toehold switch mRNA system is an example where the ribosome binding site on an mRNA is conditionally inactivated via a sequestering hairpin domain.[41–44] Trigger RNA binding opens up this hairpin allowing translation of the encoded sequence into protein. This protein could be used to perform some downstream output function such as gene silencing or simply act as a fluorescent

reporter to signal the cell's behavior has changed. The unifying theme among all these system is that they co-opt native RNA processes to execute novel functions. These technologies were possible after carefully studying the structure function of certain RNA processing proteins and their RNA partners. The fact that these technologies were possible speaks to the versatility of the RNA engineering paradigm enabled by predictions of RNA structure.

Experimentally, detection of endogenous sequences inside of living cells by synthetic RNA nanotechnology has not been trivial to implement. Prior work on the RNAi pathway showed that conditional RNAi systems can be engineered to activate in response to an mRNA input X and silence a different mRNA Y. These systems show endogenous mRNA detection is possible in human cell lysate. However, their implementation in living systems have only recently been successful after a long hiatus.[39,40,45,46]. These demonstrations were possible after the unforeseen off-target interactions prevalent in living systems were carefully identified and mitigated.

Living systems exist in a chaotic environment, and synthetic RNA sensing systems such as cgRNAs must be robust to this noise to a standard more strict than *in vitro* and *in situ* systems. Unlike DNA strand displacement (DSD) in a test tube, detection of RNA sequences must occur in a more crowded environment. In DSD experiments, the concentration of nucleic acid strands used is 50-100nM.[29] These strands were PAGE gel purified, and their exact composition in the well is carefully controlled. In living systems, there could be 0.63nM of cgRNA trying to detect 11nM to 34nM of a highly expressed mRNA, and this detection experiment must occur in a background of around 8.87μM total mRNA.[*] Unlike mRNA detection with HCR, high annealing temperatures and formamide cannot be used in living systems to suppress off-target interactions with the endogenous RNA background. This is even more challenging if you consider off-target protein interactions. HCR works on fixed tissues where proteins are inactivated or removed. In bacteria, there are on average 540 proteins per mRNA transcript in the cell.[†] Many of these proteins are involved in RNA transcription, translation, and degradation. The crowded nature of the cellular environment means the small amounts of signal leak in a test tube

---

[*]The total mRNA in bacteria grown in LB is 7800 molecules in 1.46 fL. This translates to around 8.87μM of total RNA in a cell. The average number of mRNA transcript per gene is 0.56. This is 0.63nM in bacteria and is used as the lower number for cgRNA expression. The mRNA copy number for a highly expressed gene like LacZ is at most 30. This is 34nM in bacteria and is used as the concentration for a highly expressed endogenous mRNA the cgRNA is trying to sense. These numbers were found on https://bionumbers.hms.harvard.edu.

[†]See number of protein to mRNA transcripts in a bacteria. https://bionumbers.hms.harvard.edu/bionumber.aspx?id=104186

are unacceptable in the cellular setting. These off-target interactions can be difficult to control or identify because current computational tools do not attempt to predict all the off-target interactions possible with other endogenous RNA transcripts and proteins. Efforts to engineer RNA strand displacement systems for robust operation in living systems must be cognizant of these limitations.

Pure computational design cannot replace the need for exceptional critical thinking and debugging skills in the experimental setting. This work requires out of the box thinking, boldness, and steadfast optimism in the face of failure. Incoming graduate students come with two hands and a pipette, and you must be prepared to tackle the problem from many different angles before finally getting things to work. Such is the nature of fundamental research.

1.3 Motivation for engineering conditional guide RNAs
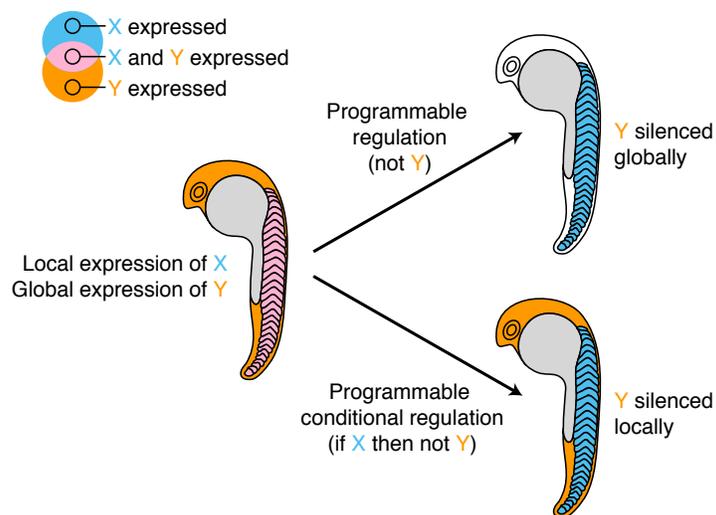
Figure 1.1: Diagram illustrating constitutive gene silencing vs conditional gene silencing. Molecular logic of programmable regulation using a standard programmable regulator, such as CRISPR/Cas gRNA ("not Y," top arrow) vs programmable conditional regulation using a small conditional RNA, such as a cgRNA ("if X, then not Y," bottom arrow). In this conceptual illustration, the standard programmable regulator silences Y in all tissues, while the programmable conditional regulator silences Y only in tissues where and when X is expressed, exerting spatiotemporal control over regulation. Figures adapted from Hanewich-Hollatz et al[47]

CRISPR/Cas9 technology plays a central role in biology by enabling scientists to perform precision gene silencing, activation, deletion, or replacement in a diverse array of model organisms[48,49]. A common application of this tool is in the study of genetic necessity in developmental gene circuits (see Figure 1.1 and 1.2)[50]. Here,

a Standard guide RNA (gRNA) structure and interactions

b Standard guide RNA (gRNA) logic and function

c Conditional guide RNA (cgRNA) logic and function
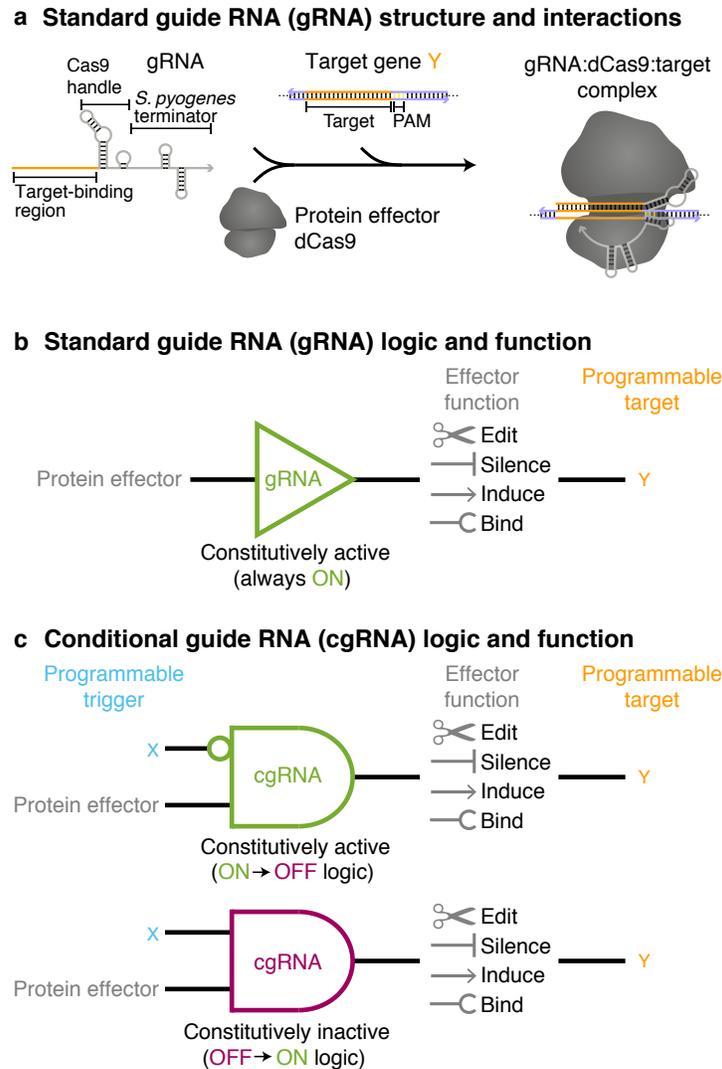
Figure 1.2: Illustration comparing canonical guide RNAs to cgRNAs. a) Illustration of dCas9 and guide RNA activity in CRISPR/Cas9. Guide RNA binds to a target gene. dCas9 binds to the guide RNA:target complex and silences the target gene. b) Illustration of canonical guide RNA logic. Expression of guide RNA and Cas9 results in constitutive silencing, activation, or editing activity. c) Illustration of cgRNA logic. Input trigger RNA can toggle cgRNA ON to OFF or OFF to ON . Tissue-specific promoters are not needed to enable cell selectivity.

Cas9 is used to knockdown a gene of interest in a developing embryo via gene deletion. The perturbation to the regulatory pathway is analyzed to understand the gene's function. Constitutive genetic perturbations with Cas9 can be less illuminating because researchers are often interested in the temporal behavior of genes in a particular genetic pathway of a certain cell type. Genes in other tissues do not need to be knocked down because this can alter or kill the embryo being studied.

CRISPR activity can be restricted to certain tissues by expressing guide RNAs on tissue-specific promoters[50] or using tissue-specific delivery vectors such as a virus[51]. Tissue-specific promoters are limited by their genomic availability. Developmental biologists often have to mine for tissue-specific enhancers or promoters from the genome using differential gene analysis and chromatin immunoprecipitation sequencing[52,53]. Sometimes promoters or enhancers for a particular cell or tissue type may not exist. Tissue-specific viruses are similarly limited by their high upfront costs. Viruses exhibiting tropism for a particular tissue need to be painstakingly engineered and screened for affinity and selectivity.[54,55] Not all labs or universities have this capability.

This thesis introduces a new modality for controlling when and where CRISPR Cas9 is activated in a whole organism. Rational design tools are used to generate conditional guide RNAs, whose activities can be modulated by the expression of their cognate RNA trigger sequence. For example, cgRNAs can be made to switch ON only if a certain RNA trigger sequence is expressed in the cell. Alternatively, cgRNAs can be designed such that the cognate trigger RNA will inactivate the cgRNA. In principle, tissue selectivity could be achieved by engineering cgRNAs that trigger against marker genes which are highly expressed in certain tissues or cell types (see Figure 1.2). Unlike tissue-specific promoters, cgRNAs are not mined from the genome and are not restricted by the availability of genomic parts. cgRNAs do not have to be engineered in the model organism where it is finally used. The output targeting sequence and promoters used to express cgRNAs can be altered to adapt cgRNA sequences toward different target applications. Unlike viral constructs, cgRNAs are designed *in silico* by computational software such as NUPACK. Most university labs have the ability to run NUPACK on a laptop and validate cgRNA sequences in mammalian or bacteria-based screening assays for activity and specificity. These features make cgRNAs a more versatile and accessible platform for implementing spatiotemporal control over the CRISPR toolbox.

1.4   Thesis overview

The goal of this thesis is to outline current methodology for engineering cgRNAs, discuss the identification of new guide RNA structure/function relationships, and show the experimental validity of cgRNA activity and specificity for the cognate trigger sequence. Chapter 2 describes prior work on Cas9 guide RNA structure/function relationships and how this led to the first breakthrough in design of cgRNAs. Chapter 3 describes new studies on guide RNA structure/function, how this information

helped with engineering next-generation cgRNA designs that have reduced design constraints, and the construction of a medium-throughput screening assay for evaluating new cgRNA mechanisms. Chapter 4 describes attempts to use next-generation cgRNA designs to detect endogenous RNA targets. This chapter details prior work and challenges associated with mRNA detection, insights learned, and new computational tools for triaging designs in a cgRNA-mRNA screen. Chapter 5 describes recommendations for future study. The appendix contains supplemental material such as plasmid maps, cgRNA sequences, minimum free energy (MFE) structures, supplemental data plots, and a doi link for the repository of genbank files and source code used in this thesis.

C h a p t e r   2

ENGINEERING CONDITIONAL GUIDE RNAS

Engineering of conditional guide RNAs (cgRNAs) was made possible by careful exploration of the Cas9 guide RNA structure/function relationships. In prior work and my own studies, key structures in the guide RNA were mutated, inserted, and conditionally broken to switch guide RNAs between active and inactive conformations. In this chapter, I describe engineering strategies that worked and contrast them with strategies that failed. The following sections discuss the overall structure and function of guide RNAs within the Cas9 system, how each portion of guide RNA structure can or cannot be manipulated to engineer a cgRNA, and how orthogonal sets of cgRNAs were engineered *in silico* with NUPACK and validated experimentally in bacteria.

2.1   CRISPR/Cas9 background, structure, and function

CRISPR/Cas9 exists in nature as a prokaryotic defense mechanism to destroy foreign RNA and DNA[48]. This bacterial immune system has since been re-engineered as a versatile tool for gene editing across many organisms from bacteria and animals to plants[49,56]. Since CRISPR/Cas9's discovery, many new gene editing modalities have been added to the Cas9 toolbox. Cas9 proteins can now insert genes[57], delete genes, correct single base pair mutations[58,59], activate genes[60], or silence genes with great precision[61]. Most modifications to the CRISPR system have focused on re-engineering the Cas9 protein to improve gene editing efficacy. Less attention has been given to the guide RNA, which serves a critical function in all CRISPR systems.

In CRISPR/Cas9, guide RNAs determine the DNA targets of Cas9 endonuclease[48]. They are essential co-factors of Cas9 endonuclease activity, and each homolog of Cas9 has its own cognate guide RNA. The specificity of each Cas9 variant to its guide RNA is determined by conserved motifs embedded in the guide RNA secondary structure[62,63]. Breaking these secondary structures indirectly breaks Cas9 endonuclease activity. Thus, if the secondary structure of the guide RNA can be manipulated to switch between active and inactive states in response to some input such as a trigger RNA or ligand, then Cas9 activity can be controlled indirectly in a programmable manner. This engineering strategy is versatile and modular because

one cgRNA can provide new novel functionality and programmable control over Cas9 activity for all editing modalities in the Cas9 toolbox.
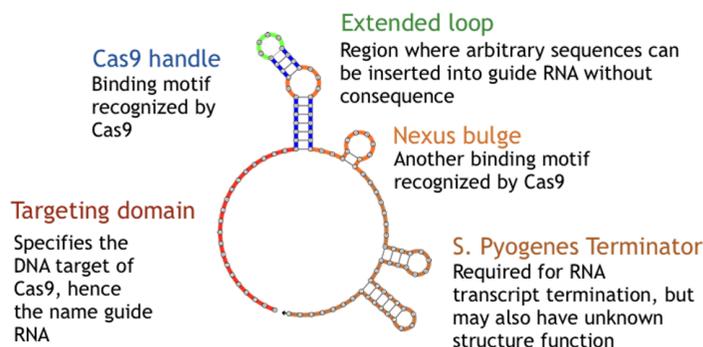
Figure 2.1: Illustration of the *S. pyogenes* Cas9 guide RNA secondary structure. Colored nucleotides denote the different functional domains on the guide RNA, which were manipulated to engineer conditional guide RNAs.

Manipulation of guide RNA activity is possible because guide RNAs contain four important motifs: a DNA targeting domain, Cas9 handle, nexus bulge, and terminator stems. The DNA targeting domain determines the DNA binding site of Cas9 endonuclease. The Cas9 handle and nexus bulge are conserved motifs required for docking with Cas9 protein. The terminator stems are responsible for bacterial RNA transcript termination. Prior work suggests that changing the sequence and structure of these motifs will alter the activity of the guide RNA[62,63].

2.2 Manipulating the guide RNA structure to create conditional guide RNAs

This section provides an overview of my attempts to re-engineer each functional domain of the guide RNA. The conditional guide RNA systems developed from structure perturbation experiments are organized by the functional domains utilized to break or unbreak the guide RNA.

Manipulating the targeting domain

Manipulating the targeting domain region was one of the strategies explored to engineer conditional guide RNAs. The target domain defines the genomic target for Cas9 binding and silencing. This domain is tolerant of sequence alterations and allows targeting of new genomic target by simply changing the targeting sequence. A canonical guide RNA binds strongest to DNA that is its targeting domain reverse-complement. The targeting domain must be located next to a protospacer adjacent motif (PAM) sequence (see Figure 2.2a). For *S. pyogenes* Cas9, the PAM sequence is NGG. The whole targeting domain on the guide RNA is 5'-target domain-NGG-

3'. Single base pair mismatches decrease the affinity of the guide RNA for its DNA target, but do not always fully abolish activity (see Figure 2.2b). The first 15 nucleotides of the targeting domain adjacent the Cas9 handle are important for determining DNA targeting efficacy. Mutations beyond these 15 nucleotides on the 5' end of the guide have little impact on guide RNA activity[48]. This suggests arbitrary sequences can be inserted into the 5' end of the guide RNA (see Figure 2.2c) to perturb native guide RNA structure in a programmable manner.
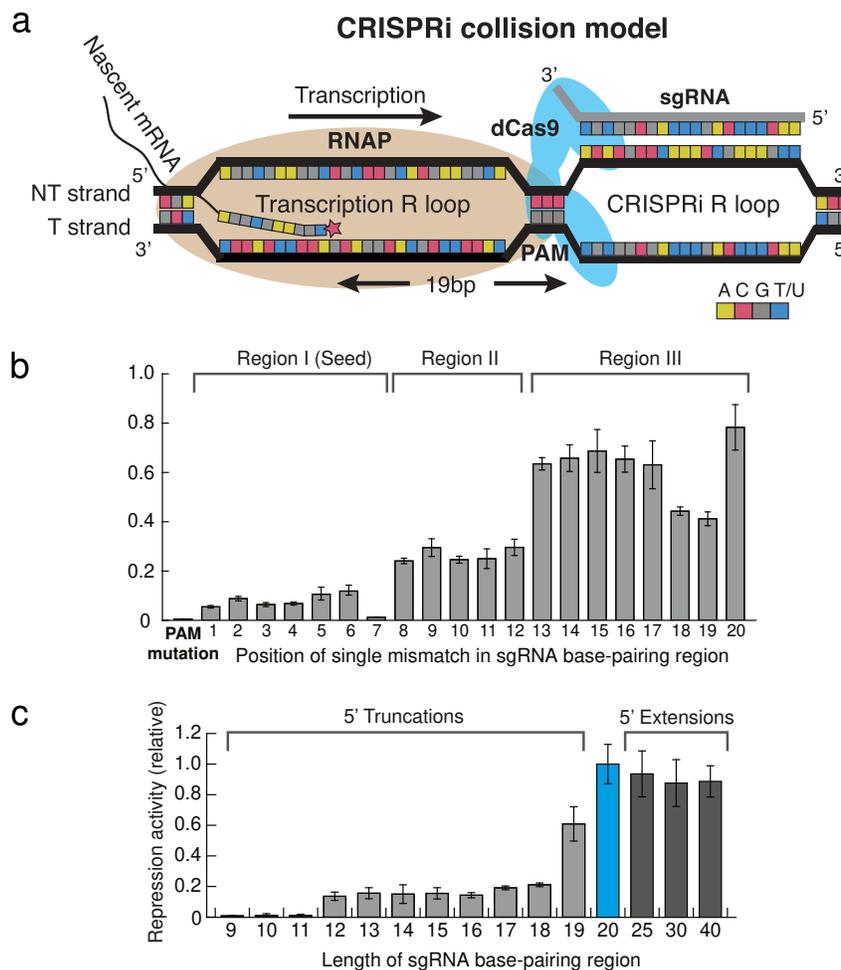
Figure 2.2: Schematic and data about the Cas9 guide RNA targeting domain. a) An illustration of how dCas9 inhibits RNA transcription via binding to DNA. dCas9 clamps down on DNA at the PAM site, blocking RNA polymerase translocation. dCas9 is guided to the PAM via the adjacent guide RNA. b) dCas9 silencing activity vs single base pair mutations on the targeting domain. Mutation near the PAM site break the guide RNA. Mutations in region III away from the PAM site have less impact on guide RNA activity. c) Truncation and extension mutations on the targeting domain of the guide RNA. 5′ extensions to the guide RNA do not break the guide RNA. A targeting domain of 20 nucleotides in length is sufficient to maintain good repression activity. Figures were adapted from Qi *et al.* [48].
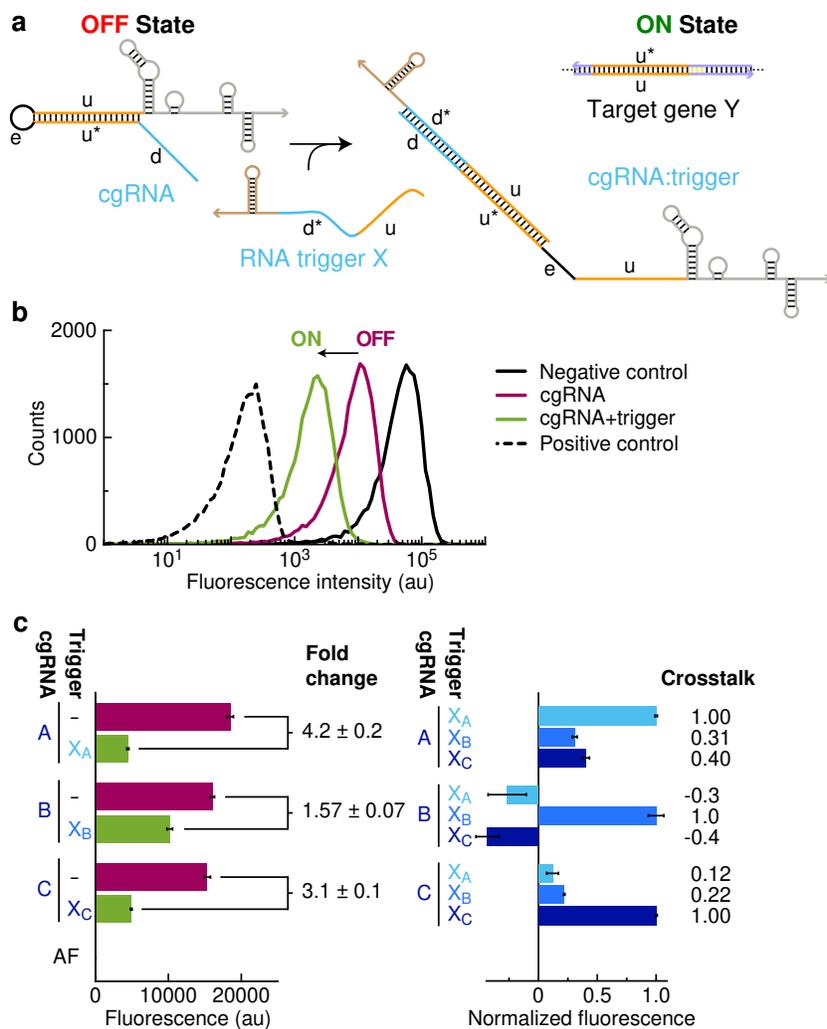
Figure 2.3: Illustration and data on the toehold switch cgRNA mechanism. This system uses OFF to ON logic (high to low fluorescence). a) Schematic of the toehold switch mechanism. The targeting domain u is sequestered in a hairpin by domain u*. Expression of a trigger RNA binding to domains d-u opens the hairpin allowing cgRNA to silence RFP via dCas9. b) Flow cytometry analysis of RFP silencing by dCas9 and toehold switch cgRNA in bacteria. Co-expression of the trigger RNA results in decreased RFP signal due to improved RFP silencing. c) Orthogonal detection of three different trigger RNAs by the toehold switch cgRNAs. Left: Raw fluorescence depicting OFF→ON conditional response to cognate trigger (fold change = OFF/ON = [no trigger−AF]/[cognate trigger−AF]). Right: Normalized fluorescence depicting orthogonality between non-cognate cgRNA/trigger pairs (crosstalk = [non-cognate trigger − no trigger]/[cognate trigger − no trigger]). Bar graphs depict mean ± estimated standard error calculated based on the mean single-cell fluorescence over 20,000 cells for each of $N = 3$ replicate wells (OFF:ON ratio and crosstalk calculated with uncertainty propagation). Figures were adapted from Hanewich-Hollatz *et al.* [47]

The toehold switch cgRNA[47,64] was one of the first working cgRNA designs (see Figure 2.3). This mechanism was designed to be constitutively inactive, and switches on only if its cognate trigger RNA was co-expressed. It worked by first sequestering

the targeting domain in an RNA hairpin engineered onto the 5' end of cgRNA. This hairpin inhibited DNA base pairing. Adding a toehold sequence to the 5' end of this hairpin allowed trigger RNAs to open the hairpin, which unblocked access to the targeting domain. This shape sequence transduction activated the cgRNA, allowing it to silence its genomic target in conjunction with silencing dCas9.

The main advantage of the toehold switch cgRNA is that it is conceptually simple and easy to design. The blocking motif is simply the reverse-complement of the targeting domain. Most of the design effort was spent on finding 5-10 nt long unstructured sequences which can act as a toehold for the trigger RNA. The disadvantage of this mechanism is that the trigger RNA must contain some sub-sequence of the targeting domain. This design cannot generate cgRNAs which have independent RNA triggers and DNA targets. If the targeting domain is a fixed sequence, design of orthogonal cgRNAs is difficult because each trigger RNA must share more than half its sequence with another trigger RNA, increasing the likelihood of crosstalk.

Manipulating the Cas9 handle

In contrast to the targeting domain, the Cas9 handle and nexus bulge were not amenable to redesign. Manipulating these regions was challenging because the Cas9 handle and nexus bulge were fixed recognition motifs involved in Cas9 binding. Single base pair mutations to the Cas9 handle could break the guide RNA. The only region amenable to modification was the Cas9 handle loop (see Figure 2.1). Truncation studies show that the handle loop can be truncated or extended without breaking the guide RNA[62,63]. The crystal structures of Cas9 and guide RNA showed that the handle loop region lies on the exterior of the protein exposed to solvent. Sequence insertions in this region were probably not deleterious because the handle loop was not involved in protein docking. The adaptability of the Cas9 handle loop and extensibility of the 5' targeting region could potentially be leveraged to engineer systems that can attempt to conditionally break the Cas9 handle motif. Several variations of this strategy were tried. The most promising cgRNA mechanisms were called the kissing loop and anti-loop systems.

The kissing loop mechanism was designed to use OFF to ON logic by inserting complementary domains in the Cas9 handle loop and the 5' region of the targeting domain (see Figure 2.4). If the insert domains were long enough, NUPACK predicted these inserts could break the secondary structure of the Cas9 handle. In theory, this would inactivate the guide RNA. Co-expression of a trigger RNA would
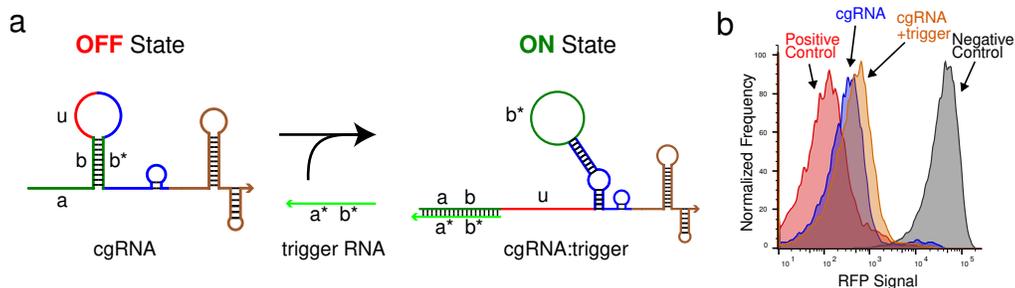
Figure 2.4: Illustration and data for the kissing loop mechanism. a) Schematic of the kissing loop mechanism. This mechanism was designed to switch from OFF to ON (high to low fluorescence in a dCas9-based RFP silencing assay). The Cas9 handle is inactivated by a duplex formed by domain b and b* in the Cas9 handle loop. Co-expression of a trigger RNA binding to domain a-b opens the hairpin and restores guide RNA activity. Red denotes the targeting domain. Blue denotes the Cas9 handle. Brown denotes the terminator domains. Green denotes the variable domains designed by NUPACK. b) Flow cytometry analysis of dCas9 silencing RFP in bacteria via the kissing loop mechanism. Domain b and b* failed to break the Cas9 handle. Co-expression of the trigger RNA did not further decrease RFP signal. This mechanism did not work. See Table A.3 for cgRNA and trigger sequences. Data from scRNA_Cas9_20160628ZC
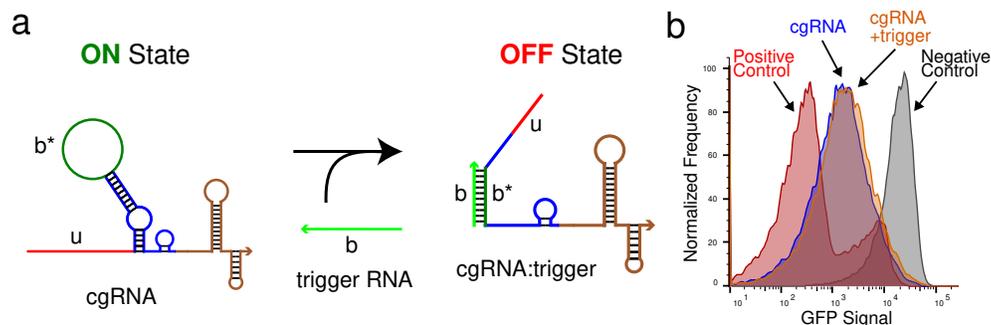
Figure 2.5: Illustration and data for the anti-loop mechanism. a) Schematic of the anti-loop mechanism. This mechanism was designed to switch from ON to OFF (low to high fluorescence in a dCas9-based GFP silencing assay). The cgRNA is constitutively active. Co-expression of a trigger RNA binding to domain b* in the Cas9 handle loop breaks the guide RNA. Red denotes the targeting domain. Blue denotes the Cas9 handle. Brown denotes the terminator domains. Green denotes the variable domains designed by NUPACK. b) Flow cytometry analysis of dCas9 silencing GFP in bacteria via the anti-loop mechanism. Co-expression of a trigger RNA binding to domain b* in the Cas9 handle loop failed to inactivate the cgRNA. See Table A.4 for cgRNA and trigger sequences. Data from scRNA_Cas9_20160628ZC

open the hairpin, allowing the Cas9 handle to form and enable docking with the Cas9 effector protein. This design did not have the onerous sequence constraints like those imposed in the toehold switch mechanism. The RNA trigger sequence could be independent from the DNA target. Unfortunately, this design did not work experimentally. Guide RNA activity could not be broken by the kissing loop do-

main, and no major shifts in RFP signal could be detected when trigger RNA was expressed (see Figure 2.4).

The anti-loop mechanism was a trans system, which inverted the kissing loop mechanism to an ON to OFF logic. Here, sequences were inserted in the handle loop. The cgRNA was initially in an active conformation. Co-expression of a trigger RNA would break the cgRNA via base-pairing to domain b* in the handle loop. NU-PACK predicted that duplex formation at the handle loop would break the secondary structure of the Cas9 handle. Experimentally, this approach failed. Loop inserts somewhat decreased the performance of the cgRNA. However, co-expression of a trigger RNA had no impact on the activity of the anti-loop cgRNA (see Figure 2.5). An OFF state could not be achieved.

These studies illustrate the shortcomings of RNA structure predictions that do not take into account the energy contributions from protein binding. Cas9 handle formation is stabilized by both the energies of base pairing and protein docking. The protein's interaction with the handle motif biases the thermodynamic ensemble toward an alternative shape that can differ greatly from the minimum free energy (MFE) structure of RNA.* Since the kissing loop mechanism and anti-loop mechanism both failed to break the Cas9 handle, the energy of Cas9 binding to handle motif must be greater than the energy of RNA duplex formation. This conceptual gap prevents us from manipulating the Cas9 handle. All strategies which attempted to break the Cas9 handle were ultimately unsuccessful.

Manipulating the terminator stem

In terms of versatility, modifications to the terminator stem domain strike a balance between the Cas9 handle domain and the targeting domain. Prior work indicates the double stem loop structure of this region cannot be changed.[63] Truncations on the terminator stems result in loss of activity, but do not fully break the guide RNA. Unlike the Cas9 handle, the terminator stem can be any sequence that forms a stem loop.[47] This means these regions can act as useful landing sites for inserting arbitrary sequences, which break or unbreak the guide RNA.

The terminator switch mechanism was an example of a system where the terminator loop is used as a landing site for base pairing with trigger RNA. This mechanism uses ON to OFF logic. The cgRNA is constitutively active. Expression of a trigger

---

*The minimum free energy structure is the highest probability structure in the structural ensemble at equilibrium.
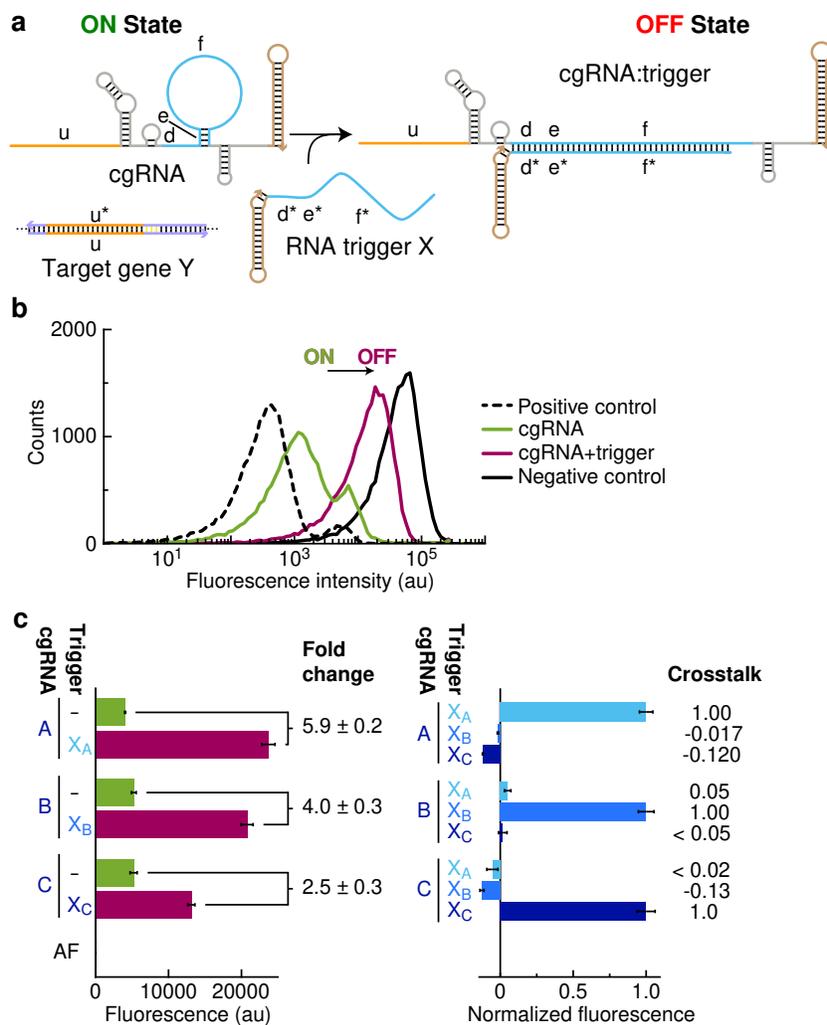
Figure 2.6: Illustration and data on the terminator switch cgRNA mechanism. This system uses ON to OFF logic (low to high fluorescence). a) Schematic of the terminator switch mechanism. The cgRNA is constitutively active. Expression of a trigger RNA binding to domains d-e-f sequesters the cgRNA. b) Flow cytometry analysis of RFP silencing by dCas9 and terminator switch cgRNA in bacteria. Co-expression of the trigger RNA results in increased RFP signal due to less RFP silencing. c) Orthogonal detection of three different trigger RNAs by the terminator switch cgRNAs. Left: Raw fluorescence depicting ON→OFF conditional response to cognate trigger (fold change = OFF/ON = [cognate trigger−AF]/[no trigger−AF]). Right: Normalized fluorescence depicting orthogonality between non-cognate cgRNA/trigger pairs (crosstalk = [cognate trigger − no trigger]/[non-cognate trigger − no trigger]). Bar graphs depict mean ± estimated standard error calculated based on the mean single-cell fluorescence over 20,000 cells for each of $N = 3$ replicate wells (OFF:ON ratio and crosstalk calculated with uncertainty propagation). Orthogonality results qualitatively agreed with computational predictions (see Figure 2.12). Figures were adapted from Hanewich-Hollatz *et al.* [47]

RNA binding to the stem loop domain breaks the cgRNA (see Figure 2.6). In a dCas9-based RFP silencing assay in bacteria, terminator switch cgRNAs containing

large loop insertions in the first terminator stem were able to silence RFP almost as effectively as an unmodified guide RNA. Co-expression of a trigger RNA binding to the loop insertions was able to increase the RFP signal (see Figure 2.6). These results suggest the stem loop motif is important for guide RNA activity even though these structures lie on the exterior of the Cas9 protein.[†]

A major advantage of the terminator switch is sequence independence between the trigger RNA and DNA target. The trigger RNA does not have to contain any sub-sequences related to the targeting domain or Cas9 binding motifs. The domain e-d-f-e* encompassing the first terminator stem can be any sequence which forms a stem-loop structure. This reduces the constraints on the trigger RNA such that triggering off of mRNA transcripts is a possibility (see Figure 4.7).

The main disadvantage of the terminator switch is the difficulty of finding sequences which can form a good stem loop structures. Strong stem formation is necessary for good ON state activity (see Figure 2.7). Highly unstructured terminator loops are important for trigger RNA binding and maintaining good OFF states. Trigger RNA binding occurs through a pseudoknot loop infiltration process, which can be orders of magnitude slower than RNA strand displacement via a toehold sequence[65]. Trigger sequences which optimally satisfy these requirements tend to have low nucleotide diversity and high GC content. The reverse complement of many endogenous sequences, especially protein coding genes, do not always fit these criteria. This limits the scope of endogenous sequences which can be used as trigger RNAs.

The exact biophysical reason for why the terminator switch cgRNA works is a mystery. The terminator stem was not expected to have any functional role in guide RNA activity other than to assist in bacterial transcript termination during RNA synthesis. This assumption runs contrary to experimental evidence. Terminator switch cgRNAs also work in eukaryotic cells[47,66], which do not use terminator stems for transcript termination. Mutagenesis studies on guide RNA function show that artificial extensions of the linker loop or deletions of the terminator stem can break the guide RNA[62,63]. In terminator switch cgRNAs, mutations that decrease the probability of terminator stem formation result in cgRNAs which have poor activity (see Figure 2.7). These studies suggest the terminator stem must play an important role in Cas9 and guide RNA docking.

Contemporary molecular dynamics studies of guide RNA and Cas9 docking have

---

[†]The crystal structure of dCas9 and guide RNA can be found at https://www.rcsb.org/3d-view/4ZT9/1
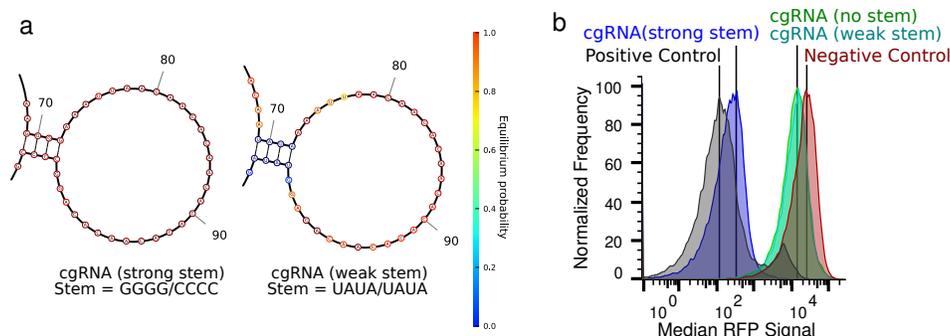
Figure 2.7: Mutational analysis of the terminator switch structure/function. a) NUPACK analysis of the minimum free energy structure and base pairing probabilities of cgRNA constructs with weak or strong terminator stems (see Table A.5 for gRNA sequences). b) Flow cytometry measurement of RFP silencing in bacteria by cgRNA constructs with weak, strong, and no stem loops. Weak and no-stem cgRNAs have no activity. Only strong stem cgRNA and positive control are able to silence RFP. Data from scRNA_Cas9_20190211ZC

ignored the role of terminator stem in guide RNA and Cas9 function. Most molecular dynamics studies on Cas9 focus on characterizing DNA off-target binding.[67–69] These studies examined how amino acid residues near the catalytic domains enable PAM site recognition and how RNA-DNA duplex formation in the targeting domain enable specificity over DNA targeting. The role of the terminator stems in guide RNA function has not been explored with molecular dynamics models. This is a unique opportunity to illuminate how nucleic acid residues lying outside of the catalytic domains influence the core structure and function of a Cas endonuclease. This information could be important for engineering high performing cgRNA mechanisms for *S. pyogenes* Cas9 and its orthologs.

Design of an *in vivo* cgRNA activity assay

A key part of engineering cgRNAs was the ability to measure the interaction between cgRNA and trigger RNA in an *in vivo* setting. These measurements must be comparable between different cgRNA and trigger sequences being tested. Engineering a reliable measurement system in living organisms was not trivial and is an underappreciated aspect of experimental work. The following describes the rationale for the choice of reporter gene, promoter, and plasmid layout for cgRNA activity assay in bacteria. The intention of this section is to convey the valuable lessons learned from building a high fidelity measurement systems for CRISPR activity in living organisms.

Monomeric red fluorescent protein (mRFP) was used as a fluorescent reporter for gene activity, and guide RNAs were designed to silence this reporter gene.
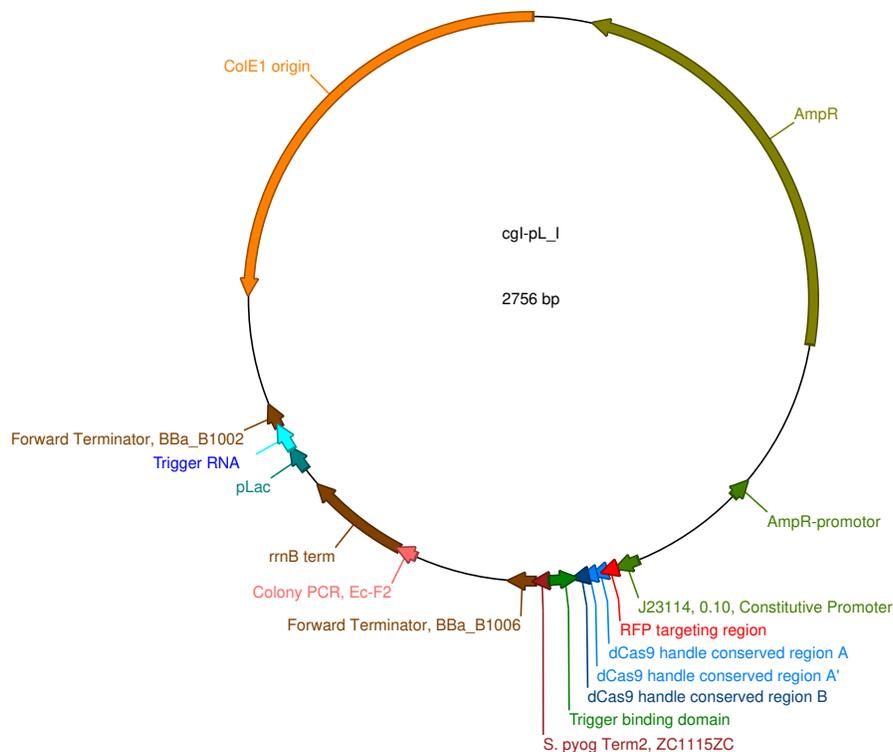
Figure 2.8: Map of the cgRNA plasmid used in the low throughput setup. cgRNA and trigger sequences are inserted behind the constitutive promoter and pLac promoter via insertion PCR, golden gate cloning, or gibson assembly. Trigger RNA expression on the pLac promoter could be induced with IPTG. AmpR gave this plasmid carbenicillin resistance. Genbank files for this plasmid are listed in Table A.17.
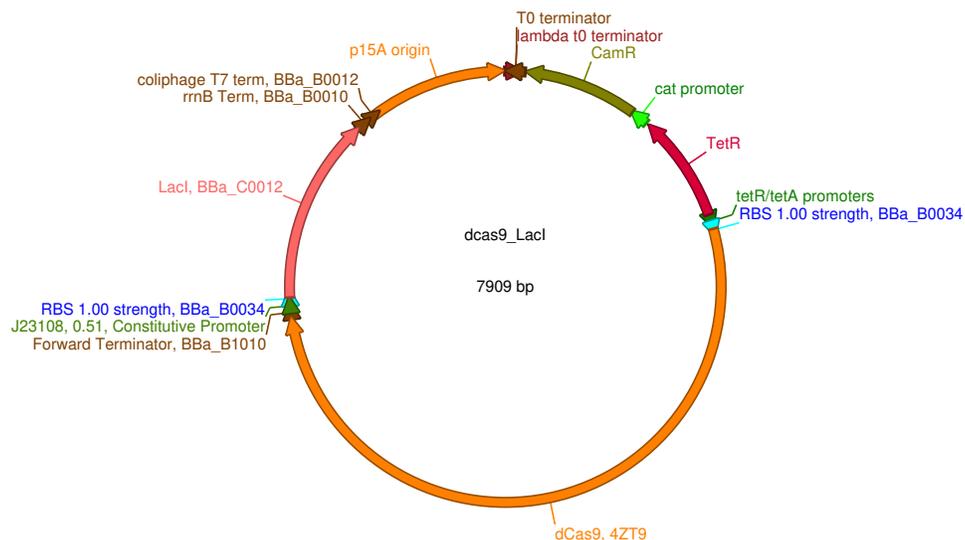
Figure 2.9: Map of the dCas9-LacI plasmid used in the low throughput setup. dCas9 is expressed on a pTet promoter, which can be induced with anhydrotetracycline (aTc). LacI was expressed on a constitutive promoter. CamR gave this plasmid chloramphenicol resistance. Genbank files for this plasmid are listed in Table A.17.
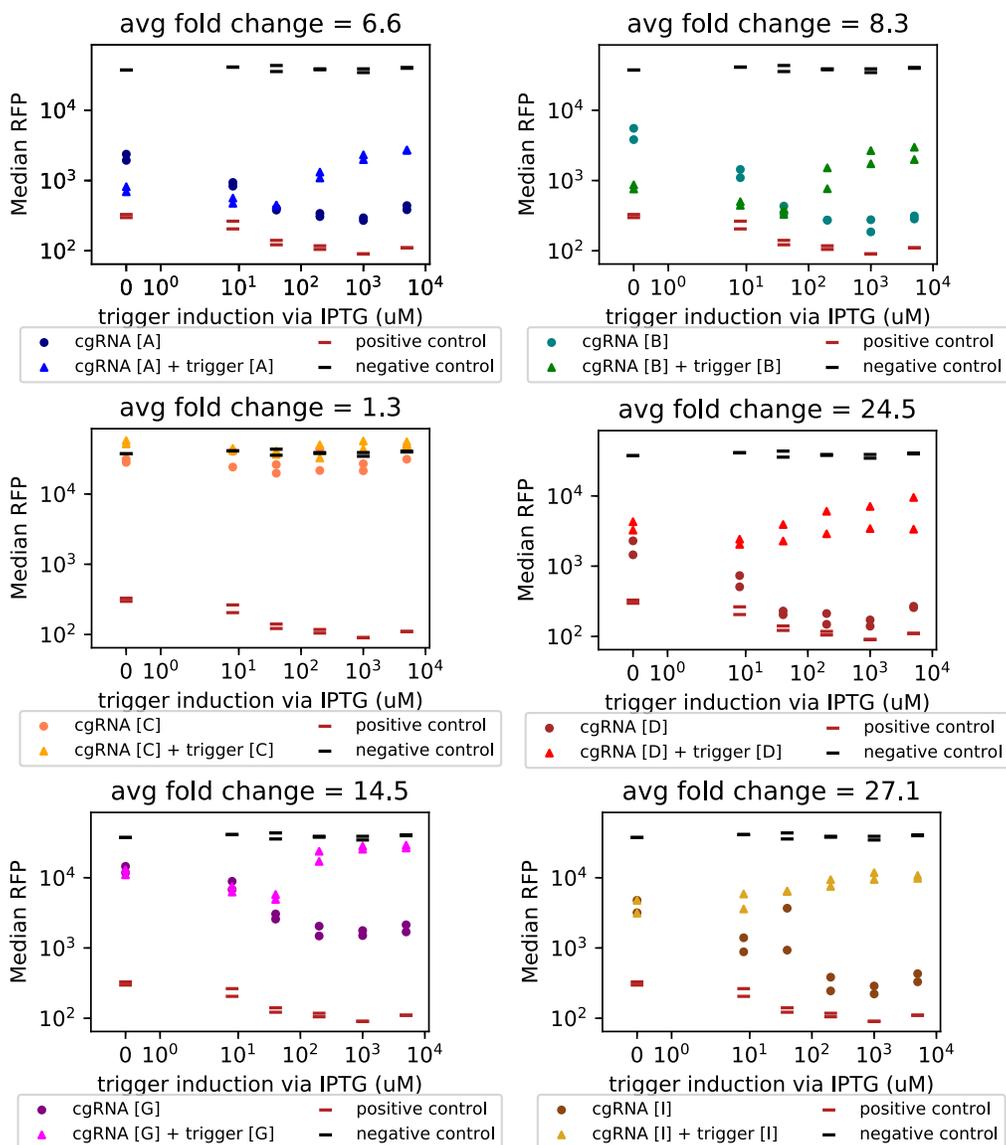
Figure 2.10: Trigger dose response for six different terminator switch cgRNAs silencing RFP in bacteria with dCas9. Trigger RNA was expressed on a pLac promoter and induced with IPTG. cgRNA was expressed with a weak constitutive promoter. RFP signal was acquired via flow cytometry. Samples represent replicates of N=2. All samples are dose responsive to trigger RNA except cgRNA[C]. cgRNA C is non-functional because the stem loop structure failed to form (see Figure A.3). cgRNA and trigger RNA sequences used in this experiment are in Table A.6. Data from scRNA_Cas9_20180512ZC.

This feature was inherited from the guide RNA activity assay used by Stanley Qi[48]. This assay uses an MG1655 strain of *E. coli* containing a single copy of mRFP genomically incorporated into an nfsA gene landing site on the bacterial chromosome (see Table A.1). Changing the sequence of the targeting domain can alter guide RNA activity. Changing the reporter gene also changes the apparent

guide RNA silencing performance. To make guide RNA activity measurements comparable across all designs, the RFP targeting domain and reporter gene was kept fixed for all subsequent assays. The same positive and negative control guide RNAs originally developed by Stanley Qi[48] were used to calibrate the full range of signal possible from the ideal ON to OFF states in the *in vivo* silencing assay (see Table A.2).

Trigger RNA and cgRNA were expressed on a pLac promoter and a weak constitutive promoter (see Figure 2.8) respectively to enable tuning and induction of cgRNA and trigger RNA expression by isopropylthio-β-galactoside (IPTG). This is important for assessing the sensitivity of a cgRNA system for a trigger RNA. Dose titration studies of cgRNA and trigger RNA (see Figure 2.10) revealed that for some designs, such as the terminator switch, the trigger RNA needs to be expressed in excess of the cgRNA for fold change activity to be measurable. Over expression of cgRNAs can reduce the measurable fold change response because not enough trigger RNA is available to fully repress the cgRNA.

dCas9 was expressed on a pTet promoter and used to inducibly silence mRFP (see Figure 2.9). dCas9 was used as the effector Cas because it will not permanently delete the mRFP reporter gene. If catalytically active Cas9 were used, the trigger RNA may not have enough time to inactivate a cgRNA before mRFP is permanently silenced. This situation is most relevant for constitutively active cgRNAs such as the terminator switch (ON to OFF logic). If the catalytic Cas9 were used, fold change signal for these designs could not be measured because the RFP would already be deleted.

Supplemental LacI expression was added to the dCas9 chloramphenicol plasmid to enable proper operation of the pLac promoter (see Figure 2.9). Native LacI expression in the MG1655 strain was weak. High LacI expression was needed to properly repress the Lac operon on the pLac promoter, otherwise induction with IPTG will fail.

The cloning vectors for the carbenicillin and chloramphenicol plasmid were derived from plasmids used by Stanley Qi[48]. LacI was inserted into the dCas9 plasmid via Gibson assembly and expressed on a strong constitutive promoter. Cloning of guide RNA and trigger RNA into the carbenicillin vector was done sequentially via insertion PCR. The fidelity of all constructs was verified by Sanger sequencing during cloning and after the fluorescence measurement. Fluorescence was measured by plate reader (Synergy Neo2) or a flow cytometer (MACS VYB or Cytoflex). For

flow cytometry data, the gating scheme used to analyze data is detailed in Figure A.1.

2.3 Computational design and validation of orthogonal cgRNAs

Orthogonality is an important test for showing that cross compatible sets of cgRNAs can be designed *in silico* by NUPACK. This assay measures the specificity of the cgRNA for its cognate trigger RNA. In situations where the fold change response to trigger RNA is weak, the orthogonality test is the only way to verify the cgRNA works as intended, meaning the cgRNA only responds to its cognate trigger RNA. If the orthogonality test fails, there is a high chance the fluorescence shifts observed in the silencing assay are due to metabolic effects[‡] [§] If the orthogonality test is passed, it suggests cgRNAs were successfully engineered with fluorescence shifts caused by a change in guide RNA activity resulting from binding with the cognate trigger RNA.

Orthogonal sets of cgRNAs for the toehold switch and terminator switch were designed using the test tube design function in NUPACK[47]. In the test tube design scheme for orthogonal systems, the formation of cgRNA and trigger RNA complexes between non-cognate pairs was penalized in a crosstalk tube while the formation of complexes between cognate pairs was rewarded in an on-target test tube (see Figure 2.11). The goal of the optimization was to minimize the multi-tube ensemble defect.

Experimental testing of a set of orthogonal cgRNA and trigger RNA sequences for the toehold switch and terminator switch systems confirmed that NUPACK can design small cross-compatible sets of riboregulators (see Figure 2.3 and 2.6). The terminator switch had better orthogonality compared to the toehold switch, in qualitative agreement with NUPACK predictions (see Figure 2.12).

2.4 Conclusions

By carefully exploring the structure function of *S. pyogenes* guide RNA, computational tools such NUPACK were successfully used to engineer guide RNAs that had the novel property of being able to sense other RNA sequences. Follow-up testing in a mammalian cell based CRISPR activation assay, confirmed the terminator switch

---

[‡]Metabolic effects are often used as a catch all phrase to describe confounding measurements in cell based assays. These effects often occur when a synthetic gene is over expressed and/or become unpredictably toxic to the host cell.

[§]The arrangement of genetic parts on a plasmid can also cause this toxicity by inhibiting the production of antibiotic resistance genes or interfering with replication of the plasmid. Predicting if a particular gene arrangement is causing metabolic effects is difficult.
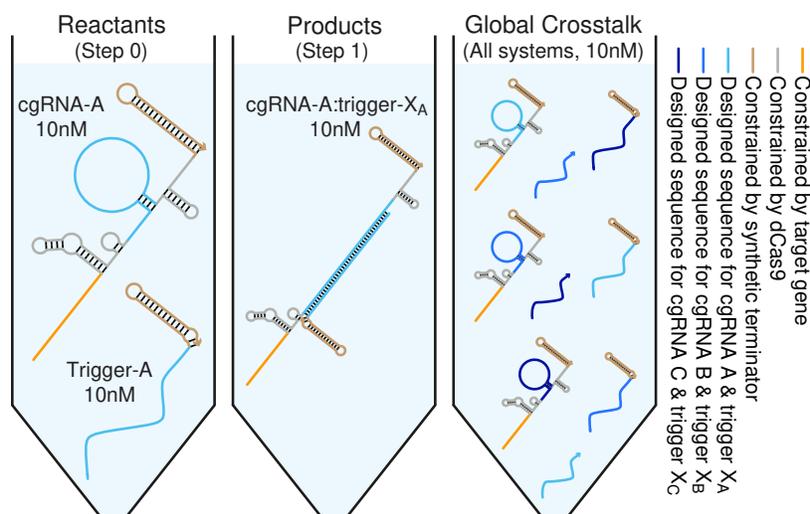
Figure 2.11: Illustration of the test tube design scheme for the terminator switch using NUPACK. a) Target test tubes for design of 3 orthogonal cgRNAs A, B, C (see Figure 2.6). Left: Elementary step tubes. Reactants tube (Step 0): cgRNA and trigger. Products tube (Step 1): cgRNA:trigger complex. Each target test tube contains a set of desired "on-target" complexes (each with the depicted target secondary structure and a target concentration of 10 nM) corresponding to the on-pathway hybridization products for a given step and a set of undesired "off-target" complexes (all complexes of up to 2 strands, each with a target concentration of 0 nM; not depicted) corresponding to on-pathway reactants and off-pathway hybridization crosstalk for a given step. To design 3 orthogonal systems, there are two elementary step tubes for each system A, B, C. Right: Global crosstalk tube. Contains the depicted on-target complexes corresponding to reactive species generated during Steps 0 and 1 as well as off-target complexes corresponding to off-pathway interactions between these reactive species. To design 3 orthogonal systems, the global crosstalk tube contains a set of on-targets and off-targets for each system A, B, C. Figures adapted from Hanewich-Hollatz *et al.* [47].

system was also functional and orthogonal in eukaryotic cells.[47]

Confirmation that cgRNAs also worked in mammalian cell systems was not easy. The mammalian cell systems initially had poorer performance compared to the bacterial systems. RNA degradation in the mammalian cell system turned out to be the limiting factor. This shortcoming was eventually addressed through the use of pseudoknotted xrRNAs structures that inhibited RNA degradation. Expressing the same trigger RNAs with 5′-xrRNAs improved the performance of the terminator switch system[66], enhancing the position of cgRNAs as a versatile platform for implementing multi-species wide conditional RNAi.

The results of this chapter are important because they showed that orthogonal sets of riboregulators can be engineered *in silico* with NUPACK and ported from bacteria to eukaryotic cells. Previously these genetic parts and sensors had to be
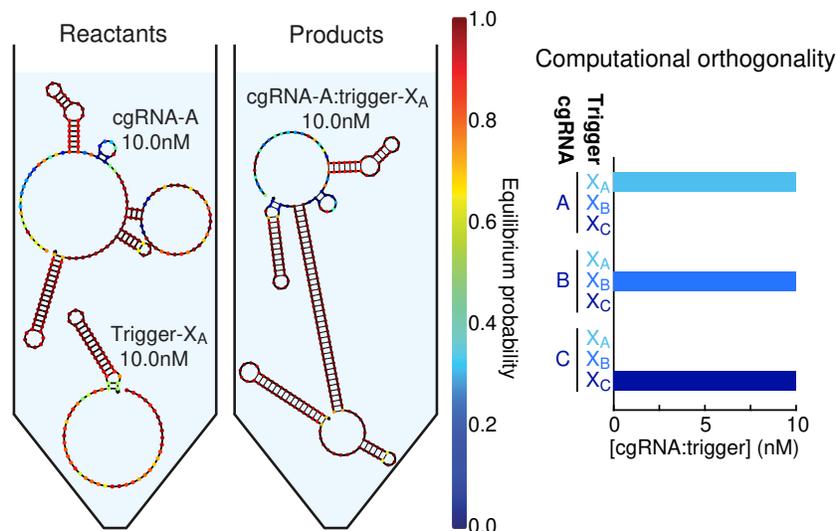
Figure 2.12: Illustration of the test tube analysis for the terminator switch using NUPACK. Left: Tubes depict the target structure and predicted concentration for each on-target complex with nucleotides shaded to indicate the probability of adopting the depicted base-pairing state at equilibrium. For this design, all on-targets are predicted to form with quantitative yield at the 10 nM target concentration but some nucleotides have unwanted base-pairing interactions (nucleotides not shaded dark red). Right: Computational orthogonality study. Predicted equilibrium concentration of each cgRNA:trigger complex for the 3 orthogonal systems of Figure 2.6. Analysis was performed with RNA at 37°C in 1M Na⁺. Figures adapted from Hanewich-Hollatz *et al.* [47].

laboriously mined from the genome. In some situations, mined components only worked in bacteria or only worked in mammalian cells. Regulatory elements such as promoters, enhancers, and repressor proteins usually are not cross-compatible between bacteria and mammalian cells. It was extremely lucky that cgRNAs worked in both bacteria and eukaryotic cells with minimal modifications. Moving forward, I focused on prototyping new cgRNA mechanisms in the bacterial setting with the assumption that they could eventually translate well into the eukaryotic setting. Prototyping cgRNAs in bacteria was preferred over developing them in mammalian cell systems because bacteria culture was easier to maintain than mammalian cell culture. Finally, although the performance of the toehold switch and terminator switch was impressive, they were not ideal for endogenous sequence detection because of burdensome design constraints. The work of the next chapter focuses on alleviating these constraints and making cgRNA designs more amenable for endogenous sequence detection.

C h a p t e r   3

ENGINEERING CGRNA SYSTEMS WITH REDUCED DESIGN CONSTRAINTS

The first generation of cgRNAs suffered from lack of sequence or structure independence. These designs were not suitable for detecting endogenous RNA sequences because overly constrained and poorly performing designs limit the scope of endogenous trigger RNAs which can be accepted as input for cgRNAs. Follow up work on cgRNAs focused on optimizing cgRNA domain dimensions to improve fold change performance and/or creating new cgRNA mechanisms which have fewer sequence and structure requirements. The following sections describe the strategies used to find these new mechanisms, development of a medium-throughput screening assay for evaluating a plethora of new cgRNA systems, and the results of the medium throughput screen.

3.1   Strategies to improve guide RNA performance

Initial efforts to enhance the performance of first generation cgRNAs focused on two approaches. The first was fragmentation of the guide RNA into multiple sub-components which were reconstituted into an active guide RNA. The second was dimensional optimization to map out the range of structural rearrangements possible while preserving cgRNA function. The aim of these approaches was to identify drivers of cgRNA performance and to reduce the number of design constraints imposed on the toehold switch and terminator switch systems.

The first strategy focused on splitting the guide RNA into multiple sub-components and conditionally reconstituting them into an active guide RNA. This idea originates from the observation that native *S. pyogenes* Cas9 guide RNAs are composed of a tracer RNA and crRNA. The crRNA contains the targeting domain and the first half of the Cas9 handle. This sub-component is generated from cleavage of the CRISPR array, which is a cassette of multiple targeting domains. The tracer RNA contains the terminator stems and the 3′ half of the Cas9 handle. These components must come together to form the active guide RNA, otherwise the system is inactive. Contemporary CRISPR/Cas9 guide RNAs improve the activity of native guide RNAs by physically joining the tracer RNA and crRNAs at the Cas9 handle loop[48,70]. The converse strategy could be useful for switching the guide RNA into an inactive

state.

In subsequent experiments, the guide RNA was split up at the Cas9 handle loop domain and at the terminator stem loop domain. Strategies to split the guide RNA along the Cas9 handle were tried, but all systems tested were unsuccessful. In computational design, NUPACK struggled to predict the formation of the Cas9 handle structure from two nucleic acid strands. The Cas9 handle is held together by several G-U wobble base pairs, which were not very stable interactions by themselves (see Figure A.2). Cas9 binding was necessary to stabilize this secondary structure. The energy contributions from protein binding were not insignificant, but these parameters could not be accounted for in the current secondary structure model underlying NUPACK. This made it challenging to generate designs which conditionally broke the Cas9 handle (see Figure 2.4 and 2.5).

Unlike the Cas9 handle, the terminator stem was more amenable towards redesign into a split guide RNA. Formation of the terminator stem was known to be important for guide RNA activity (see Figure 2.7). This structure was not sequence constrained (see Figure 2.6), which made it easier to select stem length dimensions and sequences that favored formation of the terminator stem from two nucleic acid strands. The crystal structure of Cas9 and guide RNA showed the terminator stem motif lies on the exterior of the Cas9 protein. This suggests that protein interactions were not important in stabilizing the formation of the stem loop structure, and the energies from protein binding could be neglected in the design of these structures. These factors made the terminator stem domain a more ideal location for conditionally splitting the guide RNA. Redesign of the terminator switch in this manner led to two new cgRNA systems termed the split terminator switch[66] (see Figure 3.6) and the reverse split terminator switch (see Figure 3.8).

The second strategy focused on dimensional optimization of the terminator switch structure to improve cgRNA performance. cgRNA performance is defined as the fold change ON/OFF signal ratio between active and inactive cgRNA conformations induced by a trigger RNA. Longer terminator stems improved cgRNA ON state activity, but inhibited trigger RNA binding. This means cgRNAs could not be switched OFF. Longer loops improved trigger RNA binding, but frequently led to non-functional cgRNAs. This means cgRNAs could not be turned ON. Many dimensional combinations were tried, but a more ideal balance between stem (4nt) and loop (30nt) dimensions was not found.

Despite these setbacks, the terminator stem-loop optimization experiments did reveal
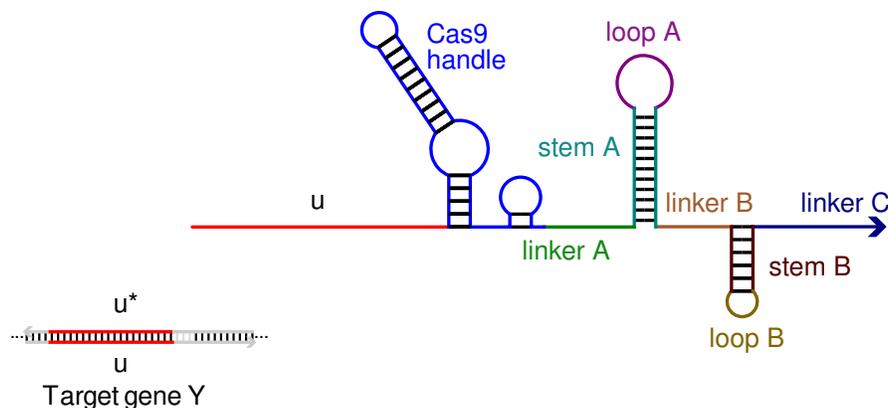
Figure 3.1: Schematic of the guide RNA stem domains being modified in Table 3.1

Name	linker a	stem a	loop a	linker b	stem b	loop b	linker c
ts10n4	6	10	4	5	10	3	7
ts10n5	6	10	4	10	10	3	7
ts10n6	6	10	4	1	10	3	1
Positive Control	6	4	4	1	6	3	7

Table 3.1: Table of dimensions for the synthetic guide RNAs with non-canonical terminator stems listed in Table A.10. The dimensions of the canonical guide RNA are on the bottom row labeled as the positive control. The sequence of the positive control is in Table A.2.
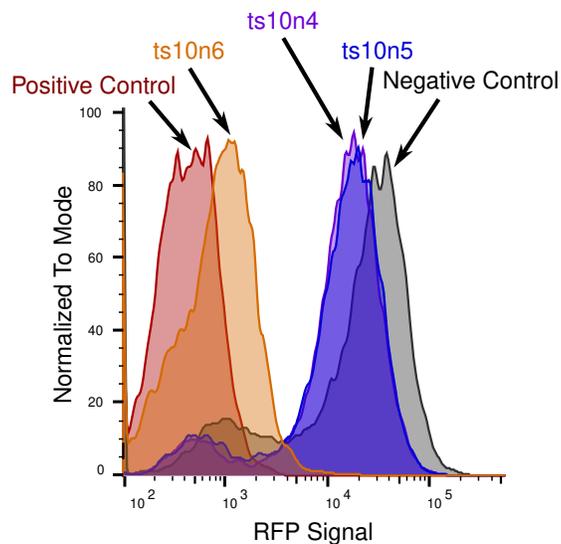
Figure 3.2: Flow cytometry plot of RFP silencing by terminator switch constructs listed in Table 3.1 and Table A.10. ts10n6 has a non-canonical terminator stem similar in shape to the canonical guide RNA and was able to silence RFP. ts10n4 and ts10n5 could not silence RFP because their terminator stem structures deviated too much from the canonical guide RNA. Data from scRNA_Cas9_20210805ZC.

two important insights about the design constraints imposed on the terminator switch. First, the entire sequence 3' of the Cas9 handle domain is unconstrained and amenable to change. Second, these sequences can be anything that have a high probability of forming double stem-loops similar in shape to the terminators on the canonical guide RNA (see Figure 2.1 and A.2). Table A.10 shows three synthetic guide RNAs with non-canonical sequences in the linker, stem, and loop regions of the terminator domains depicted in Figure 3.1. These three guide RNAs also have non-canonical terminator stem and loop lengths (see Table 3.1). Out of the three guide RNAs shown in Table A.10, one retained activity and silenced RFP as strongly as the positive control guide RNA. This guide RNA had non-canonical terminator domain dimensions and sequences (see Figure 3.2). The other two guide RNAs were non-functional like the negative control guide RNA which lacked the RFP targeting domain. These guide RNAs could not silence RFP because linker $|b|$ length was too long. Their terminator structure deviated too much from the canonical guide RNA shape (see Figure 2.1 and A.2).

The terminator stem-loop experiments showed that the terminator shape was an important driver of guide RNA activity. Previous cgRNA designs fixed the sequence of the last terminator stem-loop while changing the sequence and shape of the first stem loop (see Figure 2.6). This over constrained the terminator switch design because modifications to other regions of the cgRNA unintentionally broke the last terminator stem-loop, resulting in non-functional cgRNAs (see Figure A.3). For next-generation designs, the terminator domain was changed from a sequence constraint to a structural constraint to allow the generation of a non-canonical stem-loop sequence for each cgRNA being designed. This modification reduced the ensemble defect of each design, improving the likelihood of getting functional guide RNAs.

3.2 High performing conditional guide RNA systems

This section provides a short list of high performing conditional guide RNA systems discovered so far. These systems were invented using insights gained from the guide RNA structure function studies conducted in the previous section.

Reverse toehold switch guide RNA

The expanded adaptability of the terminator stems led to the development of the reverse toehold switch (see Figure 3.3). This system was designed to switch from OFF to ON and improved on the original toehold switch (see Figure 2.3) by seques-

tering the whole guide RNA in a hairpin via sequences appended to the 5′ and 3′ ends of the guide RNA. This design capitalized on the fact that the terminator stem and targeting domain were highly amenable to redesign. Sequences appended 5' and 3' of the guide RNA do not destroy guide RNA activity. Unlike the original toehold switch, the trigger RNA of the reverse toehold switch was not constrained by sequence or structure. Domain b simply needed to be long enough to form a hairpin that sequestered the targeting domain. Toehold a needed to be unstructured enough to enable nucleation with the trigger RNA. This configuration dramatically improved the versatility of the toehold switch design by decoupling RNA input from DNA output. This expanded the scope of endogenous RNA sequences which could be used as RNA input.
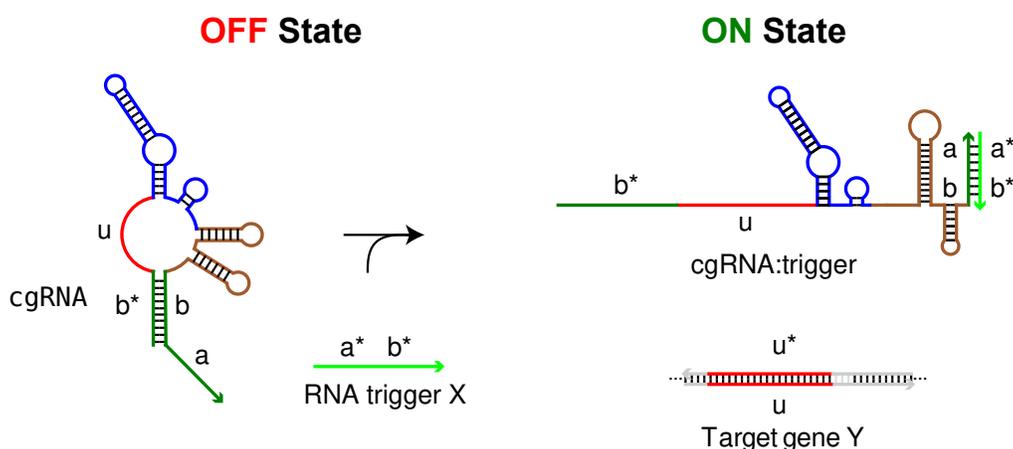
Figure 3.3: Schematic of the reverse toehold switch mechanism. cgRNA is initially held in an inactivate state by duplex b-b*. Expression of a trigger RNA binding to domain b-a opens the hairpin duplex b, enabling cgRNA binding to its DNA target u. Red denotes the targeting domain. Blue denotes the Cas9 handle. Brown denotes the terminator domains, which can be non-canonical. Green denotes the variable domains designed by NUPACK. Figure 3.4 show the experimental performance of this design.

Name	toehold a	duplex b	trigger RNA
ts45m0-1	10	20	mtrig0
ts45m0-3	10	10	mtrig0
ts45m1-1	10	20	mtrig1
ts45m2-3	10	10	mtrig2

Table 3.2: Table of dimensions for the reverse toehold switch cgRNA sequences tested in 3.5. Sequence of these cgRNAs are listed in Table A.11. Sequence of the trigger RNAs are listed in Table A.8.

Experimentally, the reverse toehold switch system is not perfect. First, the OFF state can be leaky. Over-expressing the cgRNA in some designs can lead to RFP
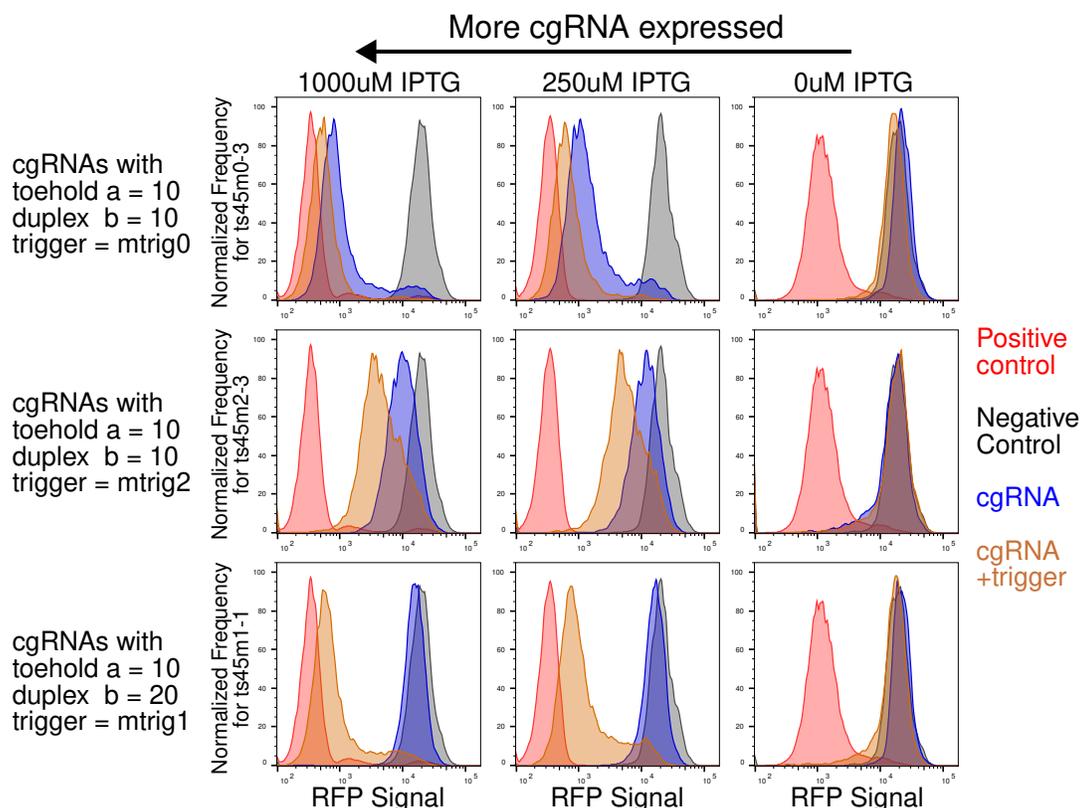
Figure 3.4: Induction of signal leak in the reverse toehold switch (see Figure 3.3). RFP silencing in bacteria by the reverse toehold switch mechanism was measured via flow cytometry. This system was designed to switch from OFF to ON (high to low fluorescence). The domain dimensions of the cgRNAs are listed in Table 3.2 and on the left axis of the histogram plots. Each row represents a different cgRNA sequence. Each column represents different amounts of IPTG used to induce expression of cgRNA in bacteria. Increasing cgRNA expression shifts RFP from high to low fluorescence. cgRNA sequences are listed in Table A.11. Trigger RNA sequences are listed in Table A.8. Data from scRNA_Cas9_20221114ZC.

silencing in absence of trigger RNA (see Figure 3.4, top row, sample ts45m0-3). Leak can be suppressed by lengthening duplex domain $|b|$ (see Figure 3.5). Domain b is responsible for holding the reverse toehold switch in an inactive state. Longer duplexes are expected to better stabilize the inactive conformation. This is supported experimentally by ts45m0-1 and ts45m0-3. These two cgRNAs share the same trigger sequences, but differ in their domain $|b|$ lengths. ts45m0-1 has a domain $|b|$ length of 10nt, which failed to inactivate the cgRNA. ts45m0-1 has a domain $|b|$ length of 20nt, which successfully held the cgRNA in an inactive state. Unfortunately, the OFF state could not be reversed by the trigger RNA for the ts45m0-1 sequence. Toehold domain $|a|$ could have been too short. This could be addressed by increasing the length of toehold domain $|a|$ to improve trigger
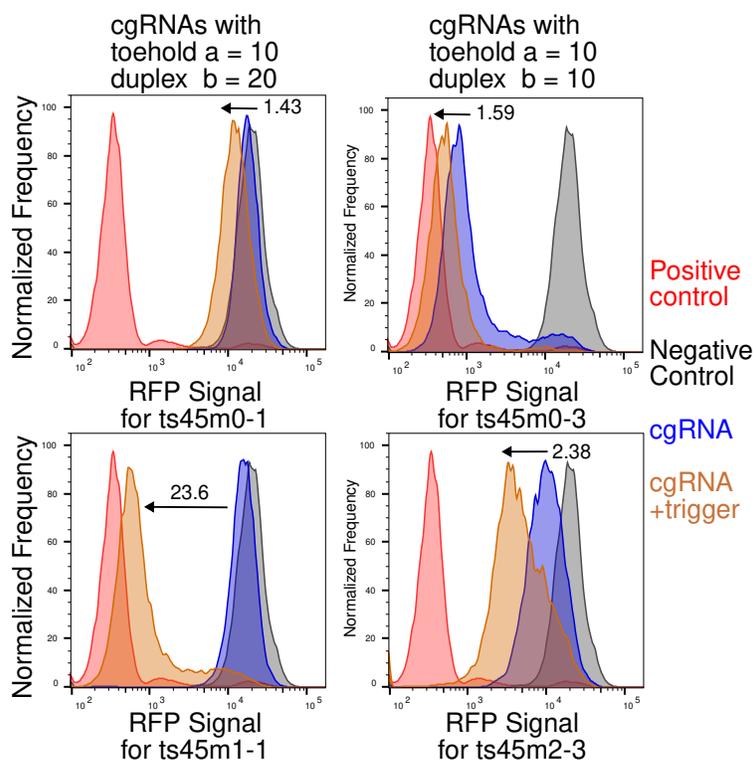
Figure 3.5: Effect of domain dimensions on reverse toehold switch performance. Flow cytometry measurement of RFP silencing in bacteria by the reverse toehold switch mechanism (see Figure 3.3) at max cgRNA induction (1000uM IPTG). This system was designed to switch from OFF to ON (high to low fluorescence). Fold change signal is denoted next to the arrow in each plot. Maximum fold change between positive and negative control is 55. cgRNA domain dimensions are listed in Table 3.2 and above each histogram. cgRNA samples within the same column share the same domain dimensions. ts45m0-1 and ts45m0-3 both use mtrig0 as the trigger RNA. ts45m1-1 uses mtrig1. ts45m2-3 uses mtrig2. cgRNA sequences are listed in Table A.11. The sequence of the trigger RNAs are listed in Table A.8. MFE structures for these sequences are shown in Figure A.4. Data from scRNA_Cas9_20221114ZC.

RNA binding to cgRNA, which increases the speed of toehold-mediated strand displacement.[71]

High performing sequences for the reverse toehold switch mechanism do exist, but generating these sequences reliably is still challenging. At max cgRNA induction (1000uM), ts45m1-1 has ON and OFF signal values which nearly match the idealized ON and OFF values demarcated by the positive and negative control guide RNAs (see Figure 3.5). The same level of performance was not achieved with ts45m0-1, which shared the same domain dimensions. ts45m0-1 could not be activated by its cognate trigger RNA. Reliably generating high performing designs is tricky because domain dimensions alone do not guarantee high performance. The free energy of

RNA duplex formation is driven by domain length and sequence content. Domain lengths which are optimal for one trigger RNA sequence will not necessarily work for another sequence. Unfortunately, test tube analysis of these cgRNA sequences was not informative in predicting this relationship (see Figure A.4). All cgRNA sequences were predicted to have good OFF state structures and could bind with their cognate trigger RNAs. These computational predictions must be taken with a grain of salt because NUPACK simulates RNA structure formation at 37°C in 1M Na$^+$. These are not necessarily the same salt conditions in bacteria. The energies of base pairing could be slightly different in a cell, making it difficult to generate high performing sequences with perfect accuracy.

NUPACK is useful for enriching candidate sequences with functional designs. Half of the reverse toehold switch sequences generated by NUPACK had some ON/OFF fold change activity (see Figure 3.5). However, computational predictions alone cannot identify high performing hits. cgRNA sequences utilizing different domain dimensions and trigger sequences must be screened in the bacterial assay. If enough samples are tested, high performing hits can be found.

Split terminator switch

The split terminator switch is a cgRNA system designed to switch from OFF to ON (high to low fluorescence in a dCas9 silencing assay). This system is a modification of the terminator switch cgRNA whereby the guide RNA is split at the terminator stem. The first half of the guide RNA termed $g1$ contains the Cas9 handle and targeting domain. The second half of the guide RNA termed $g2$ is treated as the trigger RNA. Experimental testing showed that both components are necessary for RFP silencing[66], and the cgRNA is dose responsive to trigger RNA (see Figure 3.7). The primary advantage of the split terminator switch system is high ON/OFF trigger RNA response. The fold change response of this system can span the full range of OFF and ON signal values corresponding to the positive and negative control guide RNAs. This was demonstrated in both the bacterial and mammalian cell assay settings (see Figure 3.7 and Hochrein *et al.* [66]). The discovery of this system further validates the terminator structure function studies previously discussed in Chapter 2 (see Figure 2.7).

Although strong ON/OFF response to trigger RNA is a major advantage of the split terminator switch system, the utility of the trigger RNA for endogenous sequence detection is limited because its secondary structure is somewhat constrained. As

OFF State   **ON State**
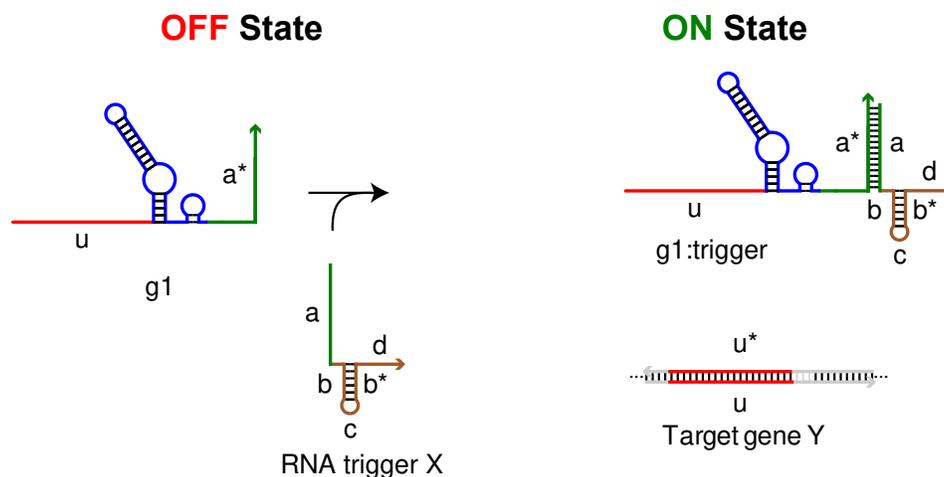
Figure 3.6: Schematic of the split terminator switch mechanism. Expression of the trigger RNA enables the formation of a fully functioning guide RNA complex. Red denotes the targeting domain. Blue denotes the Cas9 handle. Brown denotes the terminator domains, which can be non-canonical. Green denotes the variable length stem domains designed by NUPACK. Figure 3.7 shows the experimental performance of this design.
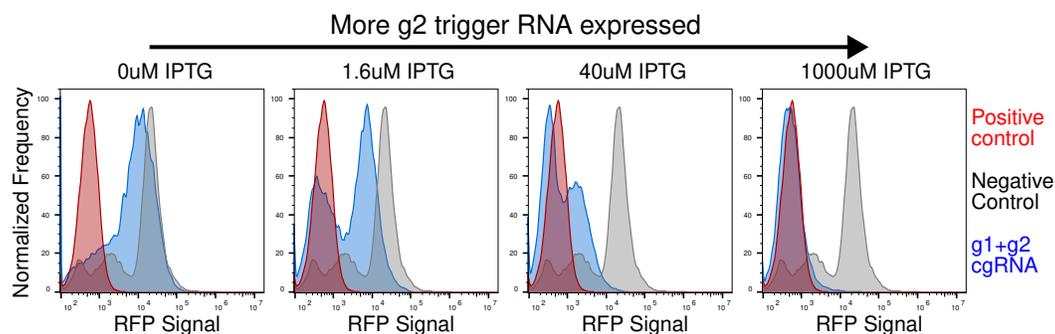
Figure 3.7: Flow cytometry measurement of RFP silencing in bacteria by the split terminator switch (see Figure 3.6). This system was designed to switch from OFF to ON (high to low fluorescence). cgRNA expression was constitutive. Trigger RNA was expressed on a pLac promoter and induced by the addition of IPTG. The bacterial population exhibited a bimodal shift from high to low RFP fluorescence. This corresponds to a shift from low guide RNA activity to high activity. The cgRNA and trigger RNA sequences are listed in Table A.12. Data from scRNA_Cas9_20210427ZC.

noted in the stem-loop optimization study (see Figure 3.2), the terminator domain can be any sequence that forms a double stem-loop. This means the trigger RNA must contain a hairpin of duplex length $|b|$ and loop length $|c|$ (see Figure 3.6). Although these domains dimensions are not fixed, the scope of short endogenous RNA sequences suitable for use as trigger RNA input could be limited (see Figure A.8 and A.9).

Reverse split terminator switch

The reverse split terminator switch is designed to switch from ON to OFF . This system is derived from the original terminator switch where the trigger RNA sequesters one half of the split guide RNA via base pairing with a toehold c and nucleating down the terminator stem b (see Figure 3.8 and 3.9). This improves on the original terminator switch by turning pseudoknot infiltration of the terminator loop into a strand displacement process which is kinetically faster and simpler to design.
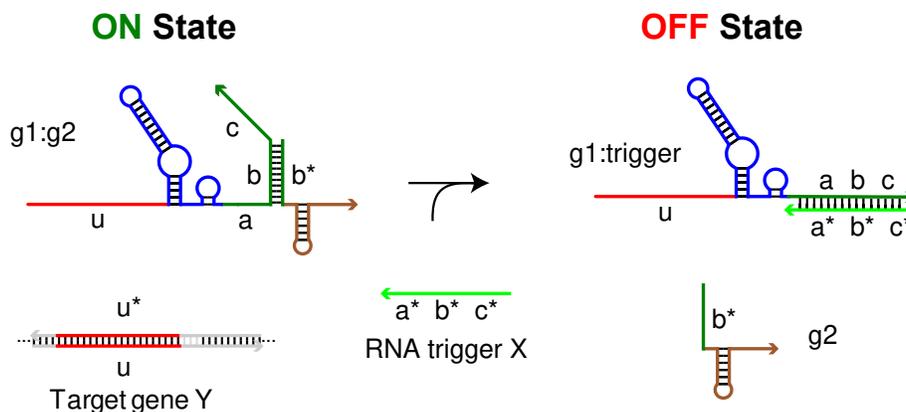
Figure 3.8: Schematic of the reverse split terminator switch mechanism. $g1$ and $g2$ halves of a functioning guide RNA are constitutively expressed. Induction of the trigger RNA sequesters $g1$, breaking the guide RNA at the terminator stem b. Red denotes the targeting domain. Blue denotes the Cas9 handle. Brown denotes the terminator domains, which can be non-canonical. Green denotes the variable length stem domains designed by NUPACK. Figure 3.9 shows the experimental performance of this design.

The main advantage of this design is the improved computational control over cgRNA behavior such as ON/OFF activity and sensitivity to the trigger RNA. cgRNA ON state activity is driven by the formation of the terminator stem from the two halves of the guide RNA. This is influenced by the length of the terminator stem and by the concentration of the $g2$ guide RNA present in the cell (see Figure 3.8 and 3.10). Longer domain b increases the affinity between $g1$ and $g2$. This is counteracted by toehold domain c on $g1$, which increases the affinity between $g1$ and trigger RNA. Trigger RNA binding sequesters the guide RNA, and longer toeholds are hypothesized to drive the system more strongly towards the OFF state, improving ON/OFF performance.[71,72] Having ways to tune cgRNA performance is important in synthetic gene circuit engineering because the relative strengths of each component can change the dynamic behavior of the circuit[34–37,73–75]. The range of possible circuit behaviors is limited by the performance characteristics of the available components. This cgRNA architecture is valuable for alleviating some

of these limitations, opening the possibility of testing genetic circuit architectures previously thought to be impractical.
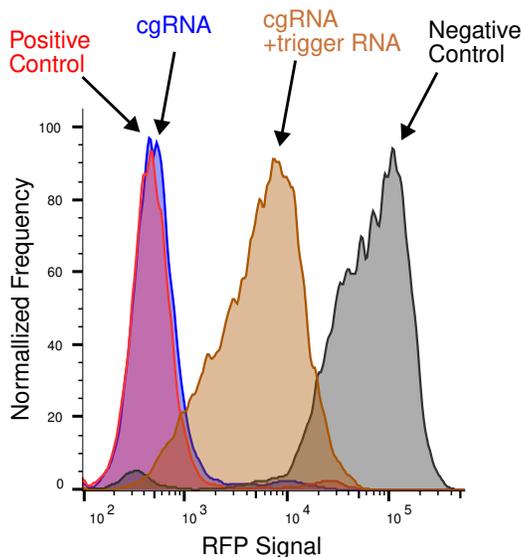
Figure 3.9: Flow cytometry measurement of RFP silencing with dCas9 in bacteria by the reverse split terminator switch (see Figure 3.8). This was designed to work in the ON to OFF logic (low to high RFP fluorescence). Constitutive expression of cgRNA leads to full RFP silencing. Co-expression of the trigger RNA sequesters the cgRNA, leading to increased RFP fluorescence. cgRNA and trigger sequences are listed in Table A.13. Data from scRNA_Cas9_20220923ZC.
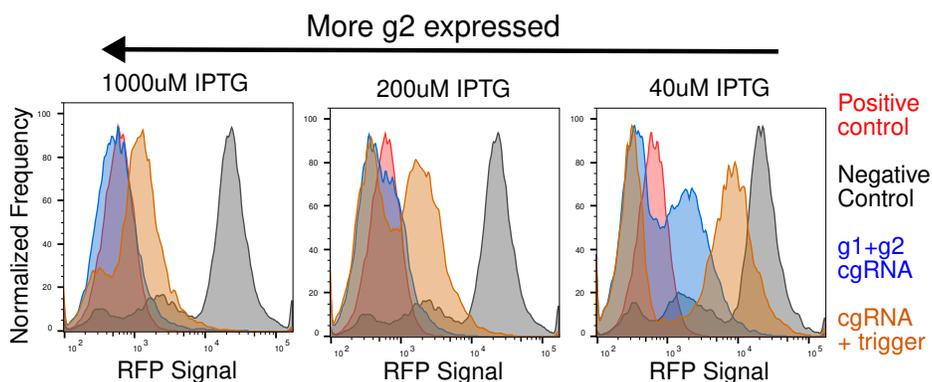
Figure 3.10: Allosteric competition between g2 and trigger RNA for the reverse split terminator switch. Flow cytometry was used to measure RFP silencing with dCas9 in bacteria by the reverse split terminator switch (see Figure 3.8). This was designed to work in the ON to OFF logic (low to high RFP fluorescence). Constitutive expression of cgRNA leads to RFP silencing. Co-expression of the trigger RNA sequesters the cgRNA, leading to increased RFP fluorescence. Increasing g2 expression shifts cgRNA towards the ON state, decreasing RFP signal. cgRNA and trigger sequences are listed in Table A.12. Data from scRNA_Cas9_20210427ZC.
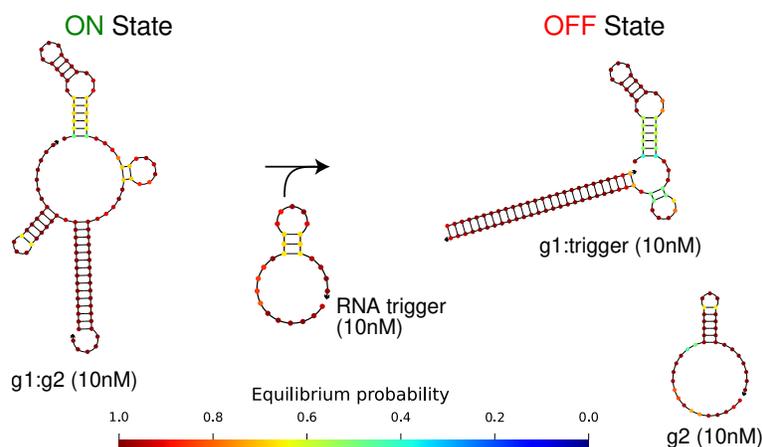
Figure 3.11: NUPACK test tube analysis of the reverse split terminator switch system in bacteria (see Figure 3.8 and 3.9). The minimum free energy (MFE) structure represents the most probable RNA structure for the given sequence in the thermodynamic ensemble. Red means base pairs are highly likely to form. Blue means the base pairs are unlikely to form. Concentrations of reactants and products for each step of the reaction are denoted next the structure label. RNA was simulated at 37°C in 1M Na$^+$.

Experimentally, over expressing $g2$ relative to the trigger RNA is detrimental to ON/OFF fold change activity because trigger RNA competes with $g2$ for $g1$ (see Figure 3.10). This phenomena may not be reflected in NUPACK predictions of test tube concentrations at equilibrium for two reasons (see Figure 3.11). First, the cellular environment may not be at equilibrium. Trigger RNA and cgRNA are constantly being produced and degraded. Second, dCas9 binding to the active $g1$:$g2$ complex could be a irreversible process. This would suggest that the kinetics of cgRNA to trigger RNA binding is more useful for determining cgRNA performance than equilibrium concentration information. The exact nature of these processes requires further study, and the reverse split terminator switch system maybe an interesting platform for exploring the biophysics of these synthetic allosteric regulatory interactions.[76] These future studies could be important for calibrating the base pairing energies and kinetics of nucleic acid strand displacement *in vivo*. The information gleaned from these studies would be useful for refining or augmenting the secondary structure models underlying NUPACK, enabling more accurate modeling of RNA strand displacement inside cells.

Single-hairpin split cgRNA

The single-hairpin split cgRNA is designed to switch from OFF to ON. This system inverts the logic of the reverse split terminator switch by sequestering $g1$ in a hairpin (see Figure 3.12). This prevents $g2$ from binding to $g1$ resulting in a constitutively

inactive state. Co-expression of the trigger RNA opens up the sequestering hairpin via toehold mediated strand displacement. This allows $g1$ to bind to $g2$, reforming into an active guide RNA complex. Creation of this system was motivated by the fact that the fluorescence change of ON to OFF systems was difficult to detect because the trigger RNA must be expressed in excess of the cgRNA. In dCas9 based RFP silencing assays, minuscule changes in constitutive cgRNA activity can be masked by a high fluorescent background. OFF to ON systems are easier to detect in dCas9 based silencing assays because the fluorescence signal change occurs above the background signal. This means less sensitive detectors can be used.
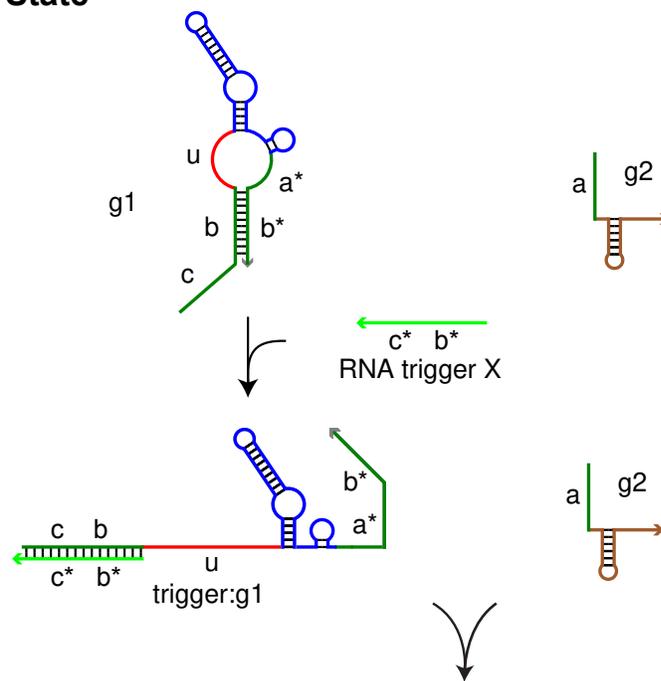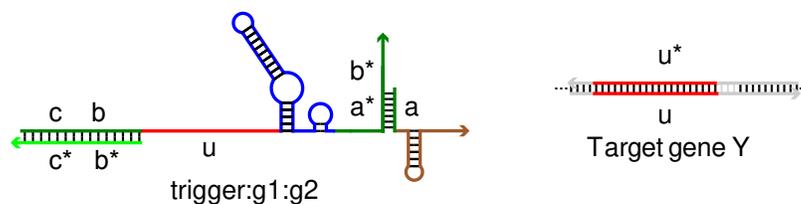
Figure 3.12: Illustration of a single-hairpin split cgRNA. $g1$ is initially held in a hairpin, which prevents binding with $g2$. Trigger RNA binding to domain C-B opens up the hairpin. $g1$ and $g2$ can now reform into an active guide RNA complex. Red denotes the targeting domain. Blue denotes the Cas9 handle. Brown denotes the terminator domains, which can be non-canonical. Green denotes the variable length stem domains designed by NUPACK. Figure 3.13 shows the experimental performance of this design.

Like the reverse toehold switch, the OFF state of the single-hairpin split cgRNA

Name	stem a	duplex b	toehold c	trigger RNA
0ts18b	7	20	10	mtrig0
28ts18b	7	20	10	mtrig2
29ts18b	7	20	5	mtrig2

Table 3.3: Table of dimensions for the single-hairpin split cgRNA sequences tested in 3.13. cgRNA sequences are listed in Table A.14. Trigger RNA sequences are listed in Table A.8.
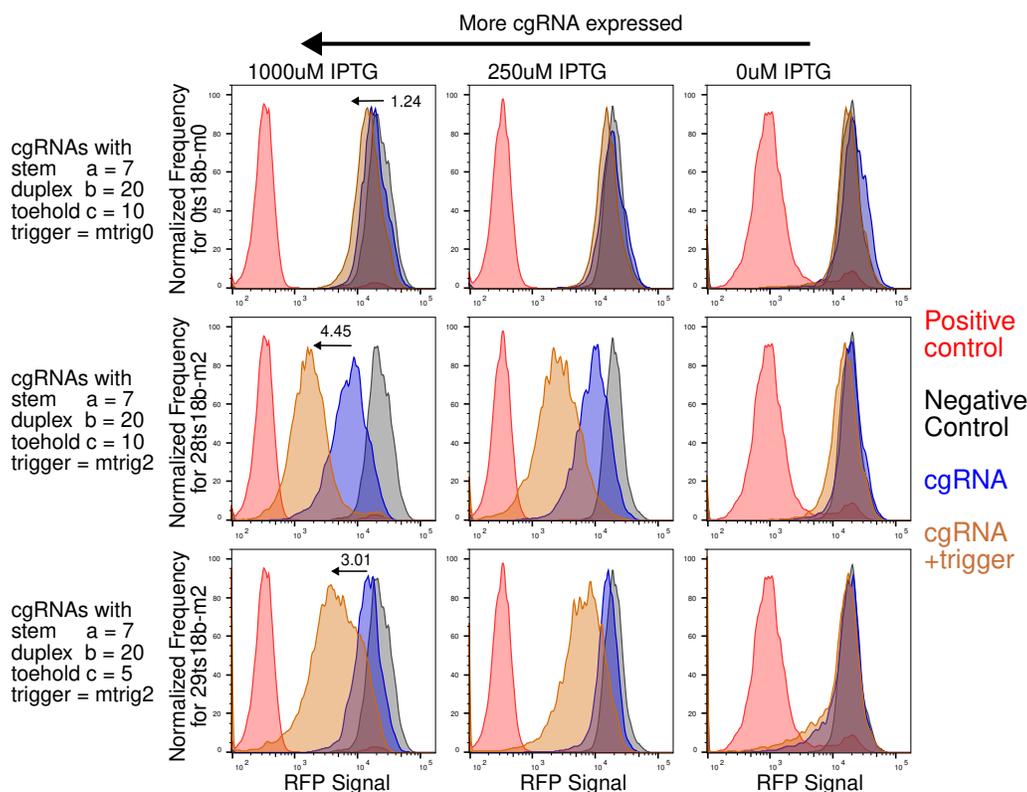
Figure 3.13: Flow cytometry measurement of RFP silencing in bacteria by the single-hairpin split cgRNA system (see Figure 3.12). This was designed to work in the OFF to ON logic (high to low RFP fluorescence). Fold change signal is denoted next to the arrow in each plot. Maximum fold change between positive and negative control is 45. The domain dimensions of the cgRNAs are listed in Table 3.3 and on the left axis of the histogram plots. Each row represents a different cgRNA sequence. Each column represents different amounts of IPTG used to induce expression of cgRNA in bacteria. Increasing cgRNA expression shifts RFP from high to low fluorescence. cgRNA sequences are listed in Table A.14. Trigger RNA sequences are listed in Table A.8. Data from scRNA_Cas9_20221114ZC.

can also be leaky. Over-expressing the cgRNA can lead to increased RFP silencing in the absence of trigger RNA (see Figure 3.13, sample 28ts18b-m2). Leak can be suppressed by lengthening duplex domain $|b|$. This domain is responsible for holding the cgRNA in an inactive state. Longer duplexes are expected to better stabilize the inactive conformation. Conversely, ON state activity maybe improved

by adjusting the length of toehold domain $|c|$, which increases the probability of trigger RNA binding to cgRNA. An additional feature of this system is the ability to adjust the length of domain $|a|$, which control the affinity between $g1$ and $g2$. Increasing the length of domain $|a|$ can exacerbate signal leak because $g1$ and $g2$ are more likely to come together in absence of a trigger RNA. Reducing the length of domain $|a|$ or reducing the expression of $g2$ can help mitigate signal leak. However, if the length of domain $|a|$ is too short, $g1$ and $g2$ will not bind, meaning the cgRNA cannot be turned on.

High performing sequences for the reverse toehold switch mechanism can exist, but generating these sequences reliably is challenging. At max cgRNA induction (1000uM), all cgRNA sequences tested were able to maintain good OFF states, meaning the hairpin duplex successfully suppressed $g1$ and $g2$ binding. 28ts18b-m2 and 29ts18b-m2 exhibited measurable fold change values, while 0ts18b-m0 could not be activated by its trigger RNA (see Figure 3.13). 0ts18b-m0 and 28ts18b-m2 shared the same domain dimensions, but had different ON/OFF performance because they sensed different trigger RNAs. This suggests that domain dimensions alone do not guarantee high performance. The free energy of RNA duplex formation is driven by domain length and sequence content. Domain lengths which are optimal for one trigger RNA sequence will not necessarily work for another sequence. Unfortunately, test tube analysis of these cgRNA sequences was not informative in predicting this relationship. All cgRNA sequences were predicted to have good OFF state structures and could bind with their cognate trigger RNAs. These computational predictions must be taken with a grain of salt because NUPACK simulates RNA structure formation at 37°C in 1M Na$^+$. These are not necessarily the same salt conditions in a bacteria. The energies of base pairing could be slightly different in a cell, making it difficult to generate high performing sequences with perfect accuracy.

NUPACK is useful for enriching candidate sequences with functional designs. More than half of the cgRNA sequences generated by NUPACK had some ON/OFF fold change activity (see Figure 3.13). However, computational predictions alone cannot identify high performing hits. cgRNA sequences utilizing different domain dimensions and trigger sequences must be screened in the bacterial assay. If enough samples are tested, high performing hits can be found.

Two-hairpin split cgRNA

The two-hairpin split cgRNA is designed to switch from OFF to ON. This system conceptually improves on the single-hairpin split cgRNA by sequestering both halves of the guide RNA into a hairpin (see Figure 3.14). This engineering strategy was expected to reduce the leakiness of the OFF state via redundancy in the hairpin sequestration mechanism. A notable feature of this system is that it can potentially be catalytic. If domain d is made to equal domain c*, the system regenerates the trigger RNA domain b*-c* upon formation of the active guide RNA complex. Additional strands of $g1$ can bind to this triggering domain and form additional guide RNAs. The nucleation and formation of a chain of cgRNAs from a single trigger RNA is analogous to hybridization chain reaction (HCR) amplification technology.[77] Like HCR, the catalyic feature of the two-hairpin split cgRNA system could potentially improve the sensitivity of cgRNA to trigger RNA via non-enzymatic amplification of the trigger RNA signal. However, this did not work out in practice (see Figure 3.15).

Like the reverse toehold switch and single-hairpin split cgRNA, the two-hairpin split cgRNA also suffers from a leaky OFF state. Over-expressing the cgRNA can lead to RFP silencing in the absence of trigger RNA (see Figure 3.15, third row, sample 3ts23b-m0). Leak can be suppressed by lengthening duplex domain b. This domain is responsible for holding both hairpins in an inactive state. Longer duplexes are expected to better stabilize the inactive conformation. Conversely, ON state activity maybe improved by adjusting the length of toeholds $|c|$ and $|a|$, which increase the probability of trigger RNA binding to cgRNA and $g2$ binding to $g1$ respectively. Increasing the length of domain $|a|$ too much can exacerbate signal leak because $g1$ and $g2$ are more likely to come together. Decreasing the length of domain $|a|$ or reducing the expression of $g2$ can help mitigate signal leak. However, if the length of domain $|a|$ is too short, $g1$ and $g2$ will not bind, meaning the cgRNA cannot be turned on. Notably, sequestering both $g1$ and $g2$ in a hairpin did not seem to significantly reduce signal leak. The catalytic version of this system had bad OFF states possibly because the problems associated with signal leak were exacerbated by auto-catalysis (see Figure 3.16).

High performing sequences for the reverse toehold switch mechanism can exist, but generating these sequences reliably is still challenging. At max cgRNA induction (1000uM), four out of the five cgRNA sequences tested were able to maintain good OFF states, meaning the hairpin duplex successfully suppressed $g1$ and $g2$

OFF State

ON State
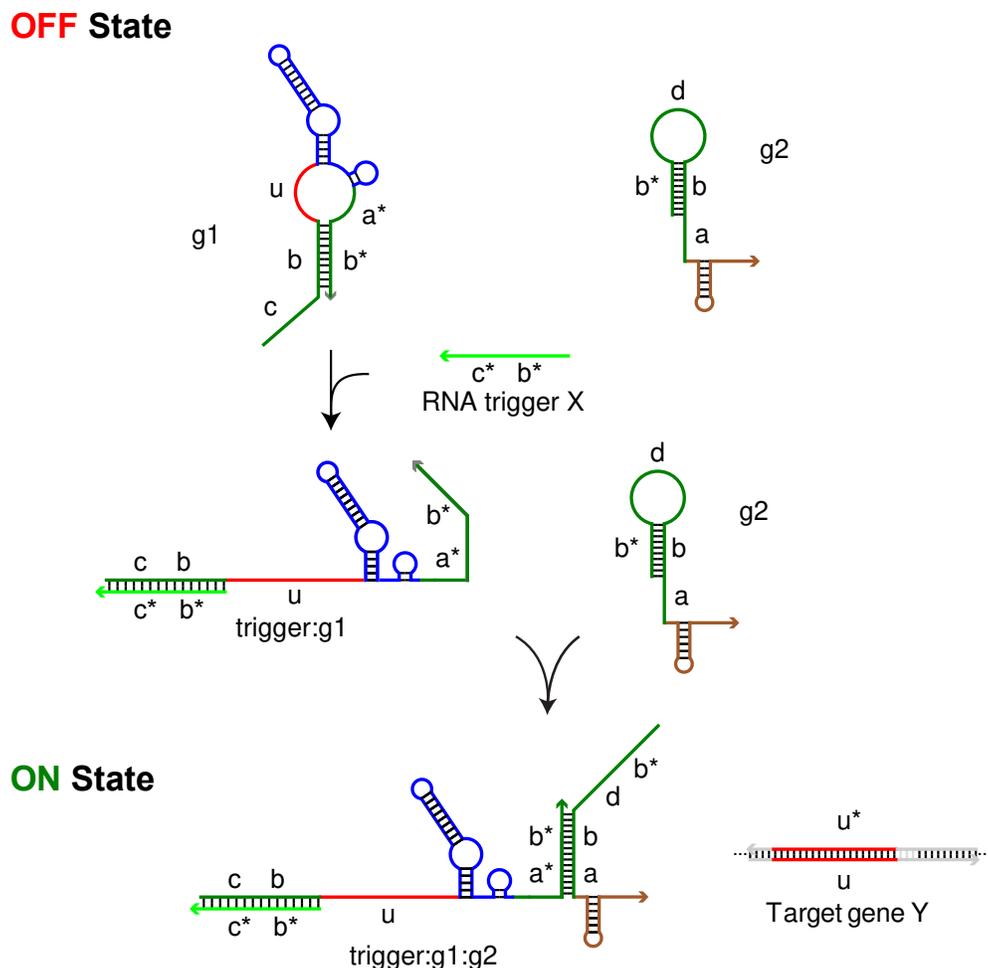
Figure 3.14: Illustration of the two-hairpin split cgRNA mechanism. $g1$ is and $g2$ are initially sequestered by a duplex domain b. Trigger RNA opens up the first duplex via binding to domain c-b. This enables toehold mediated strand displacement of domain b-a on $g2$ by the trigger:$g1$ complex, forming an active guide RNA. Red denotes the targeting domain. Blue denotes the Cas9 handle. Brown denotes the terminator domains, which can be non-canonical. Green denotes the variable length stem domains designed by NUPACK. Figure 3.15 shows the experimental performance of this design.

binding. Of the cgRNA sequences which could maintain good OFF states, 33ts23b-m2 and 30ts23b-m2 exhibited measurable fold change values (see Figure 3.15, second and fifth row), while 19ts23b-m1 and 31ts23b-m2 could not be activated by their respective trigger RNAs (see Figure 3.15, first and fourth rows). 3ts23b-m0 and 31ts23b-m2 shared the same domain dimensions, but one of them could not maintain a good OFF state. 19ts23b-m1 and 33ts23b-m2 shared the same domain dimensions, but one of them could not be switched ON by the trigger RNA. These pairs of cgRNAs each sensed different trigger RNAs. This suggests that domain dimensions alone do not guarantee high performance. The free energy of RNA
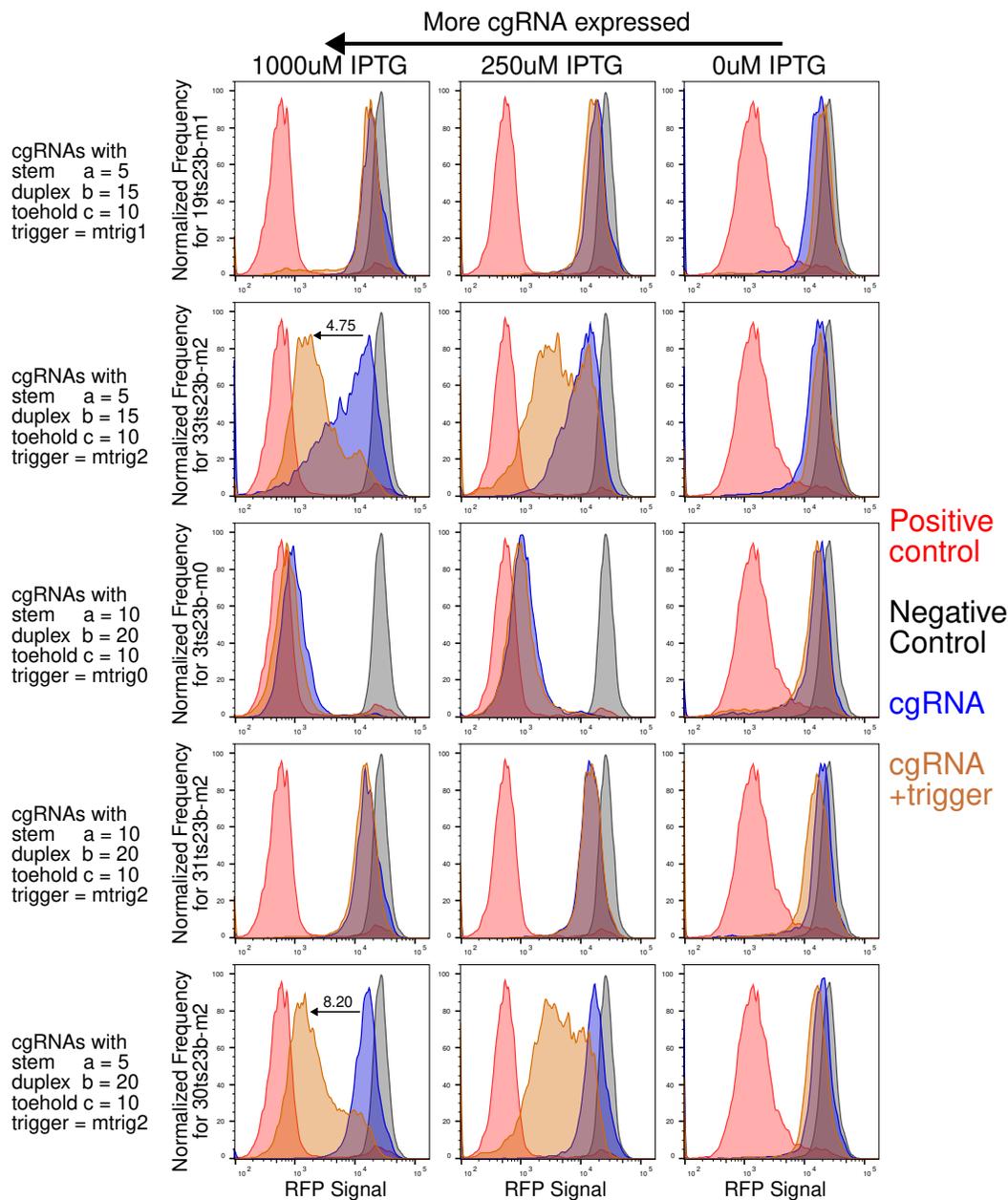
Figure 3.15: Flow cytometry measurement of RFP silencing with dCas9 in bacteria by the two-hairpin split cgRNA mechanism (see Figure 3.14). This system was designed to work in the OFF to ON logic (high to low fluorescence). Fold change signal is denoted next to the arrow in each plot. Maximum fold change between positive and negative control is 45. cgRNA domain dimensions are listed in Table A.16 and on the left axis of the histogram plots. Each row represents a different cgRNA sequence. Each column represents different amounts of IPTG used to induce expression of cgRNA in bacteria. Increasing cgRNA expression shifts RFP signal from high to low fluorescence. cgRNA sequences are listed in Table A.15. Trigger RNA sequences are listed in Table A.8. Data from scRNA_Cas9_20221114ZC.

duplex formation is driven by domain length and sequence content. Domain lengths
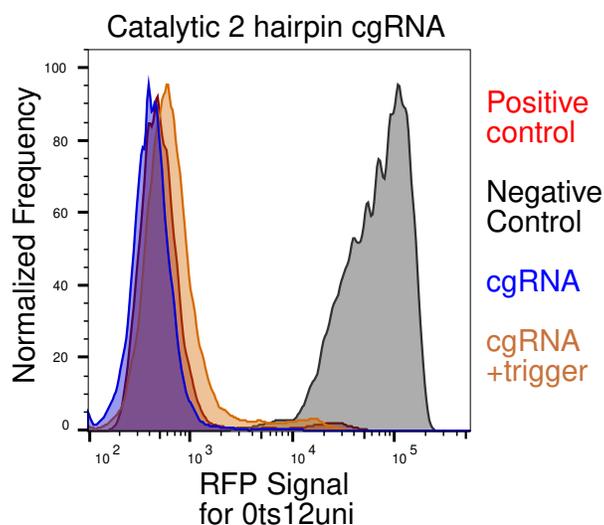
Figure 3.16: Flow cytometry measurement of RFP silencing with dCas9 in bacteria by the catalytic two-hairpin split cgRNA mechanism (see Figure 3.14). This system was designed to work in the OFF to ON logic (high to low RFP fluorescence). However, a good OFF state could not be maintained in absence of trigger RNA because the ON state was leaky. Catalytic assembly of the two hairpin system exacerbates this leak resulting in a constitutive ON state. cgRNA domain dimensions are listed in Table A.16. cgRNA and trigger RNA sequences are listed in Table A.13. Data from scRNA_Cas9_20220923ZC.

which are optimal for one trigger RNA sequence will not necessarily work for another sequence. Unfortunately, test tube analysis of these cgRNA sequences was not informative in predicting this relationship. All cgRNA sequences were predicted to have good OFF state structures and could bind with their cognate trigger RNAs. These computational predictions must be taken with a grain of salt because NUPACK simulates RNA structure formation at 37°C in 1M Na$^+$. These are not necessarily the same salt conditions in a bacteria. The energies of base pairing could be slightly different in a cell, making it difficult to generate high performing sequences with perfect accuracy.

NUPACK is useful for enriching candidate sequences with functional designs. Two out of five cgRNA sequences generated by NUPACK had some ON/OFF fold change activity (see Figure 3.15). However, computational predictions alone cannot identify high performing hits. cgRNA sequences utilizing different domain dimensions and trigger sequences must be screened in the bacterial assay. If enough samples are tested, high performing hits can be found.
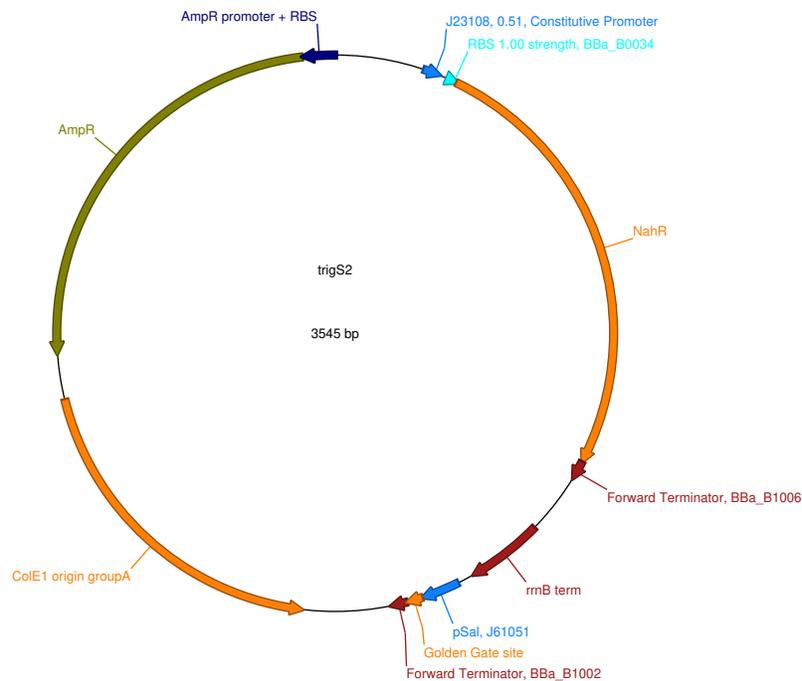
Figure 3.17: Map of the trigger RNA plasmid used in the medium throughput setup. Trigger RNA was expressed on the pSal promoter. Trigger sequences could be inserted behind the pSal promoter via Golden Gate or Gibson assembly. NahR was expressed on this plasmid to enable control over the pSal promoter with salicylate. AmpR gave this plasmid carbenicillin resistance. Genbank files for this plasmid are located in section A.8.

3.3 Strategies for building higher throughput cgRNA screening assays

This section describes the medium throughput cgRNA screening assay used to evaluate the next-generation cgRNA mechanisms discussed above. The evaluating the plethora of new cgRNA systems and sequences being generated was challenging because the assay setup of Chapter 2 was unsuitable for higher throughput screening. The cloning of trigger RNA and cgRNA could not be decoupled. Trigger RNA and cgRNA were expressed on the same plasmid, and this quadratically increased the number of plasmids needed for an orthogonality test. Orthogonality tests involved cross testing each pair of cgRNA and trigger RNA. The number of test strains required grows quadratically $O(N^2)$ with N number of orthogonal cgRNA systems being tested. A new plasmid layout was developed to reduce the cloning workload. This section describes the rationale for the choice of reporter gene, promoter, and layout of genes for the medium throughput cgRNA activity assay in bacteria. The intention of this section is to convey the key lessons learned that could be useful toward building higher throughput measurement systems for CRISPR/Cas9 activity in bacteria and other living organisms.
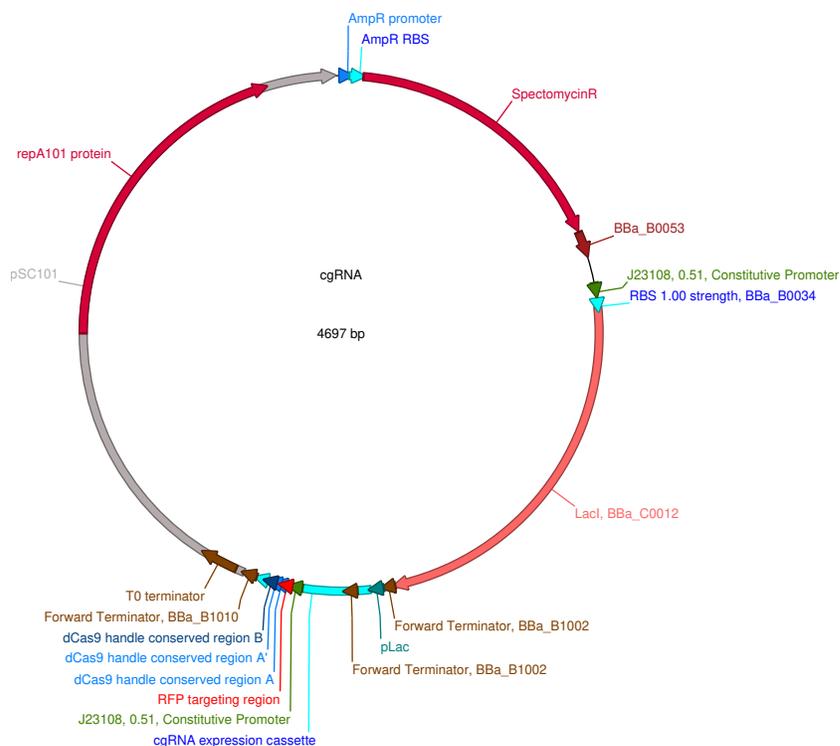
Figure 3.18: Map of the cgRNA plasmid used in the medium throughput setup. cgRNA oligos could be inserted behind the pLac promoter via Golden Gate or Gibson assembly. LacI was expressed on this plasmid to enable control over the pLac promoter with IPTG. SpectinomycinR gave this plasmid spectinomycin resistance. Genbank files for this plasmid are listed in Table A.17.

The medium throughput screening assay was designed with three goals in mind. The first goal was to decouple the cloning of trigger RNA from cgRNA. The solution was to express the trigger RNA on the carbenicillin plasmid and express cgRNAs on a spectinomycin plasmid (see Figure 3.17 and 3.18). Standardized bacterial strains for each trigger RNA could be generated independently of the cgRNA plasmids. Each trigger RNA strain was then transformed with the cgRNA plasmid to create the test strain. This simplified the generation of test strains for the orthogonality test because off-target and on-target trigger RNAs could be swapped between cgRNA and trigger RNA pairs.

The second goal was to enable titration of both cgRNA and trigger RNA in the assay. This was important for characterizing the performance of OFF to ON systems. OFF to ON systems can exhibit some level of leak into the ON state in the absence of trigger RNA. Some systems could also be stuck in a constitutively inactive state. Expressing the cgRNA on inducible promoters offered a way to check if a particular design could be active at all. This provided important hints about the fidelity of the
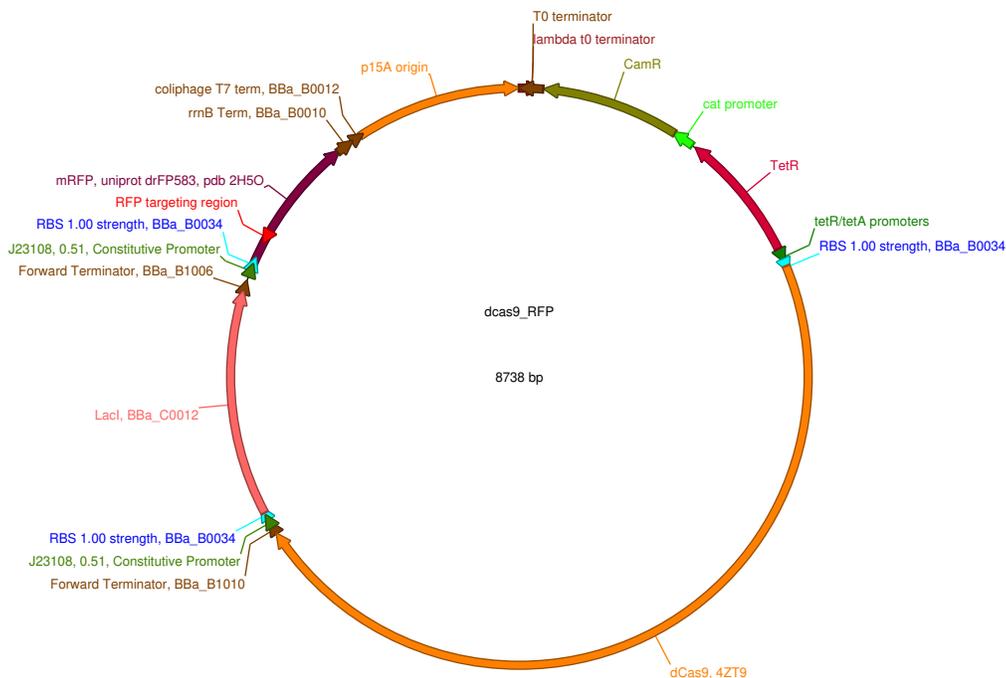
Figure 3.19: Map of the dCas9 plasmid used in the medium throughput setup. dCas9 was expressed on a pTet promoter. dCas9 could be induced with anhydrotetracycline (aTc). LacI and mRFP were expressed on constitutive promoters. CamR gave this plasmid chloramphenicol resistance. Genbank files for this plasmid are listed in Table A.17.

sequestering hairpins. If the cgRNA could not turn ON, the sequestering hairpin was too long. If the cgRNA leaked ON too easily, the sequestering hairpin was too short. Expressing the trigger RNAs on an inducible promoter allowed ON/OFF activity to be measured by inducing or not inducing the trigger RNA. This trigger dose response experiment was simpler and faster to setup than the orthogonality test because it only required one test strain per cgRNA and trigger pair. This was a bacteria expressing cognate cgRNA and cognate trigger RNA.

Finally, an mRFP reporter gene was integrated into the dCas9 plasmid behind the LacI gene (see Figure 3.19). This enabled generation of test strains from commercial bacterial cell lines such as DH10B and DH5A. DH10B and DH5A were *E. coli* strains which had higher transformation efficiencies than the MG1655 strain used in Chapter 2. This plasmid was useful for rapid and efficient generation of test strains via triple plasmid transformation with DH10B or DH5A.

3.4 Computational design of universal trigger RNAs

Higher throughput screening of cgRNA designs was also aided by the use of universal trigger RNAs, which enabled test strain standardization for the orthogonality

test. Universal trigger RNAs are a set of short unstructured trigger RNAs, whose sub-sequence reverse-complement have a low likelihood of interacting with other trigger RNAs in the same set. The unstructured nature of the universal trigger sequences makes them more accessible for RNA duplex formation with any cgRNA system being developed (see Figure 3.20). NUPACK test tube analysis of these sequences show that the sub-sequences of each trigger RNA are distinct enough to be orthogonally detected by any cgRNA system. This is illustrated by the absence of off-target binding in the heat map of crosstalk tube concentrations (see Figure 3.21).
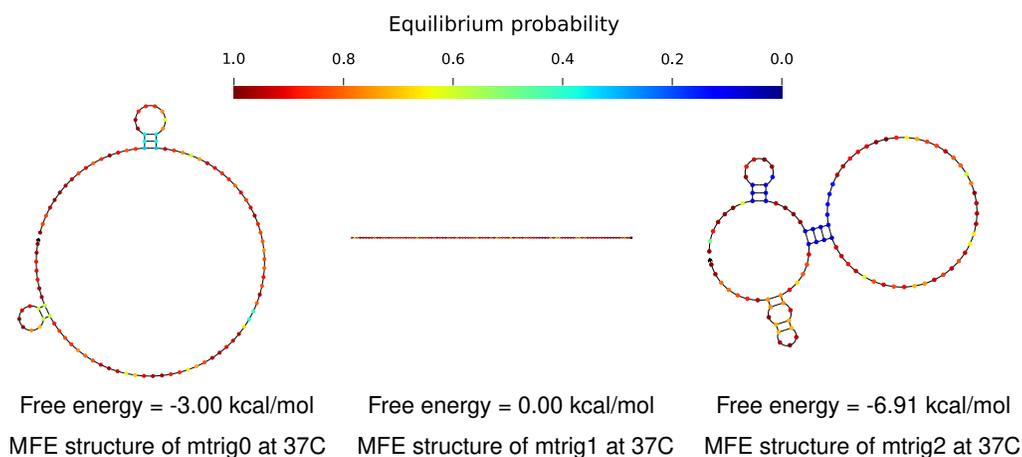
Figure 3.20: NUPACK analysis of the minimum free energy (MFE) structure of the universal trigger RNAs. The MFE structure represents the most populous RNA structure for the given sequence in the thermodynamic ensemble. Red means base pairs are highly likely to form. Blue means the base pairs are unlikely to form. The sequences used in the analysis are listed in Table A.8.

When cgRNAs are designed against sub-sequences drawn from a set of universal trigger sequences, the resulting cgRNA sequences will inherit the property of orthogonality from the universal trigger sequences. A cgRNA sequence targeted against mtrig0, for instance, will not be able to bind to mtrig1 or mtrig2. This abrogates the need for a crosstalk tube as, metaphorically speaking, the crosstalk information has been pre-computed and accounted for via the window constraints placed on the trigger sequences.

This design scheme was advantageous in two ways. First, it reduced the number of trigger RNAs required for each cgRNA mechanism being tested. Every cgRNA being tested triggers off the same set of universal trigger RNAs, which made it possible to generate a set of standardized trigger RNA strains for ON/OFF fold change and crosstalk measurements. The test strains were generated by transforming
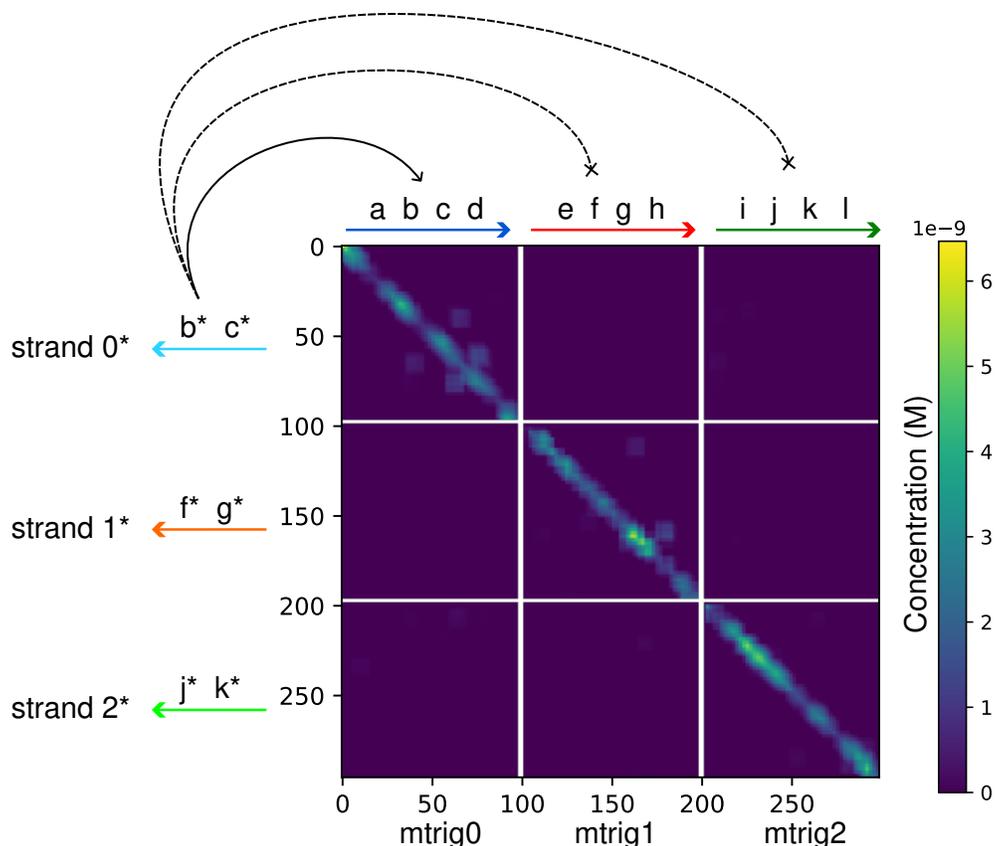
Figure 3.21: NUPACK analysis of crosstalk for the universal trigger RNAs. X axis represents the location of sub-sequences in each of the three universal trigger RNAs. Y axis represents RNA strands designed to bind against a particular sub-sequence on each universal trigger RNA. Strands designed to bind to their cognate trigger RNAs are unlikely to bind to their non-cognate trigger RNAs. This is illustrated with the solid arrow and dashed arrows. The heat map shows the average test tube concentration of strand to trigger RNA binding at each sub-sequence location. Yellow denotes high probability of interactions between the two strands. Blue denotes no interaction between the two strands. The sequences used in the analysis are listed in Table A.8.

the cgRNA plasmid into cell lines expressing dCas9 and each of the three universal trigger sequences. This strategy significantly reduced the cloning effort that was previously needed.

Second, the use of universal trigger RNAs enabled better comparison of the fold change ON/OFF performance between different cgRNA sequences. The $\Delta\Delta G$ free energy of trigger RNA binding to cgRNA influenced the achievable fold change response for each cgRNA sequence (see Figure 3.15). Trigger sequences with greater $\Delta\Delta G$ change were expected to have greater affinity for their cognate cgRNA. If one cgRNA has significantly more $\Delta\Delta G$ free energy than another design targeting a different trigger sequence, the fold change measurement would also be different.

Ideally, reliable comparison of trigger detection sensitivities between different designs required that each trigger RNA have the same sequence or $\Delta\Delta G$ free energy. This was not always feasible for every system and domain dimension tested. Use of universal trigger sequences helped normalize these differences and provided a reference standard for evaluating different cgRNA sequences and systems.

3.5 Conclusions

This chapter achieved two important goals. The first was the invention of a set of cgRNA systems with reduced design constraints. These systems are more suitable for sensing a wider array of endogenous RNA transcripts than the terminator switch and toehold switch cgRNAs presented in Chapter 2. These systems will be used for endogenous sequence detection in Chapter 4. The second outcome was the creation of a platform for quickly screening the plethora of new cgRNA systems. This cloning and screening infrastructure was crucial for identifying new cgRNA systems in a reasonable timeframe.

The utility of this screening assay should not be underestimated. In future work, this assay setup could be adapted towards protein and RNA engineering in other CRISPR systems such as prime editing[58] and base editing[59]. To test alternative Cas variants, the Cas9 protein (see Figure 3.19) could be replaced with other endonucleases such as Cpf1[78], Cas13b[79], or Cas9 reverse-transcriptase[58]. The fluorescent reporter protein could also be changed to other read out modalities such as luminescence[80] or cell survival markers such as the Hok-sok[81,82] or ccdB toxin system[83,84]. This could enable the use of alternative higher throughput screening methodologies such as live dead selection, next generation sequencing based screening, or high-throughput colony selection on a 384 or 1536 well plate. The time saved from using a faster and more reliable assay setup should be substantial. It is hoped that the lessons discussed in this chapter could save valuable time and mitigate the need for laborious cloning for the next student who takes on this project.

Chapter 4

DETECTING ENDOGENOUS RNA SEQUENCES

Showing how conditional guide RNAs (cgRNA) can be engineered to detect endogenous RNA sequences is the ultimate goal of this thesis. Endogenous sequence detection is important for enabling programmable control over when and where CRISPR/Cas9 gene editing is activated in a whole organism. cgRNAs could be engineered to turn ON or OFF in response to certain RNA transcripts which mark a particular tissue or cell type. Chapter 3 introduced a set of new cgRNA mechanisms which had reduced design constraints. These new cgRNA systems could accept a wider range of endogenous RNA sequences as trigger RNA input. The goal of this chapter is to show how these systems were used to detect endogenous microRNAs and mRNAs. This task was not trivial because it required the development of new tools for predicting mRNA sub-sequence accessibility and specificity. The following sections discuss why endogenous RNA detection is challenging, how computational tools were used to triage which mRNA sub-sequences were suitable for use as trigger RNAs, how a cgRNA design pipeline for endogenous sequence detection was created, how the marker genes for mRNA and microRNA detection were selected, and the results of my first foray into endogenous sequence detection.

4.1   Prior work and challenges with endogenous sequence detection

Detection of endogenous RNA sequences inside a living cell is possible because natural systems for RNA detection exist. The RNAi pathway in eukaryotic cells and the Cas13b RNA targeting system in bacteria are good examples[40,79]. These systems act as an immune system to protect the cell from RNA based viruses. They work via hybridization of a guide RNA strand to the target RNA transcript. The effector proteins in these respective pathways bind and cleave the guide RNA and target RNA duplex to silence the target gene. Although the natural configuration of siRNAs and Cas13b guide RNAs cannot switch between active and inactive states, the observation that their guide RNAs do interact with endogenous mRNA is encouraging because it suggests cgRNAs could also interact and trigger off of endogenous mRNAs inside of living cells.

Detection of long RNA strands is more difficult than detection of short RNA strands. This has been true for both cgRNAs (see Figure 4.18) and conditional siRNAs[39,40,45].

a

Pseudoknot

Long RNA in a test tube

b

Ribosomes
translating
protein

mRNA in a test tube

c

Endoplasmic reticulum

Nucleus

mRNA

guide RNA

mRNA in a eukaryotic cell
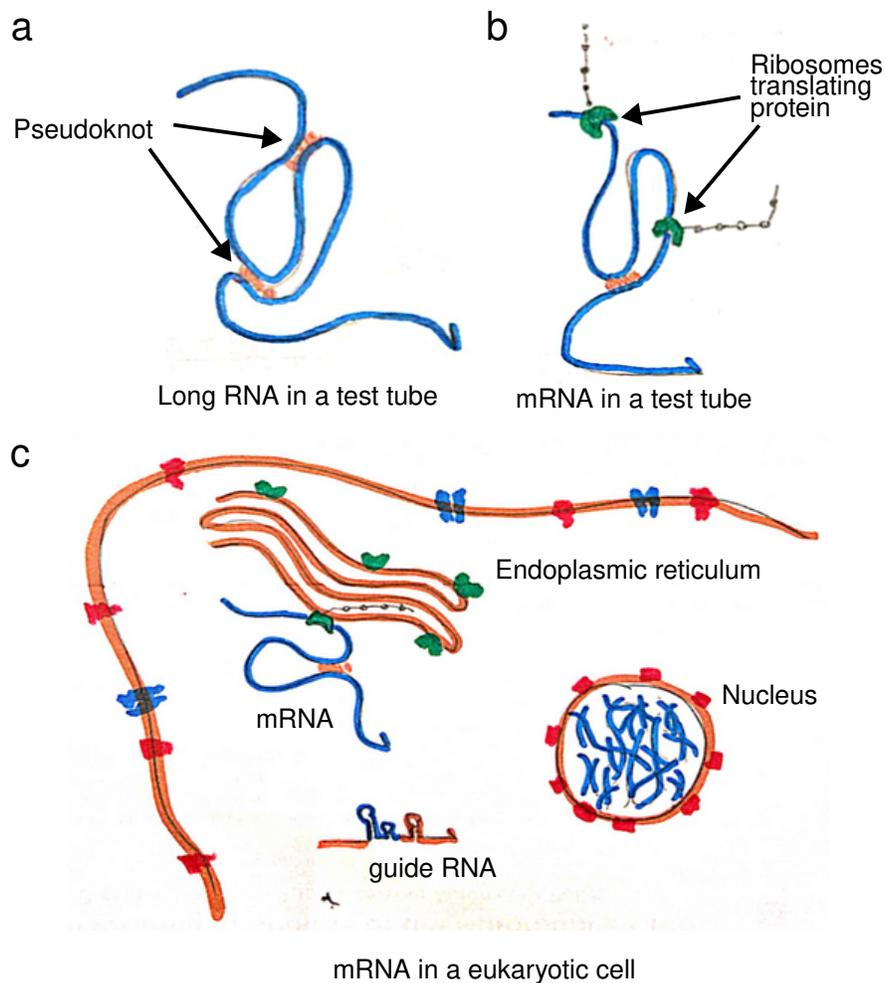
Figure 4.1: Challenge of mRNA detection at different length scales. a) At the single mRNA level, formation of pseudoknots in long RNAs complicate the prediction of RNA structure. b) Translation of mRNA into protein by ribosomes moving along the RNA transcript will also alter the local RNA structure. Hairpins and pseudoknots must be unwound for ribosomes to pass through. c) In a eukaryotic cell, different RNA species such as mRNA, non-coding RNA, and microRNA are translocated to different compartments such as the endoplasmic reticulum or nuclear spliceosome for post processing. It is unclear if small RNAs such as guide RNA can interact with mRNA before they are both transported out of the nucleus.

Long RNA strands tend to have pseudoknots (see Figure 4.1a) which complicate predictions about the accessibility of an RNA sub-sequence[45,85]. Thermodynamic models such as NUPACK cannot efficiently compute the formation of pseudoknots, yet these interactions are important for determining the structure of long RNA sequences. mRNAs add another level of complexity to site accessibility predictions. In mRNA, ribosomes are continuously sliding down the RNA transcript assembling amino acid chains from the mRNA template (see Figure 4.1b). This interaction could unwind and rewind certain regions of the mRNA in unforeseen ways. In eukaryotic

cells, mRNA and non-coding RNA are synthesized in the nucleus and exported to different cellular compartments. mRNA is first transported to the splicesome (in the nucleus) to remove introns. Afterwards, the mature transcript is exported to the endoplasmic reticulum to help make protein. Non-coding RNAs do not necessarily end up in the endoplasmic reticulum. They can remain in the nucleus to help with post-translational processing of mRNA or silencing of certain genes[45]. For cgRNAs, it is not clear if they have any opportunity to interact with mRNA before it is exported out of the nucleus. Note that this is not a problem in bacterial systems. Understanding how each of these aspects of the RNA life cycle affects cgRNA activity may be critical for getting endogenous sequence detection to work.

In this chapter, I focused on studying the difference between mRNA and non-coding RNA at different length scales in bacteria. This was done by testing cgRNAs against synthetic mRNA transcripts and endogenous coding and non-coding variants of the same RNA transcript. Synthetic mRNAs were fluorescent protein mRNAs tagged with a validated trigger sequence on the 3′ end. Previously validated cgRNA sequences could be tested against these synthetic mRNAs to check if mRNA sequence detection was at all feasible. The non-coding and coding versions of endogenous transcripts were tested to study the effect of ribosome activity on trigger sequence detection. Non-coding transcripts contained the same sequence as coding transcripts, but lacked a ribosome binding site needed to initiated protein translation. The problem of length scales was studied via truncation of the longer mRNA transcript to create shorter trigger RNAs. This equated to reducing the influence of pseudoknotted structures. I avoided the complexities of cellular compartmentalization by working only in bacteria. Bacteria are not compartmentalized like eukaryotic cells. The bacterial chromosome, ribosomes, and endogenous RNAs intermingle together in the cytosol. Adding the complexities of cellular compartmentalization to these studies would further confound our understanding of the problem and hinder our ability to find a solution. Once the problems associated with RNA detection in bacteria are solved, we will have a better foothold for tackling the same problems in the eukaryotic setting.

4.2 Predicting the accessibility and specificity of mRNA sub-sequences

Detecting long endogenous RNA transcripts is more difficult than detecting short synthetic trigger RNAs. cgRNAs can only base pair with short sequences. However, selecting an appropriate sub-sequence to base pair with is not trivial for three reasons. First, endogenous sub-sequences can have many paralogs. This is a

byproduct of evolution when proteins evolve from common progenitor genes. A random 20 to 40 nucleotide sub-sequence in a marker gene may not be suitable as a trigger sequence because it is not specific to that RNA transcript. Second, not all locations along an mRNA are good sites for cgRNA binding. DNA probe binding experiments on d2eGFP mRNA in cell lysate reveal a highly variable DNA-RNA hybridization yields due to native secondary structure within the mRNA (see Figure 4.3). Targeting the wrong sub-sequence window can make endogenous sequence detection more difficult. Alternative predictors of sub-sequence site accessibility or specificity may need to be leveraged to address these issues. Finally, the experimental workflow cannot exhaustively test all the possible combinations of mRNA sub-sequence targets, cgRNA systems, and domain dimensions. A system for triaging promising cgRNA designs is necessary. These complications motivate the search and development of alternative predictors for ranking trigger RNA sub-sequence suitability in the design process. The output of these tools are meant to be used as additional parameter inputs to the cgRNA design process for endogenous sequence detection.
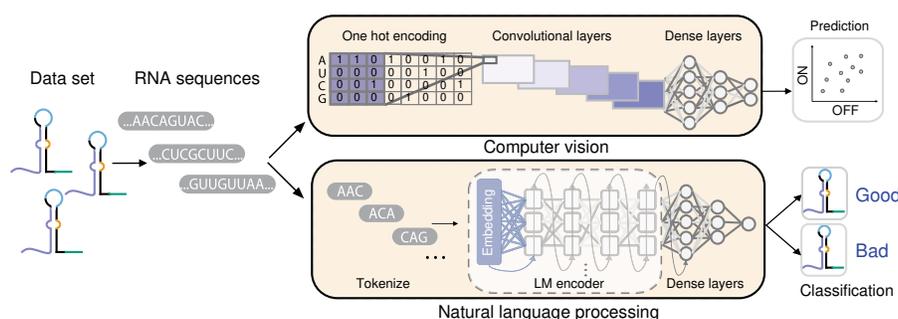
Figure 4.2: Illustration of the machine learning predictor for fold change performance called STORM. STORM takes an mRNA sequence as input. The mRNA sub-sequences are transformed into a list of toehold switch sequences. These sequences are one hot encoded and fed through a deep convolutional neural network model which predicts the ON/OFF performance of the toehold switch design. The machine learning model was trained on a large data set of toehold switch performance data obtained through a high-throughput NGS based screening assay. Figures are adapted from Valeri *et al.* [86]

One of the tools used as an alternative predictor of trigger RNA performance was the Sequence-based Toehold Optimization and Redesign Model (STORM). STORM is a machine learning based tool for predicting the performance of toehold switch mR-NAs (see Figure 4.2).[86] Toehold switch mRNAs are synthetic riboswitches designed to conditionally sequester a ribosome binding site. Co-expression of a trigger RNA sequence in bacteria opens up the ribosome binding site, enabling translation of a fluorescent reporter protein[41–43]. STORM accepts an mRNA sequence as input.
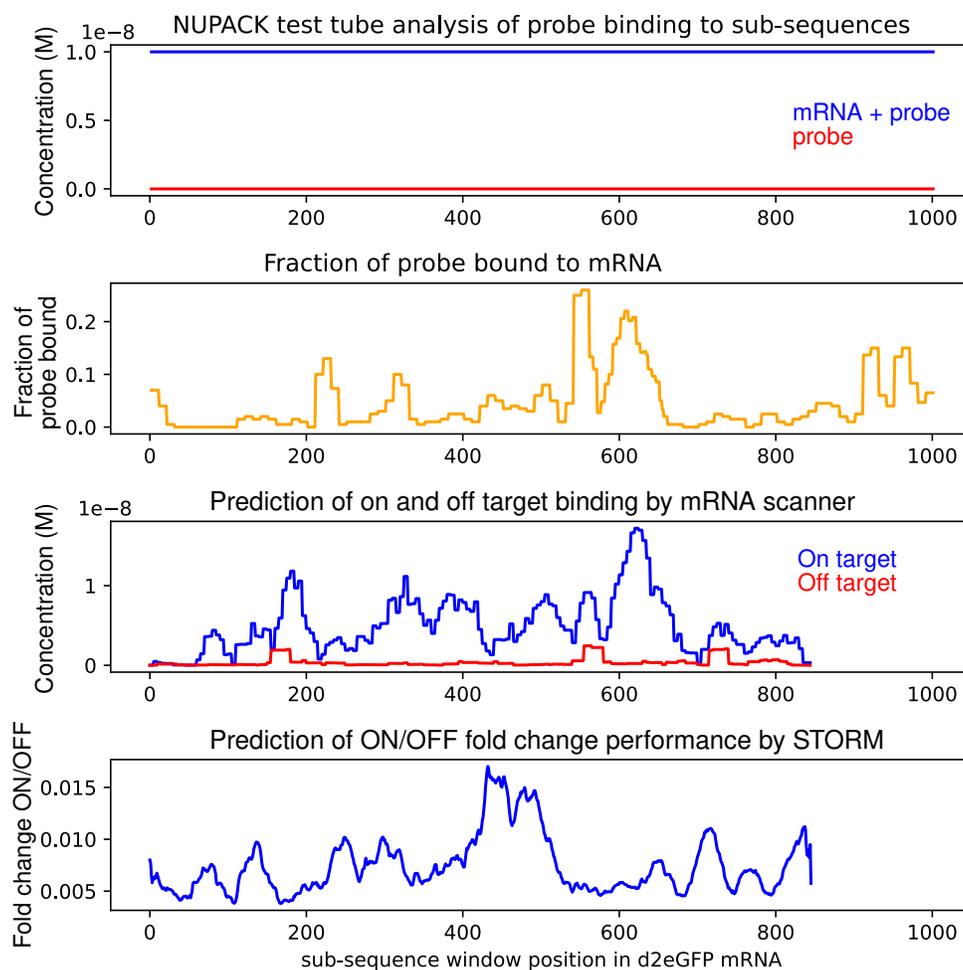
Figure 4.3: Plots of different predictors of d2eGFP sub-sequence accessibility and specificity. In the first row, NUPACK test tube analysis was used to compute the formation of a mRNA sub-sequence and probe duplex at 37C with 10nM of each reactant. All duplex pairs were predicted to form. No unpaired sub-sequences were predicted to be left at equilibrium. The second row shows experimental data from a DNA probe hybridization experiment.[40] This was used to measure site accessibility on d2eGFP mRNA. The third row shows predictions of mRNA sub-sequence specificity using the mRNA scanner tool. This was based on the propensity of RNA hairpins (toehold 5nt and duplex 20nt) binding to their cognate and non-cogate mRNAs. SNAI2, PAX7, and TFAP2b were used as non-cognate mRNAs in this analysis. These values represent the row sum of the heat map shown in Figure A.7. The fourth row shows the fold change performance predicted by STORM if each sub-sequence of d2eGFP was sensed with a toehold switch mRNA.[86]

The machine learning model scans the mRNA and builds a list of toehold switch mRNAs constrained to detect each sub-sequence in the input mRNA at a stride of one base pair. The machine learning model outputs a normalized prediction of toehold switch performance ranging from 0.0 to 1. High values indicate good ON/OFF performance. This output was averaged with a moving window filter and
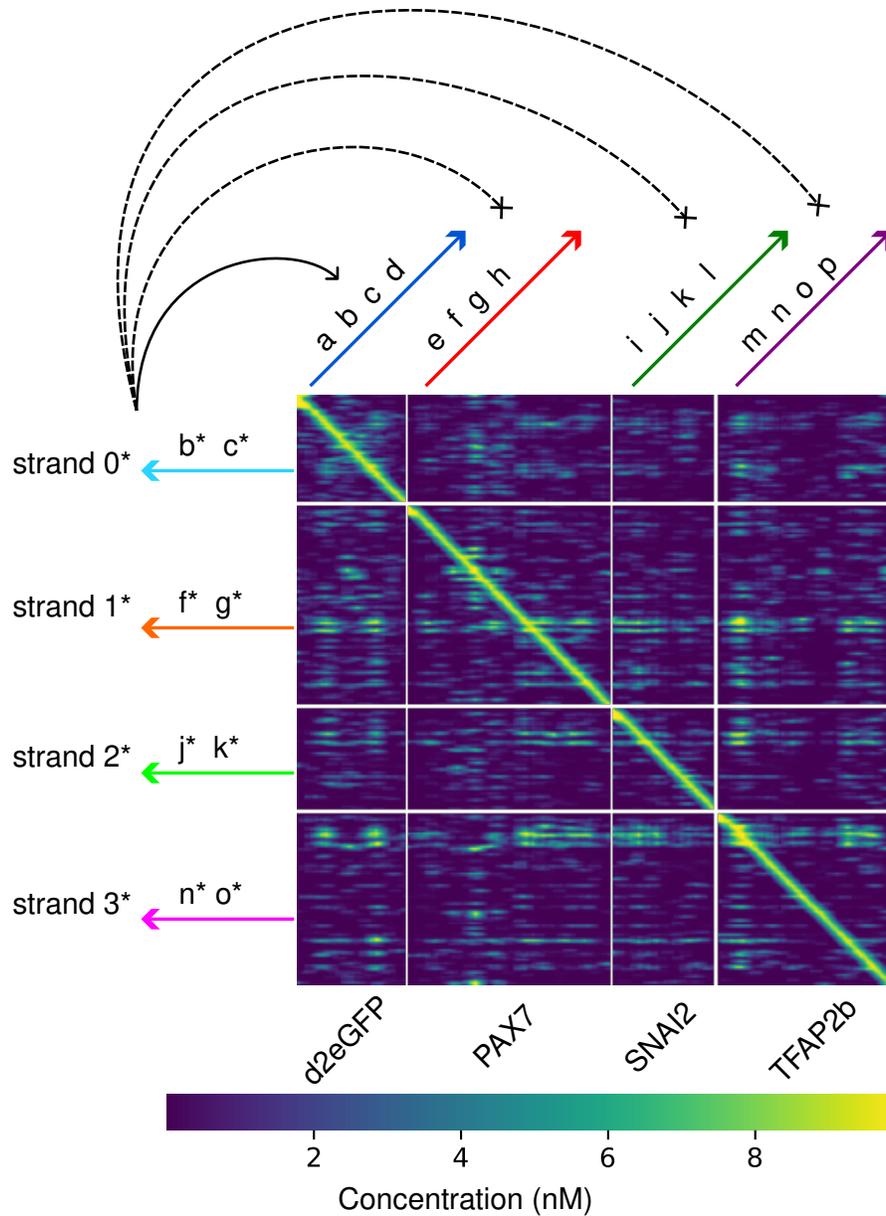
Figure 4.4: Illustration of the sub-sequence specificity analysis performed by mRNA scanner. The above heat map shows the NUPACK predicted crosstalk between mRNA sub-sequences (x axis) and their sub-sequence reverse-complements (y axis) in a crosstalk tube analysis at 37C. 10nM was used as the reactant concentration. Yellow denotes regions of high interaction, and blue denotes low interaction. The yellow diagonal represents on-target interactions. Off-diagonal values represent off-target interactions. A 20nt long strand of RNA was used as the anti-sense strand, which binds with a moving window of mRNA sub-sequences (100nt long).

used to predict fold change response for a given trigger RNA sequence (see Figure 4.3, bottom plot).

STORM was used as a predictor of trigger RNA fold change response because it

was one of the only models trained on a large set of experimentally derived performance data. This data set of approximately 91,534 toehold switches and trigger RNA pairs was obtained using cell sorting and next generation sequencing (NGS).[86] In this experiment, short trigger RNA sequences and toehold switch mRNAs were expressed as a single RNA transcript in bacteria and sorted to high or low fluorescence. Toehold switch mRNAs were designed to produce fluorescent protein only if its cognate trigger RNA sequence was present on the same transcript (see Figure 4.2). This NGS screening scheme enabled high-throughput identification of good and bad toehold switch designs, which is why this data set is so large and comprehensive. Trigger RNA sequences were drawn from the genomes of zika virus to mammalian protein coding genes.[86] The diversity of sequences used was believed to improve this model's generality toward the endogenous mRNAs and microRNAs targets tested in this chapter.

Predicting sub-sequence specificity was another important requirement for mRNA detection, and this was an area in which NUPACK excels.[87,88] In Chapter 2, NUPACK was shown to be capable of designing orthogonal cgRNA systems which have low crosstalk. This observation motivated the development of a NUPACK based tool for predicting mRNA site specificity. This tool, termed the mRNA specificity scanner (mRNA scanner), takes two inputs, a list of on/off target mRNAs and a list of cgRNA sequences. mRNA scanner generates a list of trigger RNA sub-sequences from the input mRNAs and uses these sequences in a pairwise crosstalk tube analysis (see Figure 4.4). If a list of cgRNAs was provided, mRNA scanner performs test tube analysis on each combination of cgRNA and input mRNA sub-sequences. If cgRNA sequences were not provided, the reverse-complement of the input mRNA sub-sequences is used in place of cgRNAs for the cross tube analysis. The output of this analysis is a table of test tubes, nucleic acid complexes, and complex concentrations for each cgRNA and mRNA sub-sequences provided as input. This information can be summed up to estimate the degree of on-target and off-target interaction for each cgRNA or mRNA sub-sequence window. Highly sequence specific cgRNAs or mRNA sub-sequences will have low off-target concentration values verse the on-target concentration values. These parameters can be used as input in the cgRNA design process to remove cgRNA designs which have high crosstalk.

An important caveat about each of these predictors is that they are not in agreement with each other or experimental data (see Figure 4.3). The proper way to leverage each of these predictions is not clear, so it is worthwhile to note the shortcomings of

each prediction to help guide our search. With STORM, the machine learning model was trained to predict the performance of toehold switch mRNAs and not cgRNAs. The behavior of cgRNAs and toehold switch mRNAs could be very different. One system interacts with dCas9 and the other interacts with ribosomes. It was not known if predictions about toehold switch performance could translate well to cgRNAs. With the mRNA probe hybridization data, this data set only existed for d2eGFP mRNA, which was not an endogenous mRNA transcript. Moreover, this experiment was conducted in a test tube, which may not be reflective of conditions inside of a cell. With mRNA scanner, the rationale behind testing for off-target interactions computationally is well grounded. However, computational predictions can change drastically when different probe dimensions are tested. Using unstructured probes can result in high crosstalk (see Figure 4.4). Using detection hairpins can suppress off-target interactions (see Figure A.6). Reducing the toehold domain length of the detection hairpin can further reduce predicted crosstalk (see Figure A.7). Moreover, the underlying model for RNA structure prediction uses energy parameters derived from nucleic acid melting experiments in a test tube. Again, the conditions inside a test tube may not be reflect of conditions inside the cell. In the absence of better alternatives, the only option is to just try it.

4.3 Computational design of cgRNAs for endogenous sequence detection

The cgRNA design workflow was revised so external scoring metrics, such as those generated by STORM and mRNA scanner, could be used to rank and triage promising cgRNA sequences. The need to triage sequences arose out of experimental necessity. Experimental setups have finite capacity. STORM and mRNA scanner provided an alternative way to evaluate cgRNA sequences using models trained on empirical data and alternate measures of sequence specificity. This was necessary to augment NUPACK's test tube predictions because they were not stringent enough to eliminate non-functional sequences related to mRNA site accessibility and specificity.

These modifications were not trivial to implement for two reasons. First, the window constraint function in NUPACK cannot place weights on the input mRNA sub-sequence. The importance of each sub-sequence in the cost function was spread equally across all mRNA sub-sequence windows. This means there is no internal mechanism for ranking cgRNA sequences with the aid of external scoring metrics. Second, orthogonal mRNA detection was difficult to specify. All possible combinations of off-target complexes for a given cgRNA design had to be enumerated in the crosstalk tube. Recomputing the partition function for each off-target com-

plex across each independent design was unnecessary. Optimization of on-target test tubes could fail before the crosstalk tube computation became necessary. To address these issues, the design workflow was separated into two steps (see Figure 4.5).

The first step focused only on improving the ON/OFF fold change response of cgRNAs targeted to each mRNA sub-sequence. This modification eliminated redundant calculations related to the crosstalk tube. It also allowed greater parallelization of design tasks. A json container of NUPACK design specifications could be generated for each cgRNA system, domain dimension, and mRNA sub-sequence target. These containers could be uploaded to an S3 file server and processed by a herd of NUPACK docker containers on the Pierce lab kubernetes cluster (see Figure 4.5). As the docker containers work through the queue of designs, results were continuously uploaded onto the S3 file server. These results were accumulated for processing in the second stage of the pipeline.

The second stage of the pipeline focused on sequentially filtering for cgRNA sequences which satisfy an extensible set of external constraints such as sequence specificity. Here, mRNA scanner and STORM scoring metrics could be used to filter for cgRNA sequences that target some desired mRNA sub-sequence windows, fulfill some threshold level of fold change predicted by STORM, and satisfy additional test tube constraints defined by the user. These constraints could be the degree of on-target and off-target specificity for a given cgRNA sequence, empirically derived scoring metrics such as the probe binding data, or data-driven statistical predictions such as in the case of STORM. The science of filtering and ranking promising cgRNA sequences is not perfect. Work in refining this process is still needed.

A major benefit of this pipeline is the automation of oligo library generation for experimental testing. The final output of the pipeline is a set of DNA oligos which can be submitted to Twist Biosciences for synthesis and a set of Genbank or fasta files defining the cgRNA plasmid constructs. These files are used for sequence alignment to verify the identity of the test strains and plasmid constructs. The oligo library output is also useful for transitioning into an NGS based workflow where sequence alignment against the oligo library can be used to identify cgRNA sequences and demultiplex from their experimental conditions. This puts us a step closer toward executing a high-throughput NGS based screen for cgRNA activity.
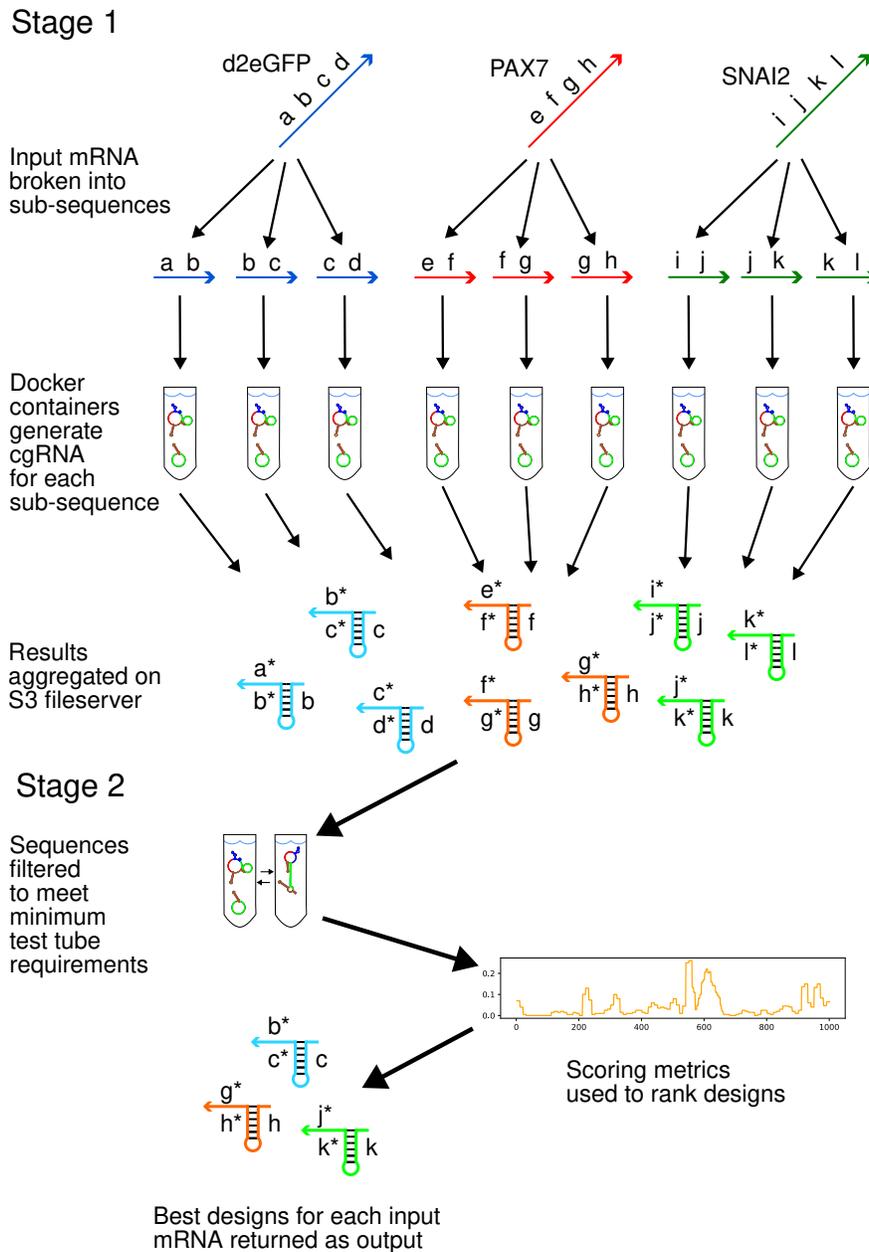
Figure 4.5: Illustration of the bioinformatic pipeline used to design cgRNAs for endogenous sequence detection. In the first stage, input mRNA is broken in sub-sequences. A cgRNA sequence is obtained for each sub-sequence. In the second stage, cgRNA sequences are filtered to meet minimum test tube requirements for ON state and OFF state concentration. Sequences which pass the test tube filter are ranked and filtered according to the external scoring metrics provided. The highest ranking designs are returned and generated into oligos for cloning.

4.4 Detection of synthetic mRNAs

Synthetic mRNA detection experiments were motivated by the need to check if each cgRNA system engineered so far have any issues detecting mRNAs. As noted at

the beginning of this chapter, ribosomes sliding down the mRNA transcript could interfere with RNA duplex formation via winding and unwinding of mRNA. Unlike siRNAs and Cas13b guide RNAs, cgRNA-mRNA complexes do not benefit from the stabilizing energies of protein binding. cgRNA-mRNA complexes could be more vulnerable to RNA strand displacement by the ribosome than siRNA and Cas13b guide RNA. This aspect of endogenous RNA detection was ignored in contemporary studies such as toehold switches[41], STAR terminators[38], RADARs[89], and cgRNAs[47,66,90,91].

Synthetic mRNAs were constructed via two approaches. The first was to add a ribosome binding site 5′ of a universal trigger RNA. The universal trigger RNAs were designed to be short unstructured RNA strands that also encode a short amino acid sequence. These transcripts began with a start codon and ended with a stop codon. Adding or removing a ribosome binding site inter-converts the transcript between short mRNA or short trigger RNA. This enabled us to study the influence of the ribosome binding site without the complexity of pseudoknot interactions at larger RNA length scales. The second approach involved inserting an unstructured trigger RNA sequence behind a fluorescent protein mRNA. This mRNA was miRFP670[*]. Sequence insertions in this region were intended to provide a known accessible site for cgRNA binding to the miRFP670-trigger. In contrast to the short mRNA approach, translation of the miRFP670-trigger into fluorescent protein could be measured by another channel on the plate reader or flow cytometer. This helped confirm the mRNA-trigger transcript was making protein. It also enabled observations of how cgRNA binding to miRFP670-trigger interfered with protein translation or vice versa how protein translation interfered with cgRNA-trigger duplex formation.

Detection of synthetic mRNAs with terminator switch cgRNAs

The work of Chapter 2 produced an orthogonal set of terminator switch cgRNA and trigger RNA pairs. These unstructured trigger sequences could be appended to miRFP670 mRNA as the 3′ untranslated region (UTR). This generates a set of orthogonal fluorescent protein mRNAs detectable by their cognate terminator switch cgRNAs. The goal of this experiment was to see if terminator switch cgRNAs could be used to detect mRNAs that actively produce protein.

The cgRNA terminator switch constructs used in Chapter 2 (see Figure 2.10 and 2.8)

---

[*]miRFP670 mRNA encodes a fusion protein of miRFP670 and heme oxygenase. miRFP670 uses biliverdin, a byproduct of heme oxidation, as a co-factor for infrared fluorescence. Heme oxygenase was necessary to produce biliverdin in bacterial culture. Otherwise, the protein would not fluoresce.
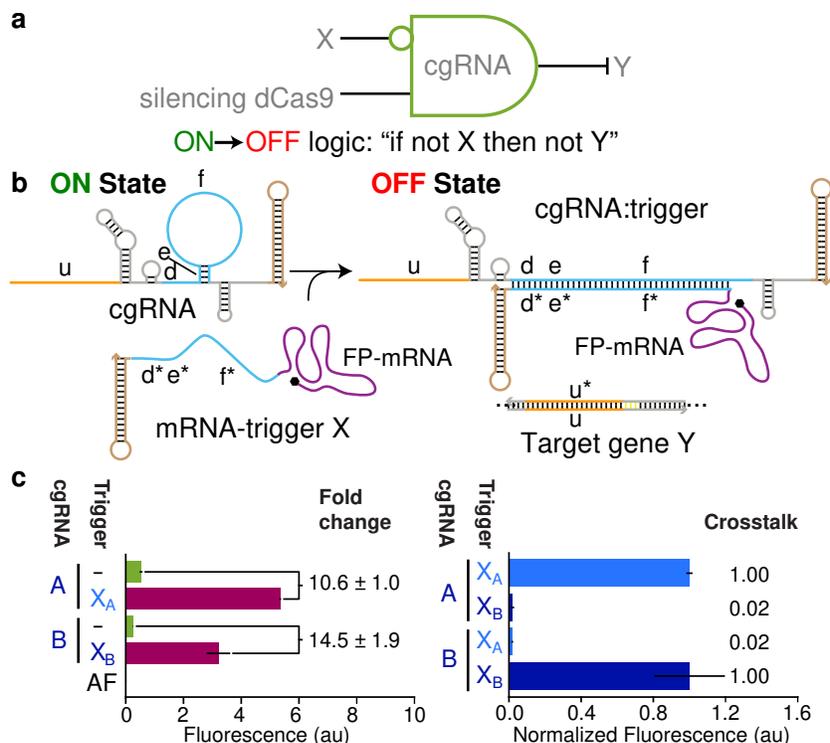
Figure 4.6: Detection of synthetic mRNAs by terminator switch cgRNAs in a dCas9-based RFP silencing assay in bacteria. a) Schematic of the cgRNA logic. Terminator switch cgRNAs are designed to switch from ON to OFF in response to trigger RNA (low to high fluorescence). b) Schematic of the terminator switch mechanism detecting a synthetic mRNA. Trigger RNA sequences are appended 3′ of the miRFP670 mRNA to create a synthetic mRNA. Binding of this trigger sequence to the cgRNA inactivates the cgRNA. c) Bar plots showing the orthogonality of synthetic mRNA detection with the terminator switch cgRNAs. The median RFP fluorescence signal of terminator switch cgRNAs silencing RFP in bacteria is plotted on the X axis. Fluorescence signal was acquired via flow cytometry. RFP signal only increases when cognate mRNA-triggers were induced. Terminator switch cgRNAs did not respond to off cognate trigger RNAs. At least three replicates were used for the error bars. See Figure 2.6 captions for equations computing fold change and crosstalk.

were modified to express the trigger RNA sequences as a mRNA-trigger. The same cgRNA activity assay described in Chapter 2 was used to measure RFP silencing by the terminator switch cgRNA. miRFP670-trigger RNA was expressed on the pLac promoter and induced with IPTG. Expression of the miRFP670-trigger RNA could be measured with infrared fluorescence. cgRNA activity could be measured by RFP silencing. Synthetic mRNA detection by the terminator switch cgRNA was successful. The experiment demonstrated both orthogonal detection of mRNA-triggers and dose responsive sequestration of terminator switch cgRNAs by the mRNA-trigger (see Figure 4.6 and 4.7). cgRNA D was inactivated only by mRNA-trigger D, and cgRNA I was inactivated only by mRNA-trigger I. These samples had
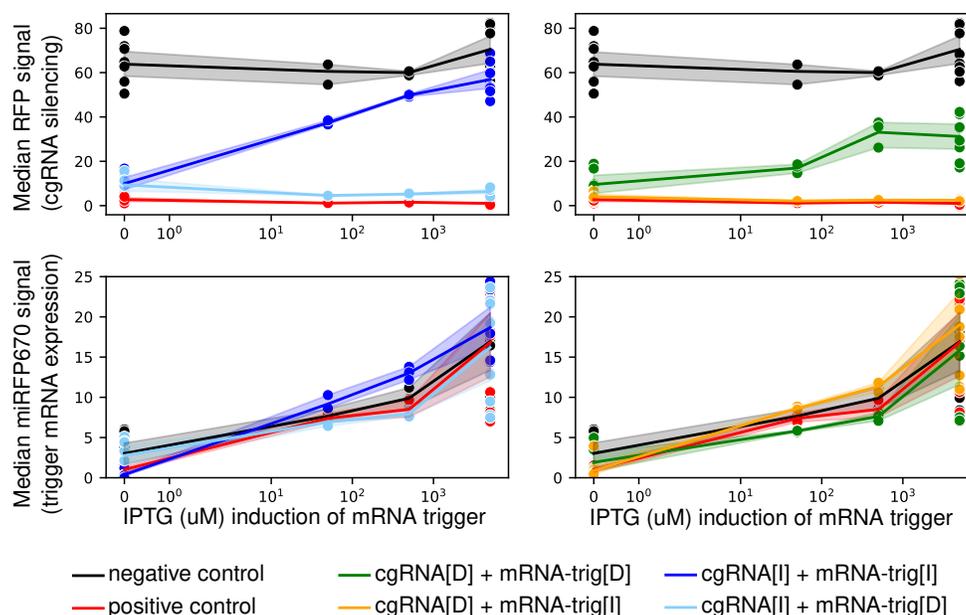
Figure 4.7: Dose response curve of terminator switch cgRNAs detecting synthetic mRNA in a dCas9-based RFP silencing assay in bacteria. Terminator switch cgRNAs are designed to switch from ON to OFF in response to trigger RNA (low to high fluorescence). RFP signal only increases when cognate mRNA-triggers were induced. Terminator switch cgRNAs did not respond to non-cognate trigger RNAs. Fluorescence values for RFP and RFP670 were obtained via flow cytometry. The plots show the median fluorescence values of the cell population vs induction of the mRNA-trigger by IPTG. cgRNA was constitutively expressed while fluorescent protein mRNA was expressed on a pLac promoter. Markers represent experimental replicates. Shaded regions represent the standard deviation of the replicates. The mRNA-trigger sequences used are listed in Table A.7. The cgRNA sequences used are listed in Table A.6. Data from scRNA_Cas9_20200112ZC

increased RFP signal as more mRNA-trigger was induced. Expression of miRFP670 fluorescent protein was not perturbed by cgRNA binding. The infrared fluorescence of non-cognate samples were the same as cognate samples. This experiment did not compare the dose responsiveness of the short trigger RNA vs mRNA-trigger in the terminator switch system. This follow up experiment would have told us if ribosome activity perturbs cgRNA binding, but was not pursued due to time constraints.

This experiment was important for showing that cgRNAs can detect mRNAs that actively produce protein. Endogenous sequence detection with the terminator switch system was not pursued because this system had structural constraints that made it challenging to use with endogenous trigger sequences. Other higher performing cgRNA systems were already developed at the time of this experiment. Time was instead dedicated towards testing these newer systems. Nevertheless, these results were encouraging because they show that detection of mRNA sub-sequences by

cgRNAs was possible.

Detection of synthetic mRNAs with the reverse toehold switch

Detection of synthetic mRNAs with the reverse toehold switch system utilized short RNA transcripts rather than long synthetic mRNAs. Unlike the terminator switch experiment above, this experiment utilized the universal trigger sequences as the mRNA transcript (see Table A.9). This was not possible with the terminator switch trigger sequences because they contained stop codons which prematurely halted protein translation. In universal trigger mRNAs, the ribosome could hypothetically slide along the whole trigger sequence, producing a short chain of 33 amino acids. Chapter 3 showed the reverse toehold switch system could detect the universal trigger RNA sequences. It was not clear if this system could detect the same sequences expressed as an mRNA.

Figure 4.8: Detection of synthetic mRNAs by the reverse toehold switch. An RFP silencing assay in bacteria was used to measure cgRNA activity. cgRNAs were designed to switch from OFF to ON in response to trigger RNA. RFP signal only decreased when the cognate trigger sequences were present. This corresponded to the OFF to ON switch. Fluorescence values for RFP silencing were obtained via flow cytometry. The mRNA-trigger sequences used are listed in Table A.9. The cgRNA sequences used are listed in Table A.11. Data from scRNA_Cas9_20221114ZC

The reverse toehold switch system was designed to switch from OFF to ON (see Figure 3.3). This corresponded to a shift from high to low signal in the RFP silencing assay in bacteria. Detection of universal trigger mRNAs by the toehold switch was successful (see Figure 4.8). The short trigger mRNA samples induced a smaller RFP shift than the non-coding trigger. For some samples such as ts45m2-3, these

differences were negligible. For ts45m1-1, the loss in trigger RNA sensitivity was significant. This meant the presence of a ribosome binding site disrupted cgRNA binding to trigger RNA. Its degree of influence on mRNA detection was unclear because of a lack of data. Nevertheless, these results showed it was feasible to detect short synthetic mRNAs with the reverse toehold switch system.

Detection of synthetic mRNAs with the single-hairpin split cgRNA

The single-hairpin split cgRNA system was tested against the universal trigger mRNAs. This system was designed to switch from OFF to ON (see Figure 3.12). This corresponded to a shift from high to low signal in a dCas9-based RFP silencing assay in bacteria. Detection of universal trigger RNAs by this system was successful. In some test conditions, the trigger mRNA induced a smaller RFP shift than the non-coding trigger (see Figure 4.9). This meant the presence of the ribosome binding site could somewhat disrupt cgRNA binding to trigger RNA. However, its influence on mRNA detection did not seem to be significant for these cgRNA sequences. These results showed it was feasible to detect short synthetic mRNAs with the single-hairpin cgRNA system.
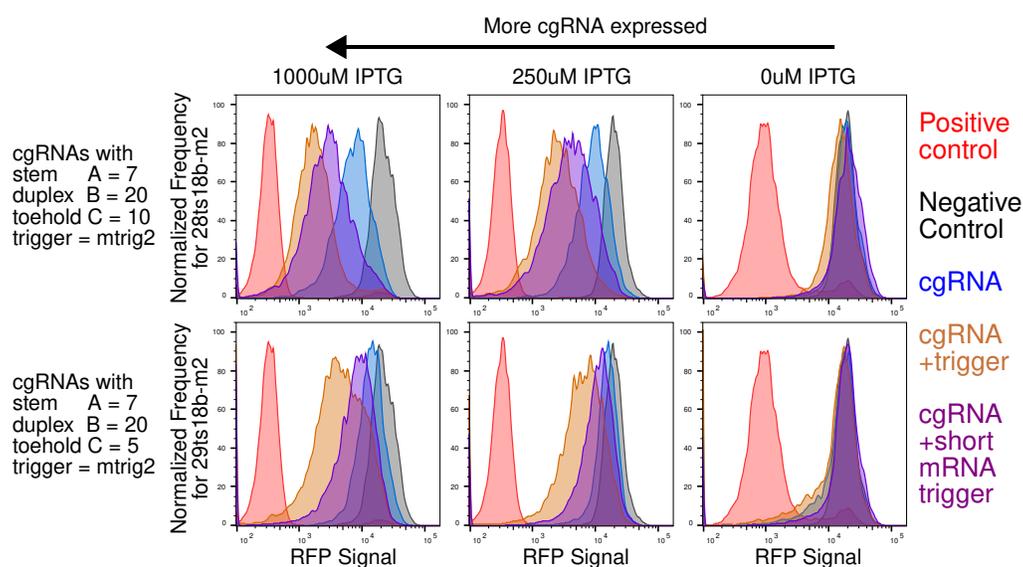
Figure 4.9: Detection of synthetic mRNAs by the single-hairpin split cgRNA. A dCas9-based RFP silencing assay in bacteria was used to measure cgRNA activity. cgRNAs were designed to switch from OFF to ON in response to trigger RNA. RFP signal only decreased when the cognate mRNA-triggers were induced. This corresponded to the OFF to ON switch. Fluorescence values for RFP silencing were obtained via flow cytometry. The mRNA-trigger sequences used are listed in Table A.9. The cgRNA sequences used are listed in Table A.14. Data from scRNA_Cas9_20221114ZC

Detection of synthetic mRNAs with the two hairpin guide RNA

The two hairpin cgRNA system was tested against the universal trigger mRNAs. This system was designed to switch from OFF to ON (see Figure 3.14). This corresponded to a shift from high to low signal in the RFP silencing assay in bacteria. Detection of universal trigger RNAs was successful (see Figure 4.10). In some test conditions, the trigger mRNA induced a smaller RFP shift than the non-coding trigger (see Figure 4.10, 250uM IPTG induction for 30ts23b-m2). These differences were sometimes negligible depending on the amount of cgRNA expressed. This meant the presence of the ribosome binding site could sometimes disrupt cgRNA binding to trigger RNA. Further study on this phenomena using fluorescence resonance energy transfer (FRET) microscopy and super resolution microscopy would useful for understanding why this occurs. Nevertheless, these results showed it was feasible to sense short synthetic mRNAs with the two hairpin cgRNA system.
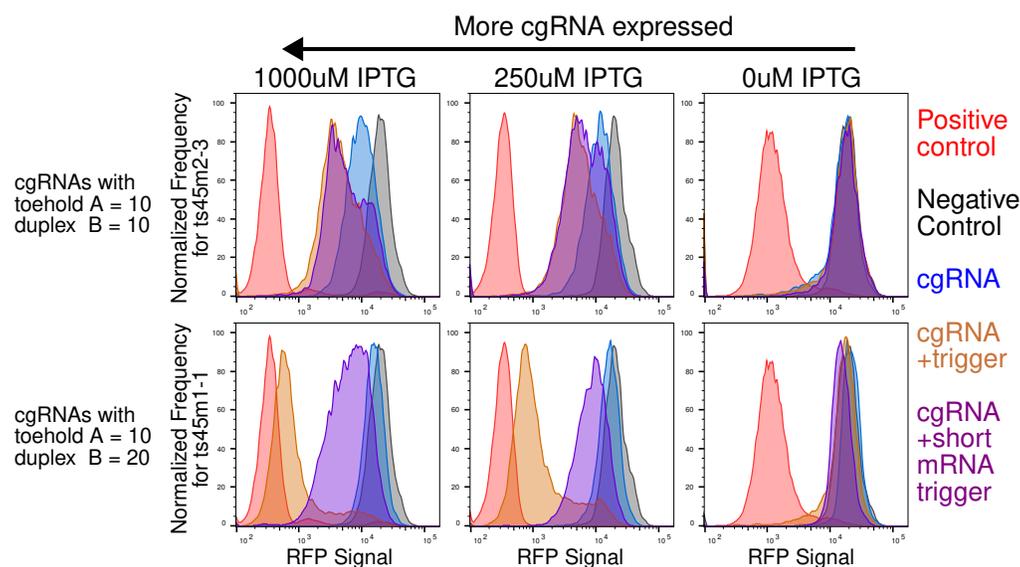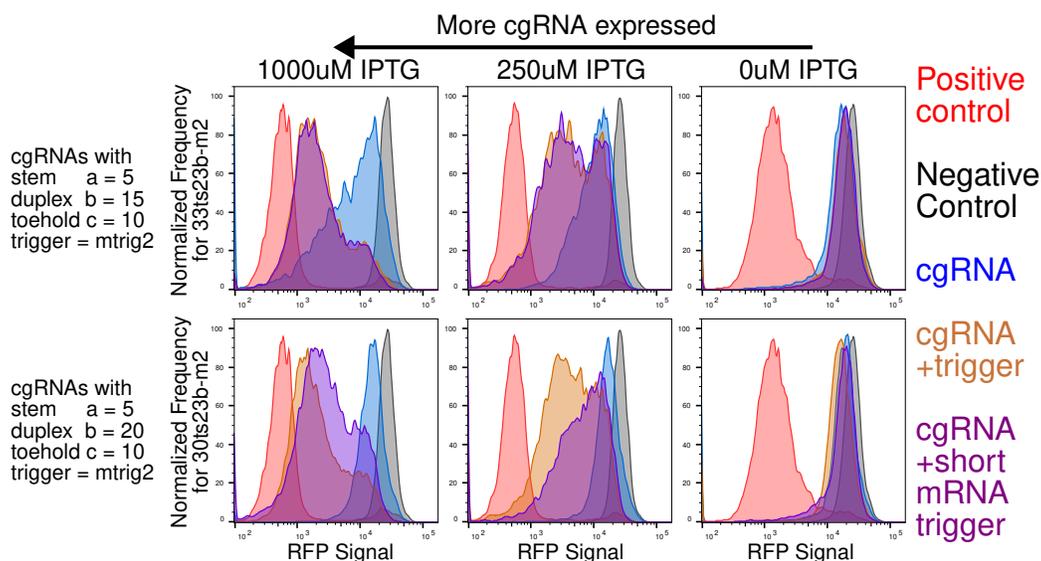

Figure 4.10: Detection of synthetic mRNAs by the two hairpin guide RNA. An RFP silencing assay in bacteria was used to measure cgRNA activity. cgRNAs were designed to switch from OFF to ON in response to trigger RNA. RFP signal only decreased when the cognate mRNA-triggers were induced. This corresponded to the OFF to ON switch. Fluorescence values for RFP silencing were obtained via flow cytometry. The mRNA-trigger sequences used are listed in Table A.9. The cgRNA sequences used are listed in Table A.15. Data from scRNA_Cas9_20221114ZC

4.5   Detection of endogenous RNA sequences

cgRNAs (cgRNA) were designed to detect a set of endogenous mRNA and microRNA transcripts (see Table 4.1) specific to the neural crest cells of a developing chicken embryo. The ultimate aim was to show cell-selective activation of

CRISPR machinery with cgRNAs in a whole organism. This developmental biology experiment would have been valuable because alternative methods to implement cell-selective gene editing in an embryo involved mining for cell type specific promoters via chromosome immunoprecipitation sequencing (ChIP-Seq).[52] These methods could lack sensitivity if DNA to transcription factor binding was weak. Cell-selective promoters also may not exist for some cell lineages. If cgRNAs sensing endogenous mRNAs could be engineered, this would have been game changing because it meant cgRNAs could enable cell-selective targeting against any cell type distinguishable by a set of marker genes.

The endogenous sequence detection experiments were conducted in bacteria using a variant of the medium throughput cgRNA screening assay described in Section 3.3. Here, the universal trigger RNAs were substituted for endogenous mRNA and microRNA transcripts from *galus galus* (chicken). Direct screening in chicken embryos and mammalian cells was avoided because these assay platforms were very low throughput and laborious to setup. The previous section showed it was possible to sense synthetic mRNA targets with the bacterial assay. Testing in bacteria was also advantageous because there was no background expression of *galus galus* endogenous mRNAs or microRNAs. Prior experience showed that if any hits were detected in bacteria, they were expected to translate well to the eukaryotic setting. Follow up validation experiments could be done in mammalian cell assays and chicken embryos to confirm the hits. This strategy was expected to reduce the overall workload.

The RNA transcripts (see Table 4.1) used in the endogenous sequence detection experiment were carefully chosen to hedge against various experimental uncertainties. d2eGFP was used as an mRNA detection control. This mRNA encoded green fluorescent protein, which could be used as a transfection control to confirm trigger mRNA expression status in chicken embryos. d2eGFP also had DNA probe binding data[40] (see Figure 4.3), which was used to identify accessible binding sites on the mRNA and triage for promising cgRNA sequences. PAX7, SNAI2, and TFAP2b were highly expressed endogenous mRNA transcripts specific to neural crest cells in developing chicken embryos. miR-130b, miR-203a, and miR-let7b were microRNAs also highly expressed in the neural crest cells of chicken embryos. Multiple mRNA and microRNA targets were screened to increase the chance of finding at least one cgRNA, which could selectively activate CRISPR machinery in neural crest cells. microRNAs were tried as an alternative to mRNAs because NUPACK

Name	Ascension number	Description
d2eGFP	MW247153.1	mRNA encoding green fluorescent protein
PAX7	NM_205065.1	a highly expressed mRNA in chicken neural crest cells
SNAI2	XM_419196.7	a highly expressed mRNA in chicken neural crest cells
TFAP2b	NM_204895.1	a highly expressed mRNA in chicken neural crest cells
mir130B	NR_031463.1	a highly expressed microRNA in chicken neural crest cells
mir203A	NR_031438.1	a highly expressed microRNA in chicken neural crest cells
mirLET7B	NR_031397.1	a highly expressed microRNA in chicken neural crest cells

Table 4.1: List of mRNA and microRNA sequences used in endogenous sequence detection experiments.

could predict the structure of short RNA sequences better than long RNA transcripts. However, generating designs against microRNAs was difficult because they formed hairpin structures which resisted cgRNA binding. Most cgRNA designs failed if the toehold domain length was too short. The $\Delta\Delta G$ energy gained from cgRNA binding was not high enough to overcome the microRNA duplex.

Exhaustive screening of all possible domain dimensions, cgRNA systems, and endogenous trigger RNA combinations was not possible within the timeframe of this thesis. Computational tools such as STORM and mRNA scanner (see Section 4.3) were used to triage the cloning and testing of promising cgRNA sequences. The following subsections show the results of these experiments.

Detection of endogenous mRNA with the split terminator switch

Chapter 3 and Hochrein *et al.* [66] showed that *S. pyogenes* Cas9 guide RNAs could be split into two sub-components along the terminator stem domain. This system was designed to switch from OFF to ON . In endogenous sequence detection experiments, the 3′ half of the guide RNA was treated as the trigger RNA. The 3′ half of the guide RNA was not sequence constrained. This meant mRNA sub-sequences could potentially act as the 3′ half of the guide RNA and help form the active guide RNA complex. Computational analysis of the endogenous sequences showed that the 3′ guide RNA motif was prevalent in many of the endogenous mRNA transcripts being tested (see Figure A.8 and A.9). A non-overlapping subset
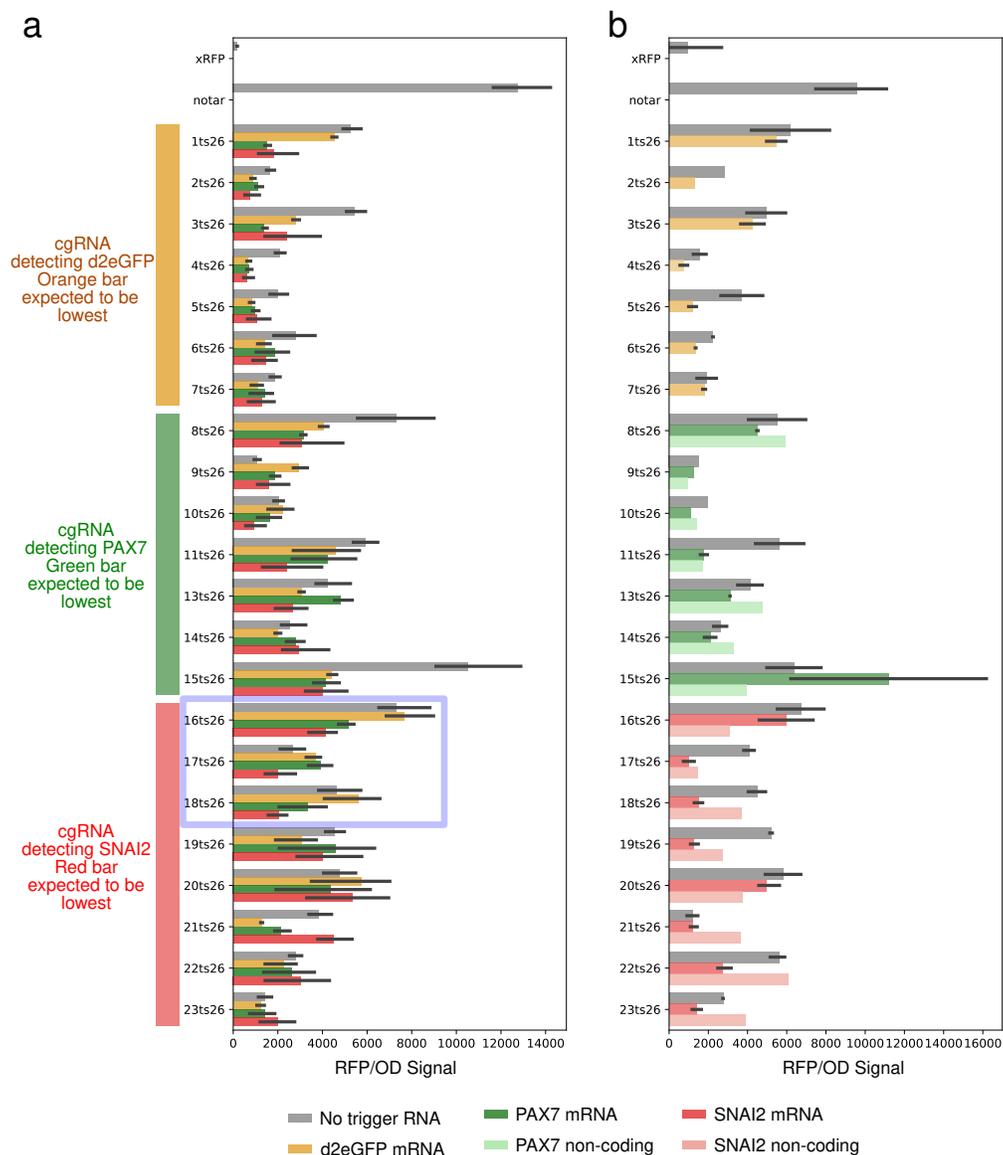
Figure 4.11: Detection of endogenous mRNAs with the split terminator switch. A dCas9-based RFP silencing assay in bacteria was used to measure cgRNA activity. Fluorescence was read on a plate reader and normalized with OD. This system was designed to switch from OFF to ON (high to low RFP signal). a) Orthogonality test for specificity of mRNA detection. Promising hits are boxed in blue . Error bars represent the standard deviation of N=3 replicates. b) Comparison of detection sensitivity between coding and non-coding transcripts. Colored bars are expected to be lower than the gray bars if the cgRNA sequence successfully detected the mRNA. Error bars represent the standard deviation of n=2 replicates.

of these sequences was selected using the STORM fold change metric and tested in the bacterial screening assay.

Detection of endogenous mRNA sub-sequences was not very successful. Out of all

cgRNA sequences tested, three sequences (16ts26, 17ts26, and 18ts26) seemed to be promising and exhibited some specificity for their cognate mRNA, but their fold change performance was poor (less than 2 fold change). Low fold change suggest the results are an artifact of metabolic noise rather than a true signal change. Non-specific detection was prevalent among all the samples tested (see Figure 4.11a, red boxes denote cgRNAs with significant off-target interactions). In these samples, the RFP signal for cognate and non-cognate mRNAs was significantly lower than the RFP signal of the OFF state where no mRNA was expressed. This suggests the cgRNA were promiscuously triggering off of cognate and non-cognate transcripts. Test tube analysis of the cgRNA sequences (using the mRNA scanner tool) revealed the orthogonality of the split terminator switch was predicted to be poor (see Figure A.10). Domain a was a long unpaired domain, and long unpaired domains have a tendency to bind non-specifically to many off-target sequences (see Figure 4.4). This could explain why this design lacked specificity.

Detection of endogenous mRNA with inhibited split terminator switch

The inhibited split terminator switch was created to improve the sequence specificity of the original split terminator switch system. This system was designed to switch from OFF to ON . Here, the 5′ half of the guide RNA is sequestered in an RNA hairpin. rendering it inactive. This acts as a lock and key like mechanism to ensure only the cognate mRNA sub-sequence could bind and activate the cgRNA.

Computational analysis predicted the inhibited split terminator switch system could outperform the split terminator switch system in orthogonality. Computational results predicted toehold length was a strong driver of sequence selectivity. Computationally, shorter toeholds reduced crosstalk at the cost of trigger sensitivity (see Figure A.11).

Unfortunately, toehold length and orthogonality predictions could not be verified experimentally because the sequences tested so far all failed to turn ON in response to co-expression of their cognate mRNAs. cgRNA sequences using the 5nt toehold all failed to sense their trigger RNAs (see Figure 4.13 and 4.14) because the toehold was too short. This occurred for both long and short RNA transcripts. However, a subset of cgRNAs using 10nt toeholds were successful at sensing truncated versions of their cognate endogenous transcripts. Samples 14ts38_1510 and 22ts38_1510 responded the best against short trigger RNAs (about 70nt long) derived from PAX7 and TFAP2b endogenous sequences respectively. Expanding the endogenous
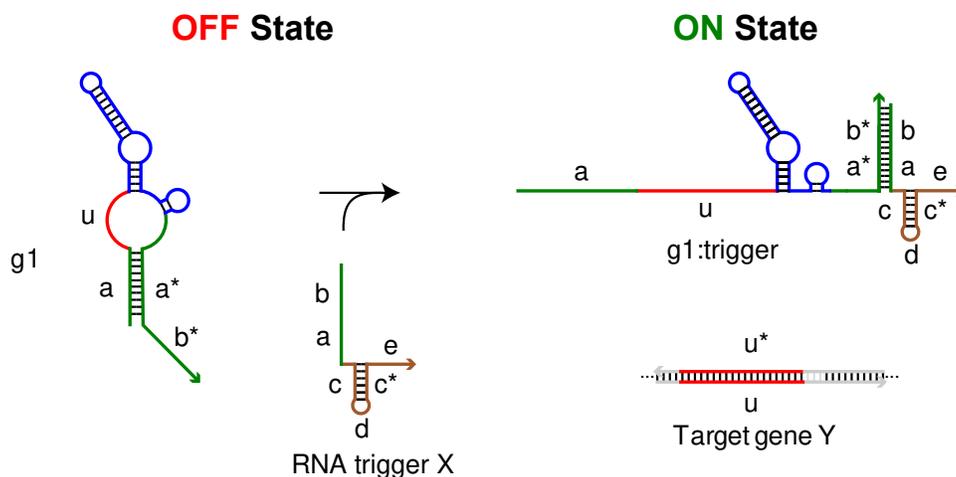
Figure 4.12: Schematic of the inhibited split terminator switch. $g1$ is initially sequestered by the duplex domain a. Co-expression of cognate trigger RNA binding to toehold b opens up the hairpin, forming the fully active guide RNA. Domain a provides some level of sequence specificity via substrate competition. A trigger RNA must unwind domain a, nucleotide by nucleotide, to form the active guide RNA complex. Red denotes the targeting domain. Blue denotes the Cas9 handle. Brown denotes the terminator domains, which can be non-canonical. Green denotes the variable length stem domains designed by NUPACK. Figure 4.13 shows the experimental performance of this design.

sequence window by 50nt each side transformed the trigger into a medium length RNA transcript (about 150nt long). This had lower activity than the short trigger RNA and suggested that trigger RNAs lose sensitivity as they get longer. Inhibited split guide RNA designs could not trigger off of full length endogenous RNAs (see Figure 4.14).

These detection experiments could have failed for any number of reasons. Ribosome activity could interfere with trigger sequence detection. The local structure of the sub-sequence could be different between the truncated and full length transcripts. The energies of nucleic acid base pairing inside cells verse the test tube could be very different. These inconsistencies make it difficult to find optimal domain lengths and trigger RNA sub-sequences with our existing tools. Follow up work on endogenous sequence detection for this system must focus on finding an optimal hairpin dimension experimentally. This can be expedited via high-throughput cgRNA screening.
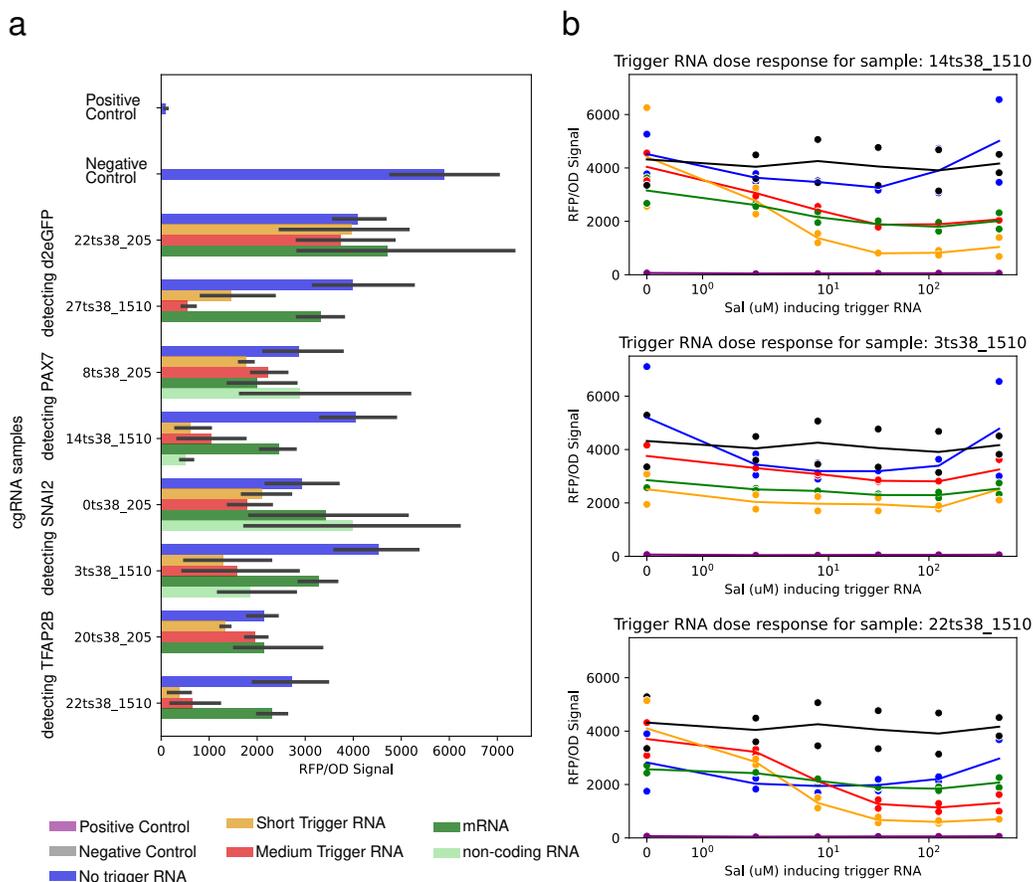
Figure 4.13: Detection of short and long endogenous RNA sequences with the inhibited split terminator switch. A dCas9-based RFP silencing assay in bacteria was used to measure cgRNA activity. Fluorescence was read on a plate reader and normalized with OD. This system was designed to switch from OFF to ON (high to low RFP signal). a) Bar plots of RFP signal for cgRNAs in an RFP silencing assay in bacteria. Blue bars representing no trigger RNA are expected to be the highest. Other colored bars are expected to be low if the cgRNA responded to the trigger input. Short trigger and medium trigger RNAs are sub-sequences of cognate mRNA for a given cgRNA and trigger pair. Error bars represent standard deviation of N=2 replicates. b) Dose response curves for a subset of cgRNA sequences shown in the bar plots. cgRNA expression is fixed while trigger RNA is expressed on a pSal promoter and induced with Salicylate. Markers represent replicates while lines represent the average of the replicate values. Data from scRNA_Cas9_20221110ZC

Detection of endogenous mRNAs and microRNAs with reverse split terminator switch system

The reverse split terminator switch system was designed to switch from ON to OFF (low to high fluorescence). This system inverts the logic of split guide RNA. Here, endogenous RNA sub-sequences would bind to *g*1 and break the guide RNA (see Figure 3.8). This system was designed to detect endogenous mRNAs and microRNAs. The mRNA targets consisted of d2eGFP, PAX7, and SNAI2. The

microRNA targets consisted of mir-130b, mir-203a, and mir-let7b. The work of the previous section showed that ribosome activity could disrupt trigger RNA binding. To help control for this phenomena, the mRNA sequences were expressed in bacteria with and without ribosome binding sites to see how protein translation may affect the ability of this cgRNA to detect the transcript.

A small set of reverse split terminator switch cgRNAs detecting mRNAs and microRNAs were designed and triaged using the STORM scoring metric. These sequences were cloned, and tested in a dCas9-based RFP silencing assay in bacteria. Fluorescence read out was acquired using a plate reader and normalized with OD. The RFP silencing assay failed to find successful hits against endogenous mRNAs (see Figure 4.15) and microRNAs (see Figure 4.16). This system could have failed because it was not very sensitivity to trigger RNA. In synthetic trigger RNA detection experiments, trigger RNA often had to be over expressed relative to the cgRNA. Otherwise, fold change signal could not be detected. Only one domain dimension for this design could be tried. The limited sampling size could be another reason this approach failed, as we could be using the wrong dimensions needed to sense endogenous mRNA. Unfortunately, testing a combination of domain dimensions and mRNA targets was not feasible within the timeframe of this thesis. These ideas could be pursued further in a high-throughput cgRNA screening setup. This will be discussed in detail in the Future Work chapter.

Detection of endogenous mRNA with two hairpin cgRNA

Several two hairpin systems were designed to detect PAX7 mRNA. These cgRNA sequences were triaged using the STORM scoring metric and tested in a dCas9-based RFP silencing assay in bacteria. This system was designed to switch from OFF to ON (high to low fluorescence). For short RNA triggers, the two hairpin cgRNA system had better sensitivity than the reverse split terminator switch system. Unlike the split terminator switch, the trigger sequence is not constrained by structural motifs. These features were expected to improve the prospects of finding a good cgRNA hit.

The two hairpin cgRNA system was able to detect short trigger RNAs representing sub-sequences in the endogenous mRNA of PAX7 (see Figure 4.17 and 4.18). Samples 0ts23, 1ts23, and 13ts23 were able to switch from OFF to ON (high to low fluorescence) in response to expression of the short trigger RNA. These samples could also switch ON in response to full length PAX7, but their activity was very

reduced. This confirmed endogenous sub-sequence detection was possible. However, the fold change ON/OFF performance achieved was not meaningful enough to warrant their use in follow up validation studies with mammalian cell assays and chick embryo experiments.

This experiment could have failed for several reasons. First, the two hairpin cgRNAs tested so far all used the same domain dimensions. The affinity of trigger RNA and cgRNA is a function of both sequence content and domain length. Fixing the domain length could have excluded other promising sub-sequence windows that could have worked with a different domain length. Second, the sampling size could have been too small. To have a better chance of success, we should screen cgRNAs against multiple mRNAs and use different domain dimensions. This was not pursued because the current assay setup cannot handle processing of many samples. We plan to address this issue with a high-throughput screening workflow, which will be discussed in the Future Work chapter. Finally, the loss of trigger RNA sensitivity suggests that pseudoknot interactions may play an out-sized role in the structure of long RNA transcripts.[40] NUPACK cannot predict the structure of long RNA sequences because pseudoknots and tertiary interactions are important at these length scales. If endogenous sequence detection is to be successful, these complexities cannot be ignored.

4.6 Conclusions

Our preliminary foray into endogenous sequence detection offered tantalizing clues that endogenous mRNA detection was possible. Some cgRNA hits were found. Unfortunately, the fold change activity of these cgRNA sequences was suboptimal. These sequences would not be useful in chicken embryos studies, and more work is needed to screen for better hits. We may have been unlucky because the cgRNA sampling size was small. In the next round of work, we may try different combinations of domain dimensions, sub-sequence windows, and cgRNA system. The volume of cloning involved necessitates a transition towards a high-throughput screening platform. Work on this system will be discussed in the Future Works chapter. Nevertheless, the results of this chapter highlighted a few important lessons about the complexities of mRNA detection.

First, our inability to predict the structure of long RNA transcripts was a major hurdle in finding good cgRNA hits. Most of the short synthetic mRNA detection experiments were successful. These experiments used synthetic trigger sequences

which were non-endogenous and known to be unstructured. The two hairpin cgRNA detection experiments showed that the high performing cgRNA systems developed in Chapter 3 had no problem detecting short trigger RNAs. However, they still had great difficulty sensing the same sub-sequences in the context of a longer transcripts. It is still not clear how pseudoknot interactions and ribosome activity make this harder to implement in long RNA transcripts. Perhaps coarse-grained molecular dynamics simulations such as oxDNA[7] could shed some light on this phenomena.

Second, the presence of a ribosome binding site was shown to somewhat disrupt trigger RNA binding. The reverse toehold switch, single hairpin split guide RNA, and two hairpin guide RNAs all lost some sensitivity when a ribosome binding site was added to the short RNA transcript. It was not clear if this would happen in longer RNA transcripts. It was also not clear if this would happen in a mammalian cell setting. To mitigate this issue, future endogenous sequence detection experiments could perhaps focus on detecting non-coding RNAs such as introns or microRNAs. If detection of mRNA transcripts is necessary, it is recommended that sub-sequences be drawn from the 3′ UTR. This was shown to work in synthetic mRNA detection with the terminator switch (see Figure 4.7).

Lastly, our best chance at getting endogenous RNA detection to work is to transition the screening process into a high-throughput workflow. The computational tools developed in this chapter have already laid some of the ground work for implementing an NGS based screening assay. We can already generate thousands of cgRNA oligos for cloning. The last step is to convert the fluorescence read out into an NGS read out. This is expected to dramatically increase our capability to screen for cgRNA sequences with a wide mixture of domain dimensions and sub-sequence targets. This workflow could expedite our ability to engineer cgRNAs for endogenous mRNA detection.

Figure 4.14: Orthogonal detection of endogenous mRNAs with inhibited split terminator switch cgRNAs. A dCas9-based RFP silencing assay in bacteria was used to measure cgRNA activity. Fluorescence was read on a plate reader and normalized with OD. This system was designed to switch from OFF to ON (high to low RFP signal). Promising hits are boxed in blue. Error bars represent the standard deviation of N=2 replicates. Data from scRNA_Cas9_20221024ZC
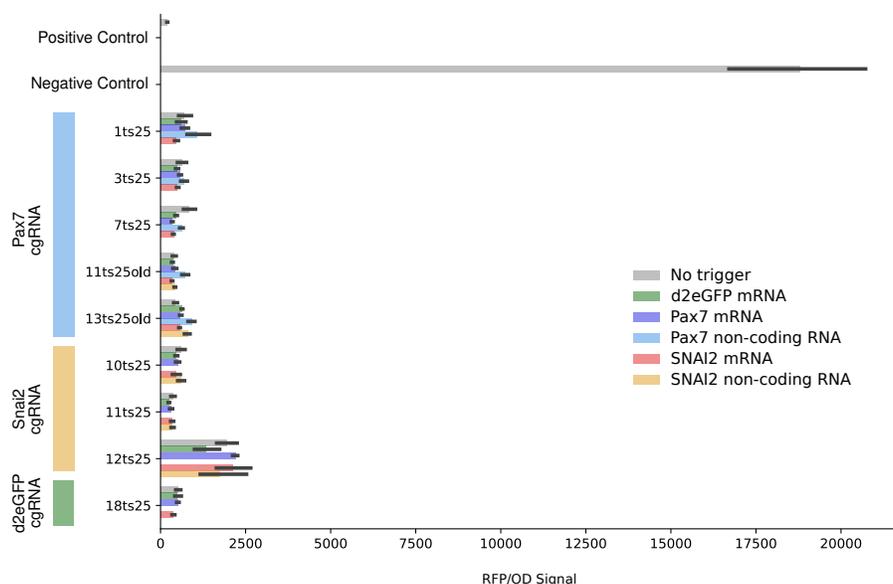
Figure 4.15: Detection of endogenous mRNA sequences with the reverse split terminator switch system. A dCas9-based RFP silencing assay in bacteria was used to measure cgRNA activity. Fluorescence was read on a plate reader and normalized with OD. Error bars represent standard deviation of N=3 replicates. This system was designed to switch from ON to OFF (low to high RFP signal). No promising hits were found. All cgRNA samples silenced RFP very well. Co-expression of cognate trigger RNAs could not inactivate the cgRNA. Data from scRNA_Cas9_20220921ZC
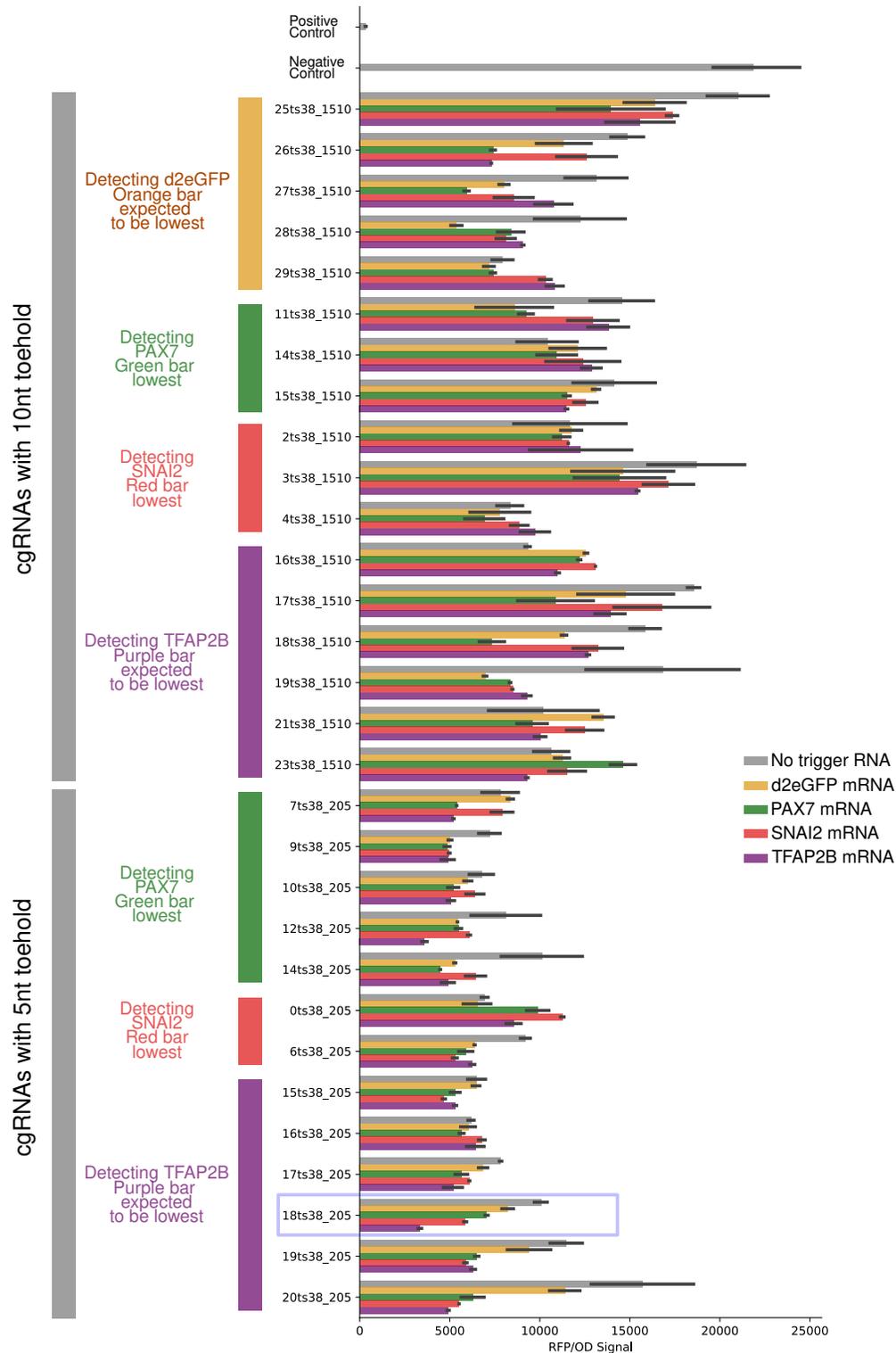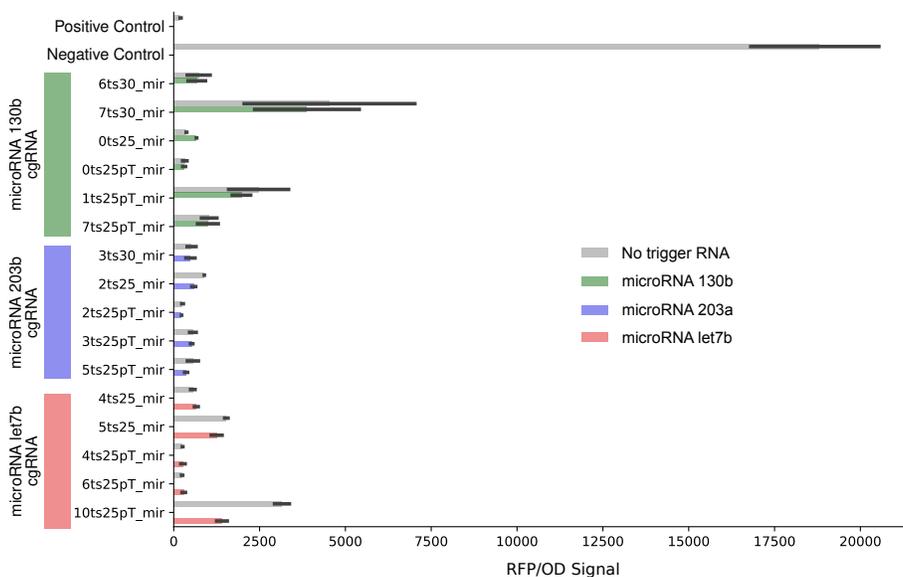

Figure 4.16: Detection of endogenous microRNAs with reverse split terminator switch system. A dCas9-based RFP silencing assay in bacteria was used to measure cgRNA activity. Fluorescence was read on a plate reader and normalized with OD. Error bars represent standard deviation of N=2 replicates. This system was designed to switch from ON to OFF (low to high RFP signal). No promising hits were found. All cgRNA samples silenced RFP very well. Co-expression of cognate trigger RNAs could not inactivate the cgRNA. Data from scRNA_Cas9_20220921ZC.
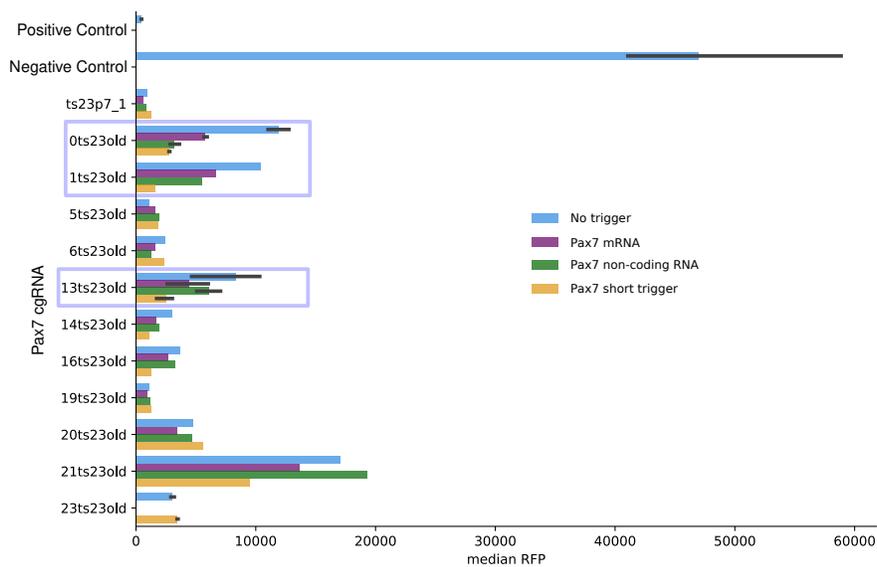
Figure 4.17: Detection of endogenous RNA sequences with the two hairpin cgRNA. A dCas9-based RFP silencing assay in bacteria was used to measure cgRNA activity. Fluorescence was read on a flow cytometer. Error bars represent the standard deviation of N=2 replicates. This system was designed to switch from OFF to ON (high to two RFP signal). Promising hits are boxed in blue . Detection of short trigger RNA was successful in many of these samples. However, detection of the same sequence in the longer transcript as an mRNA or non-coding RNA was not always possible. Data from scRNA_Cas9_20220923ZC.
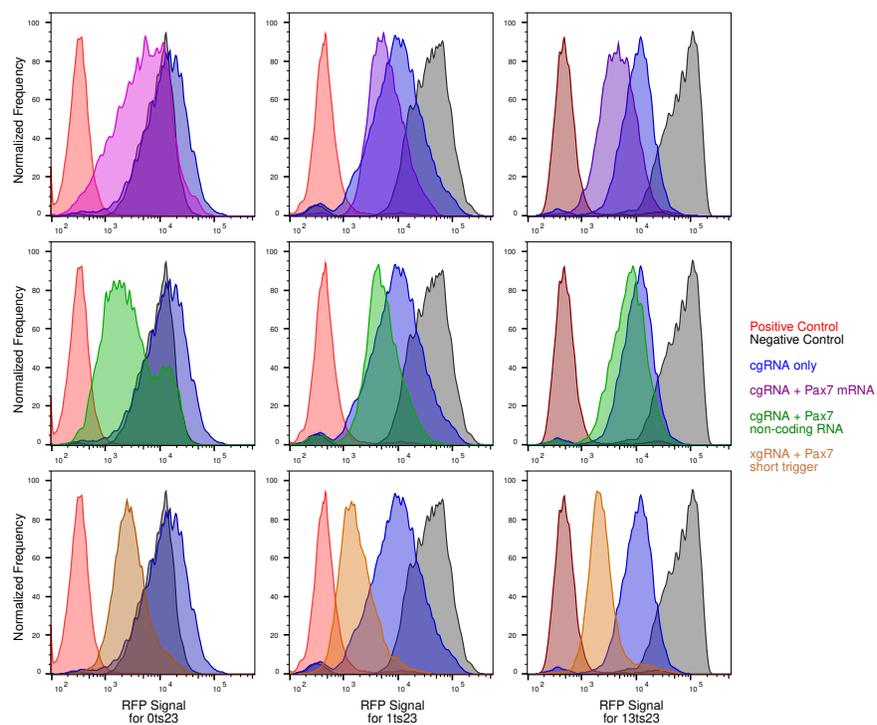
Figure 4.18: Flow cytometry plot of the two hairpin cgRNA detecting Pax7 mRNA. This system was designed to switch from OFF to ON (high to two RFP signal). These plots show the fluorescence of the cell population from the promising hits boxed in Figure 4.17. Columns represent the same cgRNA samples. The rows denote the different RNA transcripts tested. Data from scRNA_Cas9_20220923ZC.

C h a p t e r   5

FUTURE WORK

cgRNAs have the potential to revolutionize the way we interface with the cell and perturb the genome. Over the course of this thesis, I was able to show that NUPACK can engineer conditional guide RNAs (cgRNAs) *in silico* and have them work *in vivo*. These first generation systems were then redesigned to have fewer design constraints, enabling them to sense a wider variety of trigger RNA sequences and structures. Finally, several proof of concept experiments with synthetic mRNA showed that mRNA detection with cgRNAs was feasible. However, this project could not be brought to a satisfying conclusion because robust detection of endogenous mRNAs could not be achieved. The limited screening capacity and/or shortcomings in the computational model remain major hurdles towards making cgRNAs useful. The purpose of this chapter is to discuss future work that could help overcome these hurdles. Here, I detail the development of a high-throughput screening assay for cgRNA activity. I also discuss how the thermodynamic parameters underpinning NUPACK could be re-calibrated with empirical screening data so that high-performing cgRNA sequences could be found quicker in the *in silico* screening step.

5.1   High-rhroughput cgRNA screening

A major bottleneck in developing new cgRNA systems and finding high-performing hits is the pace at which each sequence can be assessed experimentally. There is no shortage of possible sequences which need to be tested. However, one student simply cannot clone fast enough.

A next generation sequencing (NGS) based high-throughput screening approach offers a promising path forward. In an NGS workflow, cloning, sequence verification, and cgRNA performance assessment can be merged into one step. This approach would eliminate many of the labor intensive tasks associated with traditional cloning such as colony PCR and plasmid purification (see Figure 5.2). In this workflow, cloning of each cgRNA into a construct can be done in bulk via ligation of a cgRNA library pool into a cloning vector. The vectorized cgRNA library can then be transformed directly into a bacterial test strain expressing the cognate or non-cognate trigger RNA on an inducible promoter. At this stage of the assay, each bacteria may

contain an empty vector, a misassembled cgRNA expression cassette, or a correctly assembled cgRNA cassette. Flow cytometry is used to eliminate non-functional constructs from this library pool.
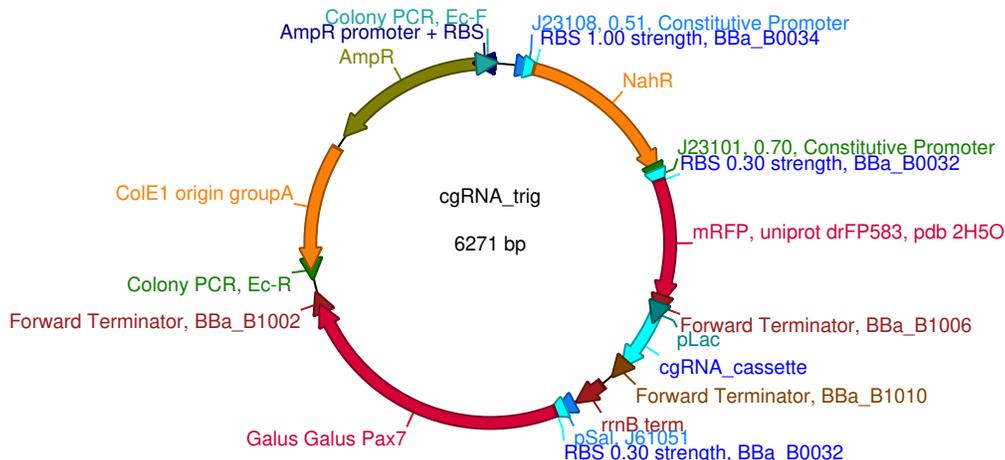
Figure 5.1: Illustration of the cloning vector used in the the high-throughput screening setup (see Figure 5.2). cgRNA oligos are inserted into this cloning vector behind the pLac promoter. This plasmid expresses galus galus PAX7 as the trigger RNA on a pSal promoter. PAX7 can be induced with salicylate. NahR expression is supplemented into this plasmid so the pSal promoter can function with low promoter leak. mRFP is used as the fluorescence read out for the dCas9-based RFP silencing assay in bacteria. The genbank file for this plasmid is in Table A.17.

In the first step of the NGS screen, cells with high or low fluorescence are selected via cell sorting. In a dCas9-based RFP silencing assay, high fluorescence samples would correspond to cgRNAs with a good OFF state, meaning the cgRNA is inactive. Low fluorescence samples correspond to samples with a good ON state, meaning the cgRNA is actively silencing RFP. These cells are collected with the cell sorter and cultured overnight. At this stage of the assay, the trigger RNA is induced with varying concentrations of salicylate or IPTG. This shifts the cgRNA conformation from an inactive to active state or vice versa.

In the second step of the NGS screen, cells grown with inducer are sorted for low and high fluorescence. A promising hit for an OFF to ON system in the dCas9-based RFP silencing assay would start out with high fluorescence in the first stage of the NGS screen and end with low RFP signal in the second stage of the NGS screen. The converse would happen for an ON to OFF cgRNA system. This stage enriches for functional phenotypes. If trigger RNA was induced at different concentrations after the first stage of the screen, this could be used to infer the sensitivity of the cgRNA sequence for the trigger RNA. High-performing cgRNA systems would require less
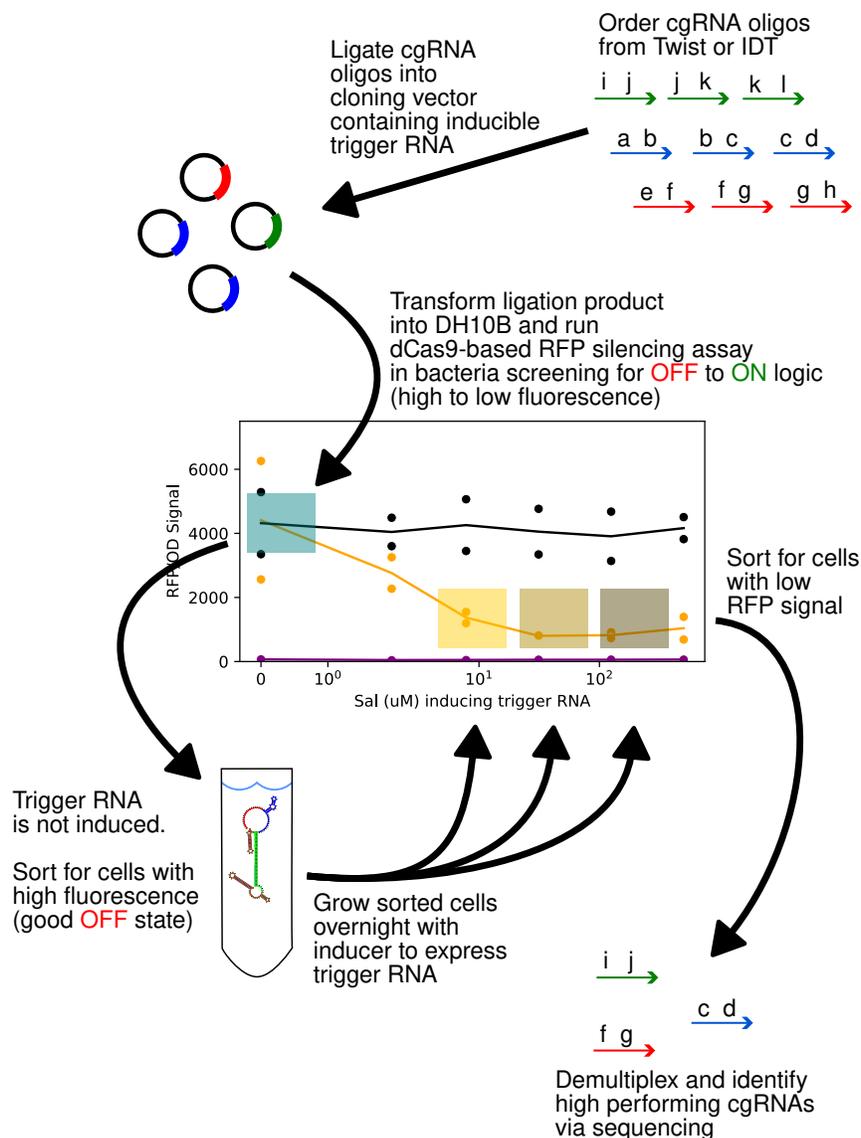
Figure 5.2: Illustration of the high-throughput screening setup. Step 1, cgRNA oligos are ordered from Twist Biosciences or IDT and inserted into the cloning vector (see Figure 5.1) via Golden Gate or Gibson assembly. Step 2, the cgRNA library is bulk transformed into DH10B bacteria containing the dCas9 plasmid and run in the dCas9-based RFP silencing assay for OFF to ON logic cgRNAs. Trigger RNA initially is not induced. Cells with high fluorescence are collected to enrich for cgRNAs with good OFF states. Step 2, these cells are grown overnight with inducer to express trigger RNA. Cells with low fluorescence are collected to enrich for cells with good ON states. Colony PCR is performed on the cells. The PCR product is sequenced to identify the high performing cgRNA oligos which had good OFF states and good ON states. Data for the dose response curve was from Figure 4.13, which was running a similar assay setup.

inducer to switch from inactive to active conformations or vice versa. Cells which fall in the lower range of inducer titrations would be enriched with high-performing

cgRNA sequences.

In the final stage of the NGS screen, the sorted cells from the second step would be cultured in bulk and purified for plasmids in preparation for sequencing with Nanopore or Illumina MiSeq. Colony PCR can be performed on the purified plasmids to attach sequencing adapters and unique molecular identifiers (UMI) to the cgRNA expression cassettes. UMIs are barcoded sequences attached 5′ and 3′ of a DNA transcript, which are used to identify the sample source. This could be used to de-multiplex samples from different trigger inducer concentrations and/or identify the endogenous trigger mRNAs used. If samples are sent out for MiSeq, the paired end sequencing workflow should be used to recover cgRNA cassettes up to 500bp in length via merger of the paired end reads.* If cgRNA cassettes are longer than 500bp, Nanopore sequencing would be the preferred workflow. Nanopore sequencing and consensus error correction techniques could be used to recover extremely long transcripts with accuracies as good as MiSeq.[92] Whatever the sequencing platform used, the final output from the NGS workflow is a set of sequencing reads which can be aligned against the UMIs and cloning oligos to identify the experimental condition and cgRNA identity.

Data analysis in the NGS screen would involve building a count matrix of cgRNA sequences and experimental conditions. Experimental conditions could represent the amount of inducer used to express trigger RNA and the identity of the trigger RNA. After count normalization, this information could be used to generate a dose response curve and/or an orthogonality matrix for each cgRNA design. The notable advantage of this kind of platform is the ability to compare cgRNA performance in the same experiment across a large number of samples and test conditions. This is important for generating a comprehensive data set which could be used to calibrate the RNA secondary structure model underlying to make it more suitable for modeling cgRNA performance inside living cells.

5.2   Data driven approaches for RNA structure prediction and rational design

Computational predictions of RNA and protein folding are fraught with inaccuracies, and these inaccuracies could be a major hurdle towards getting RNA nanotechnology to work robustly inside of living cells. The work of Chapter 2 and 3 revealed many discrepancies between test tubes predictions and experimental results. Thermodynamic models of RNA structure modeled in a dilute solution of nucleic acid

*Illumina MiSeq SBS kit v3 enables paired end reads of 2x301bp.

strands at 1M Na⁺ are not necessarily suitable for predicting RNA structure inside of a cell. Although NUPACK could generate high-performing cgRNA sequences (see Figure 3.4, sample ts45m1-1), not all cgRNA sequences generated were useful or functional. This was frustrating because laborious experimental screening was required to weed out low-performing sequences. If the parameters of the RNA secondary structure model could be improved via some sort of calibration experiment, this could dramatically reduce the experimental workload and improve our ability to tackle challenging problems such as mRNA detection and multi-layered RNA biocircuit design.

The advent of *in vivo* RNA strand displacement systems means there are now many ways to characterize the dynamics and kinetics of RNA strand displacement inside of a cell. The toehold switch mRNA system is one such example where a comprehensive biological data set of toehold switch performance is available for characterizing the biophysics of RNA strand displacement inside of a cell. This data set has been used to train machine learning models which predicted toehold switch performance with greater accuracy than the RNA structure model underlying NUPACK.[86] More of these data sets will become available in the next few years as people try to engineer cgRNAs or similar systems, and we should begin thinking about how these data sets can be used to refine or augment thermodynamic models of RNA secondary structure.

In the last decade, advances in computational power and algorithms to train statistical models have helped bring about many data-driven approaches for predicting RNA folding. CycleFold is an example of a data-driven approach for prediction of small RNA tertiary structures.[93] Instead of thermodynamic nearest neighbor parameters, base pairing probabilities are computed from a training set of RNA tertiary structures curated from the Protein Data Bank. This model is better at capturing non-canonical base pairing in extended secondary structures such as pseudoknots. SPOT-RNA is an example where deep neural networks are used to predict RNA secondary structure.[94] These models were trained on an annotated set of RNA secondary structures curated from the Protein Data Bank. For their given validation set, this method could predict canonical base pairing, noncanonical base pairing, and pseudoknots better than CycleFold and RNAfold. RNAfold and NUPACK share similar underlying models for predicting RNA secondary structure. These results suggest data-driven deep learning models could outperform free energy models of RNA structure. Data set scarcity is a major hurdle in improving the generality of

these data-driven approaches. This is expected to change in the next decade because NGS screening approaches such as those described in Valeri *et al.* [86] and the previous section could help alleviate this bottleneck.

My approach in Chapter 4 of using STORM to rank and triage for high-performing cgRNA sequences was an imperfect attempt to merge data driven models with free energy models. This approach may have been flawed because STORM was designed to predict the performance of toehold switch mRNAs. Those predictions may not be generalizable to cgRNAs. If the high-throughput screening approach described in the previous section could be properly implemented, the results from a cgRNA NGS screen could be used to training a deep learning model specific for predicting cgRNA performance. The existing bioinformatic framework described in Chapter 4 already provides a promising path forward for testing this hypothesis.

Ultimately, we want a way to calibrate free energy models against experimental data. This first principles approach is important for controlling against biases in a data set that do not make physical sense, such as a machine learning model claiming an RNA hairpin could be formed from a long sequence of uracil nucleotides. This approach could simply be performing high-dimensional optimization on the underlying thermodynamic parameters until test tube predictions match experimental results. Example experiments could be toehold length optimization, stem length optimization for the split terminator switch, and allosteric competition between trigger RNA and $g2$ in the reverse split terminator switch (see Figure 3.8). These systems are ideal for studying the biophysics of RNA strand displacement because their interaction energies are not believed to be confounded by protein interactions. The terminator stem structure and trigger RNA strands lie on the exterior of the Cas9 protein. If these data driven approaches could be integrated into the RNA structure prediction model underlying NUPACK, we could have a more robust and general tool for engineering nucleic acid strand displacement in living cells. This would enable us to simulate, test, re-calibrate, and engineer a wide range of DNA/RNA nanomachines in living systems with greater accuracy and speed than current rational design approaches.

BIBLIOGRAPHY

1. Markham, N. R. & Zuker, M. UNAFold. *Bioinformatics,* 3–31. doi:10.1007/978-1-60327-429-6_1 (2008).

2. Fornace, M. E., Porubsky, N. J. & Pierce, N. A. A unified dynamic programming framework for the analysis of interacting nucleic acid strands: Enhanced models, scalability, and speed. *ACS Synthetic Biology* **9,** 2665–2678. doi:10.1021/acssynbio.9b00523 (2020).

3. Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry* **32,** 170–173. doi:10.1002/jcc.21596 (2011).

4. Xayaphoummine, A., Bucher, T. & Isambert, H. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research* **33,** W605–W610. doi:10.1093/nar/gki447 (2005).

5. Watkins, A. M., Rangan, R. & Das, R. FARFAR2: improved de novo rosetta prediction of complex global RNA folds. *Structure* **28,** 963–976. doi:10.1016/j.str.2020.05.011 (2020).

6. Kappel, K. & Das, R. Sampling native-like structures of RNA-protein complexes through Rosetta folding and docking. *Structure* **27,** 140–151. doi:10.1016/j.str.2018.10.001 (2019).

7. Šulc, P., Romano, F., Ouldridge, T. E., Doye, J. P. & Louis, A. A. A nucleotide-level coarse-grained model of RNA. *The Journal of Chemical Physics* **140,** 06B614_1. doi:10.1063/1.4881424 (2014).

8. Poppleton, E., Romero, R., Mallya, A., Rovigatti, L. & Šulc, P. OxDNA.org: a public webserver for coarse-grained simulations of DNA and RNA nanostructures. *Nucleic acids research* **49,** W491–W498. doi:10.1093/nar/gkab324 (2021).

9. Russo, J. *et al.* SAT-assembly: A new approach for designing self-assembling systems. *Journal of Physics: Condensed Matter.* doi:10.1088/1361-648X/ac5479 (2022).

10. Bohlin, J. *et al.* Design and simulation of DNA, RNA and hybrid protein–nucleic acid nanostructures with oxView. *Nature Protocols,* 1–27. doi:10.1038/s41596-022-00688-5 (2022).

11. Bohlin, J., Turberfield, A. J., Louis, A. A. & Šulc, P. Designing the self-assembly of arbitrary shapes using minimal complexity building blocks. *arXiv preprint arXiv:2207.06954.* doi:10.48550/arXiv.2207.06954 (2022).

12. Dirks, R. M. & Pierce, N. A. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry* **24,** 1664–1677. doi:10.1002/jcc.10296 (2003).

13. SantaLucia Jr, J. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences* **95,** 1460–1465. doi:10.1073/pnas.95.4.1460 (1998).

14. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology* **288,** 911–940. doi:10.1006/jmbi.1999.2700 (1999).

15. Lu, Z. J., Turner, D. H. & Mathews, D. H. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic acids research* **34,** 4912–4924. doi:10.1093/nar/gkl472 (2006).

16. Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research* **38,** D280–D282. doi:10.1093/nar/gkp892 (2010).

17. Dirks, R. M. & Pierce, N. A. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *Journal of Computational Chemistry* **25,** 1295–1304. doi:10.1002/jcc.20057 (2004).

18. Zadeh, J. N., Wolfe, B. R. & Pierce, N. A. Nucleic acid sequence design via efficient ensemble defect optimization. *Journal of Computational Chemistry* **32,** 439–452. doi:10.1002/jcc.21633 (2011).

19. Douglas, S. M. *et al.* Rapid prototyping of 3D DNA-origami shapes with caDNAno. *Nucleic acids research* **37,** 5001–5006. doi:10.1093/nar/gkp436 (2009).

20. Kim, D.-N., Kilchherr, F., Dietz, H. & Bathe, M. Quantitative prediction of 3D solution shape and flexibility of nucleic acid nanostructures. *Nucleic acids research* **40,** 2862–2868. doi:10.1093/nar/gkr1173 (2012).

21. Pan, K., Bricker, W. P., Ratanalert, S. & Bathe, M. Structure and conformational dynamics of scaffolded DNA origami nanoparticles. *Nucleic acids research* **45,** 6284–6298. doi:10.1093/nar/gkx378 (2017).

22. Choi, H. M. T. *et al.* Programmable in situ amplification for multiplexed imaging of mRNA expression. *Nature Biotechnology* **28.** PMCID: PMC4046802, 1208–12. doi:10.1038/nbt.1692 (2010).

23. Choi, H. M. T., Beck, V. A. & Pierce, N. A. Next-generation in situ hybridization chain reaction: higher gain, lower cost, greater durability. *ACS Nano* **8.** PMCID: PMC4046802, 4284–4294. doi:10.1021/nn405717p (2014).

24. Choi, H. M. T. *et al.* Mapping a multiplexed zoo of mRNA expression. *Development* **143.** PMCID: PMC5087610, 3632–3637. doi:10.1242/dev.140137 (2016).

25. Trivedi, V., Choi, H. M. T., Fraser, S. E. & Pierce, N. A. Multidimensional quantitative analysis of mRNA expression within intact vertebrate embryos. *Development* **145.** PMCID: PMC5825878. doi:10.1242/dev.156869 (2018).

26. Kishi, J. Y. *et al.* SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nature methods* **16,** 533–544. doi:10.1038/s41592-019-0404-0 (2019).

27. Yin, P., Choi, H. M., Calvert, C. R. & Pierce, N. A. Programming biomolecular self-assembly pathways. *Nature* **451,** 318–322. doi:10.1038/nature06451 (2008).

28. Thubagere, A. J. *et al.* A cargo-sorting DNA robot. *Science* **357,** eaan6558. doi:10.1126/science.aan6558 (2017).

29. Cherry, K. M. & Qian, L. Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. *Nature* **559,** 370–376. doi:10.1038/s41586-018-0289-6 (2018).

30. Rothemund, P. W. Folding DNA to create nanoscale shapes and patterns. *Nature* **440,** 297–302. doi:10.1038/nature04586 (2006).

31. Geary, C., Grossi, G., McRae, E. K., Rothemund, P. W. & Andersen, E. S. RNA origami design tools enable cotranscriptional folding of kilobase-sized nanoscaffolds. *Nature chemistry* **13,** 549–558. doi:10.1038/s41557-021-00679-1 (2021).

32. Tikhomirov, G., Petersen, P. & Qian, L. Fractal assembly of micrometre-scale DNA origami arrays with arbitrary patterns. *Nature* **552,** 67–71. doi:10.1038/nature24655 (2017).

33. Jungmann, R. *et al.* Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. *Nature methods* **11,** 313–318. doi:10.1038/nmeth.2835 (2014).

34. Hu, C. Y. & Murray, R. M. Layered feedback control overcomes performance trade-off in synthetic biomolecular networks. *Nature communications* **13,** 1–13. doi:10.1038/s41467-022-33058-6 (2022).

35. Gao, X. J., Chong, L. S., Kim, M. S. & Elowitz, M. B. Programmable protein circuits in living cells. *Science* **361,** 1252–1258. doi:10.1126/science.aat5062 (2018).

36. Vlahos, A. E. *et al.* Protease-controlled secretion and display of intercellular signals. *Nature communications* **13,** 1–12. doi:10.1038/s41467-022-28623-y (2022).

37. Olsman, N., Xiao, F. & Doyle, J. C. Architectural principles for characterizing the performance of antithetic integral feedback networks. *Iscience* **14,** 277–291. doi:10.1016/j.isci.2019.04.004 (2019).

38. Meyer, S., Chappell, J., Sankar, S., Chew, R. & Lucks, J. B. Improving fold activation of small transcription activating RNAs (STARs) with rational RNA engineering strategies. *Biotechnology and bioengineering* **113,** 216–225. doi:10.1002/bit.25693 (2016).

39. Hochrein, L. M., Schwarzkopf, M., Shahgholi, M., Yin, P. & Pierce, N. A. Conditional Dicer substrate formation via shape and sequence transduction with small conditional RNAs. *J. Am. Chem. Soc.* **135.** PMCID: PMC3842090, 17322–17330. doi:10.1021/ja404676x (2013).

40. Hochrein, L. M., Ge, T. J., Schwarzkopf, M. & Pierce, N. A. Signal transduction in human cell lysate via dynamic RNA nanotechnology. *ACS Synthetic Biology* **7.** PMCID: PMC6305621, 2792–2802. doi:10.1021/acssynbio.8b00424 (2018).

41. Green, A. A., Silver, P. A., Collins, J. J. & Yin, P. Toehold switches: de-novo-designed regulators of gene expression. *Cell* **159,** 925–939. doi:10.1016/j.cell. 2014.10.002 (2014).

42. Ma, D., Shen, L., Wu, K., Diehnelt, C. W. & Green, A. A. Low-cost detection of norovirus using paper-based cell-free systems and synbody-based viral enrichment. *Synthetic Biology* **3,** ysy018. doi:10.1093/synbio/ysy018 (2018).

43. Ma, D. *et al.* Multi-arm RNA junctions encoding molecular logic unconstrained by input sequence for versatile cell-free diagnostics. *Nature biomedical engineering* **6,** 298–309. doi:10.1038/s41551-022-00857-7 (2022).

44. Zhao, E. M. *et al.* RNA-responsive elements for eukaryotic translational control. *Nature Biotechnology* **40,** 539–545. doi:10.1038/s41587-021-01068-2 (2022).

45. Hochrein, L. M. *Towards Conditional RNAi: Shape and Sequence Transduction with Small Conditional RNAs* PhD thesis (California Institute of Technology, 2013). doi:10.7907/NTZT-8Q67.

46. Han, S.-p. *et al.* Programmable siRNA pro-drugs that activate RNAi activity in response to specific cellular RNA biomarkers. *Molecular Therapy-Nucleic Acids* **27,** 797–809. doi:10.1016/j.omtn.2021.12.039 (2022).

47. 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. **5.** PMCID: PMC6661866, 1241–1249. doi:10.1021/acscentsci.9b00340 (2019).

48. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183. doi:10.1016/j.cell. 2013.02.022 (2013).

49. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823. doi:10.1126/science.1231143 (2013).

50. Gandhi, S., Piacentino, M. L., Vieceli, F. M. & Bronner, M. E. Optimization of CRISPR/Cas9 genome editing for loss-of-function in the early chick embryo. *Developmental biology* **432,** 86–97. doi:10.1016/j.ydbio.2017.08.036 (2017).

51. Flytzanis, N., Goeden, N., Pickel, J. & Gradinaru, V. Engineered AAVS for CNS Transduction and Peripheral Organ De-Targeting across Species after Systemic Delivery. *Molecular Therapy* **27,** Art–No. doi:10.1016/j.ymthe.2019.04.004 (2019).

52. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858. doi:10.1038/nature07730 (2009).

53. Ng, A. H. *et al.* A comprehensive library of human transcription factors for cell fate engineering. *Nature biotechnology* **39,** 510–519. doi:10.1038/s41587-020-0742-6 (2021).

54. Challis, R. C. *et al.* Adeno-associated virus toolkit to target diverse brain cells. *Annual Review of Neuroscience* **45.** doi:10.1146/annurev-neuro-111020-100834 (2022).

55. Chuapoco, M. R. *et al.* Intravenous gene transfer throughout the brain of infant Old World primates using AAV. *bioRxiv.* doi:10.1101/2022.01.08.475342 (2022).

56. Park, J., Demirer, G. S. & Cheung, L. S. Toolboxes for plant systems biology research. *Current Opinion in Biotechnology* **75,** 102692. doi:10.1016/j.copbio.2022.102692 (2022).

57. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *elife* **3,** e04766. doi:10.7554/eLife.04766 (2014).

58. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576,** 149–157. doi:10.1038/s41586-019-1711-4 (2019).

59. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nature reviews genetics* **19,** 770–788. doi:10.1038/s41576-018-0059-1 (2018).

60. Zalatan, J. G. *et al.* Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160,** 339–350. doi:10.1016/j.cell.2014.11.052 (2015).

61. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529,** 490–495. doi:10.1038/nature16526 (2016).

62. Briner, A. E. *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Molecular cell* **56,** 333–339. doi:10.1016/j.molcel.2014.09.019 (2014).

63. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156,** 935–949. doi:10.1016/j.cell.2014.02.001 (2014).

64. Siu, K.-H. & Chen, W. Riboregulated toehold-gated gRNA for programmable CRISPR–Cas9 function. *Nature chemical biology* **15,** 217–220. doi:10.1038/s41589-018-0186-1 (2019).

65. Bois, J. S. *et al.* Topological constraints in nucleic acid hybridization kinetics. *Nucleic Acids Research* **33.** PMCID: PMC1180668, 4090–4095. doi:10.1093/nar/gki721 (2005).

66. Hochrein, L. M., Li, H. & Pierce, N. A. High-performance allosteric conditional guide RNAs for mammalian cell-selective regulation of CRISPR/Cas. *ACS Synthetic Biology* **10,** 964–971. doi:10.1021/acssynbio.1c00037 (2021).

67. Saha, A., Arantes, P. R. & Palermo, G. Dynamics and mechanisms of CRISPR-Cas9 through the lens of computational methods. *Current Opinion in Structural Biology* **75,** 102400. doi:10.1016/j.sbi.2022.102400 (2022).

68. Ricci, C. G. *et al.* Deciphering off-target effects in CRISPR-Cas9 through accelerated molecular dynamics. *ACS central science* **5,** 651–662. doi:10.1021/acscentsci.9b00020 (2019).

69. Zheng, W. Probing the structural dynamics of the CRISPR-Cas9 RNA-guided DNA-cleavage system by coarse-grained modeling. *Proteins: Structure, Function, and Bioinformatics* **85,** 342–353. doi:10.1002/prot.25229 (2017).

70. Jinek, M. *et al.* A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *science* **337,** 816–821. doi:10.1126/science.1225829 (2012).

71. Zhang, D. Y. & Winfree, E. Control of DNA strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society* **131,** 17303–17314. doi:10.1021/ja906987s (2009).

72. Srinivas, N. *et al.* On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic acids research* **41,** 10641–10658. doi:10.1093/nar/gkt801 (2013).

73. Antebi, Y. E. *et al.* Combinatorial signal perception in the BMP pathway. *Cell* **170,** 1184–1196. doi:10.1016/j.cell.2017.08.015 (2017).

74. Klumpe, H. E. *et al.* The context-dependent, combinatorial logic of BMP signaling. *Cell Systems.* doi:10.1016/j.cels.2022.03.002 (2022).

75. Su, C. J. *et al.* Ligand-receptor promiscuity enables cellular addressing. *Cell systems.* doi:10.1016/j.cels.2022.03.002 (2022).

76. Razo-Mejia, M. *et al.* Tuning transcriptional regulation through signaling: a predictive theory of allosteric induction. *Cell Systems* **6,** 456–469. doi:10.1016/j.cels.2018.02.004 (2018).

77. Dirks, R. M. & Pierce, N. A. Triggered amplification by hybridization chain reaction. *Proceedings of the National Academy of Sciences* **101,** 15275–15278. doi:10.1073/pnas.0407024101 (2004).

78. Zetsche, B. *et al.* Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array. *Nature biotechnology* **35,** 31–34. doi:10.1038/nbt.3737 (2017).

79. Smargon, A. A. *et al.* Cas13b is a type VI-B CRISPR-associated RNA-guided RNase differentially regulated by accessory proteins Csx27 and Csx28. *Molecular cell* **65,** 618–630. doi:10.1016/j.molcel.2016.12.023 (2017).

80. McNabb, D. S., Reed, R. & Marciniak, R. A. Dual luciferase assay system for rapid assessment of gene expression in Saccharomyces cerevisiae. *Eukaryotic cell* **4,** 1539–1549. doi:10.1128/EC.4.9.1539-1549.2005 (2005).

81. Gerdes, K., Gultyaev, A. P., Franch, T., Pedersen, K. & Mikkelsen, N. D. Antisense RNA-regulated programmed cell death. *Annual review of genetics* **31,** 1. doi:10.1146/annurev.genet.31.1.1 (1997).

82. Faridani, O. R., Nikravesh, A., Pandey, D. P., Gerdes, K. & Good, L. Competitive inhibition of natural antisense Sok-RNA interactions activates Hok-mediated cell killing in Escherichia coli. *Nucleic acids research* **34,** 5915–5922. doi:10.1093/nar/gkl750 (2006).

83. Van Melderen, L. Molecular interactions of the CcdB poison with its bacterial target, the DNA gyrase. *International journal of medical microbiology* **291,** 537–544. doi:10.1078/1438-4221-00164 (2001).

84. McCardell, R. D., Huang, S., Green, L. N. & Murray, R. M. Control of bacterial population density with population feedback and molecular sequestration. *Biorxiv,* 225045. doi:10.1101/225045 (2017).

85. Zhang, Y., Zhang, J. & Wang, W. Atomistic analysis of pseudoknotted RNA unfolding. *Journal of the American Chemical Society* **133,** 6882–6885. doi:10.1021/ja1109425 (2011).

86. Valeri, J. A. *et al.* Sequence-to-function deep learning frameworks for engineered riboregulators. *Nature communications* **11,** 1–14. doi:10.1038/s41467-020-18676-2 (2020).

87. Sternberg, J. B. & Pierce, N. A. Exquisite Sequence Selectivity with Small Conditional RNAs. *Nano Lett.* **14.** PMCID: PMC4134187, 4568–4572. doi:10.1021/nl501593r (2014).

88. Vieregg, J. R., Nelson, H. M., Stoltz, B. M. & Pierce, N. A. Selective nucleic acid capture with shielded covalent probes. *J. Am. Chem. Soc.* **135.** PMCID: PMC3703666, 9691–9699. doi:10.1021/ja4009216 (2013).

89. Jiang, K. *et al.* Programmable eukaryotic protein synthesis with RNA sensors by harnessing ADAR. *Nature Biotechnology,* 1–10. doi:10.1038/s41587-022-01534-5 (2022).

90. Oesinghaus, L. & Simmel, F. C. Switching the activity of Cas12a using guide RNA strand displacement circuits. *Nature communications* **10,** 1–11. doi:10.1038/s41467-019-09953-w (2019).

91. Oesinghaus, L. & Simmel, F. C. Controlling gene expression in mammalian cells using multiplexed conditional guide RNAs for Cas12a. *Angewandte Chemie International Edition* **60,** 23894–23902. doi:10.1002/anie.202107258 (2021).

92. Baloğlu, B. *et al.* A workflow for accurate metabarcoding using nanopore MinION sequencing. *Methods in Ecology and Evolution* **12,** 794–804. doi:10.1111/2041-210X.13561 (2021).

93. Sloma, M. F. & Mathews, D. H. Base pair probability estimates improve the prediction accuracy of RNA non-canonical base pairs. *PLoS computational biology* **13,** e1005827. doi:10.1371/journal.pcbi.1005827 (2017).

94. Singh, J., Hanson, J., Paliwal, K. & Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature communications* **10,** 1–13. doi:10.1038/s41467-019-13395-9 (2019).

A p p e n d i x   A

MATERIALS AND METHODS

A.1   Cell culture, cloning, and assay setup

Cloning of plasmid constructs was performed via insertion PCR, Golden Gate assembly, or Gibson assembly of DNA oligos containing cgRNA constructs. Primers used in cloning were order from Integrated DNA technologies (IDT) or Twist Biosciences. Oligos were ligated into the plasmid with Quick Ligase (cat# M2200L) and digested with DpnI (cat# R0176L) to remove background vector. Ligated PCR products were transformed into DH10B (cat# C3019I) and DH5A (cat# C2987I) via heat shock at 42°C for 30 seconds. Cells were then chilled to 4°C and incubated with SoC recovery media (cat# B9020S) for 1 hour before plating on antibiotic resistance plates. Colony PCR was performed on bacterial colonies to verify the identity of the plasmid constructs. PCR products were sent to Laragen or Primordium labs for sequence verification with Sanger or Nanopore sequencing. Plasmids were purified from bulk culture using the standard protocol from the Qiagen miniprep kit (cat# 27106X4).

Bacterial test strains were generated via triple transformation of plasmid constructs into MG1655 from Stanley Qi[48] or DH10B bacteria. Cells were heat shocked at 42°C for 30 seconds and chilled at 4°C to induce uptake of the plasmids. After recovery in SoC media for 1 hour, bacteria were plated on triple resistance antibiotic plates and grown overnight. Test strains were picked from antibiotic plates and incubated for 6 hrs in preparation for seeding into the assay plate.

Assay plates was setup at 300uL per well in a 96 well plate with EZ-RDM (cat# M2105), antibiotics, anhydrotetracycline (aTc), IPTG, and Salicylate. 2uL/well of test strains were seeded into each well of the 96 well plate and grown from 12 to 24 hours at 37°C. Fluorescence read out was acquire using a Synergy Neo2 plate reader, MAC VYB flow cytometer, or Cytoflex S flow cytometer.

A.2   Flow cytometry and data analysis

Flow cytometry data was analyzed using FlowJo. Cells were gated using FSC-H and SSC-H. Singlets were gated using SSC-H and SSC-A (see Figure A.1). No gain compensation was used because fluorescence overlap between reporter genes

was low. Median fluorescence intensity of the cell populations was obtained with FlowJo. Fluorescence measurements were analyzed and plotted using pandas and matplotlib in python.
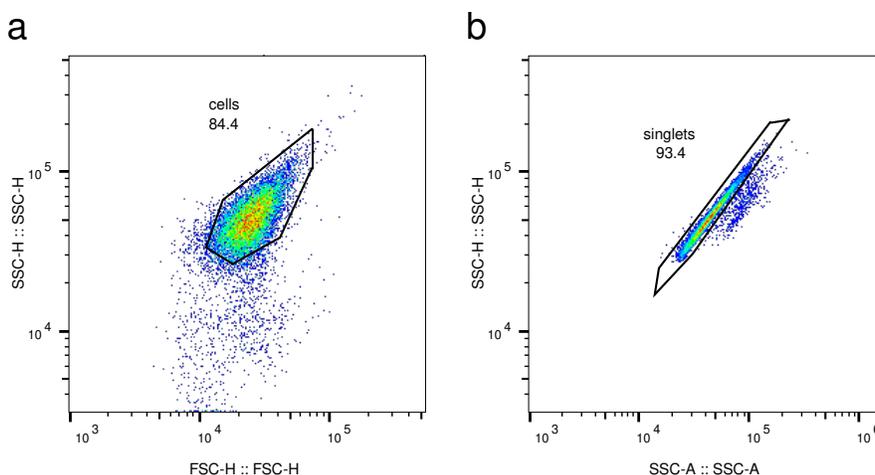
Figure A.1: Illustration of gating for cells in FlowJo. a) Gating of cells using FSC and SSC. b) Gating of singlets using SSC-H and SSC-A. Singlets are a sub-population within the live cell gate used to separate single cells from clusters of cells.

A.3 Rational design of orthogonal cgRNAs

Rational design of orthogonal cgRNAs for the terminator switch and toehold switch systems described in Chapter 2 are detailed in Hanewich-Hollatz *et al.* [47], supplemental materials section S1.1. cgRNA systems described in Chapter 3 utilized universal trigger RNAs (see Figure 3.21) to implemental orthogonality.

A.4 Computation of cgRNA performance metrics

Fold change, fractional dynamic range, crosstalk, and uncertainty propagation were computed according to the methods described in Hanewich-Hollatz *et al.* [47], supplemental materials section S1.4.

A.5 Sequences

Sequences inserted into the nfsA site

mRFP and sfGFP sequences inserted into the nfsA gene fragment of MG1655 bacteria

5'-ATGACGCCAACCATTGAACTTATTTGTGGCCATCGCTCCATTCGCCATTTCACTGATGAA
CCCATTTCCGAAGCGCAGCGTGAGGCGATTATTAACAGCGCCCGTGCGACGTCCAGTTCCAGT
TTTTTGGCATGCCCGATCAACGTCTCATTTTCGCCAGATATCGACGTCGACACCATCGAATGG
TGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAGTCAATTCAGGGTGGTGAAT

TTGACAGCTAGCTCAGTCCTAGGTATAATAGATCTGAATTCATTAAAGAGGAGAAAGGTACCA
TGAGCAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGTGATG
TTAATGGGCACAAATTTTCTGTCCGTGGAGAGGGTGAAGGTGATGCTACAAACGGAAAACTCA
CCCTTAAATTTATTTGCACTACTGGAAAACTACCTGTTCCGTGGCCAACACTTGTCACTACTC
TGACCTATGGTGTTCAATGCTTTTCCCGTTATCCGGATCACATGAAACGGCATGACTTTTTCA
AGAGTGCCATGCCCGAAGGTTATGTACAGGAACGCACTATATCTTTCAAAGATGACGGGACCT
ACAAGACGCGTGCTGAAGTCAAGTTTGAAGGTGATACCCTTGTTAATCGTATCGAGTTAAAGG
GTATTGATTTTAAAGAAGATGGAAACATTCTTGGACACAAACTCGAGTACAACTTTAACTCAC
ACAATGTATACATCACGGCAGACAAACAAAGAATGGAATCAAAGCTAACTTCAAAATTCGCC
ACAACGTTGAAGATGGTTCCGTTCAACTAGCAGACCATTATCAACAAAATACTCCAATTGGCG
ATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCGACACAATCTGTCCTTTCGAAAGATC
CCAACGAAAAGCGTGACCACATGGTCCTTCTTGAGTTTGTAACTGCTGCTGGGATTACACATG
GCATGGATGAGCTCTACAAATAAGGATCCAAACTCGAGTAAGGATCTCCAGGCATCAAATAAA
ACGAAAGGCTCAGTCGAAAGACTGGGCCTTTCGTTTTATCTGTTGTTTGTCGGTGAACGCTCT
CTACTAGAGTCACACTGGCTCACCTTCGGGTGGGCCTTTCTGCGTTTATACCTAGGGTACGGG
TTTTGCTGCCCTCTAGAGGTGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCGGAAGAGAG
TCAATTCAGGGTGGTGAATTTGACAGCTAGCTCAGTCCTAGGTATAATAGATCTGAATTCAT
TAAAGAGGAGAAAGGTACCATGGCGAGTAGCGAAGACGTTATCAAAGAGTTCATGCGTTTCAA
AGTTCGTATGGAAGGTTCCGTTAACGGTCACGAGTTCGAAATCGAAGGTGAAGGTGAAGGTCG
TCCGTACGAAGGTACCCAGACCGCTAAACTGAAAGTTACCAAAGGTGGTCCGCTGCCGTTCGC
TTGGGACATCCTGTCCCCGCAGTTCCAGTACGGTTCCAAAGCTTACGTTAAACACCCGGCTGA
CATCCCGGACTACCTGAAACTGTCCTTCCCGGAAGGTTTCAAATGGGAACGTGTTATGAACTT
CGAAGACGGTGGTGTTGTTACCGTTACCCAGGACTCCTCCCTGCAAGACGGTGAGTTCATCTA
CAAAGTTAAACTGCGTGGTACCAACTTCCCGTCCGACGGTCCGGTTATGCAGAAAAAAACCAT
GGGTTGGGAAGCTTCCACCGAACGTATGTACCCGGAAGACGGTGCTCTGAAAGGTGAAATCAA
AATGCGTCTGAAACTGAAAGACGGTGGTCACTACGACGCTGAAGTTAAAACCACCTACATGGC
TAAAAAACCGGTTCAGCTGCCGGGTGCTTACAAAACCGACATCAAACTGGACATCACCTCCCA
CAACGAGGACTACACCATCGTTGAACAGTACGAACGTGCTGAAGGTCGTCACTCCACCGGTGC
TTAACAGTGCAGTAGCATTATTCGCATTACCGACAAAGCGTTACGTGAAGAACTGGTGACGCT
GACCGGCGGGCAAAAACACGTAGCGCAAGCGGCGGAGTTCTGGGTGTTCTGTGCCGACTTTAA
CCGCCATTTACAGATCTGTCCGGATGCTCAGCTCGGCCTGGCGGAACAACTGTTGCTCGGTGT
CGTTGATACGGCAATGATGGCGCAGAATGCATTAATCGCAGCGGAATCGCTGGGATTGGGCGG
GGTATATATCGGCGGCCTGCGCAATAATATTGAAGCGGTGACGAAACTGCTTAAATTACCGCA
GCATGTTCTGCCGCTGTTTGGGCTGTGCCTTGGCTGGCCTGCGGATAATCCGGATCTTAAGCC
GCGTTTACCGGCCTCCATTTTGGTGCATGAAAACAGCTATCAACCGCTGGATAAAGGCGCACT
GGCGCAGTATGACGAGCAACTGGCGGAATATTACCTCACCCGTGGCAGCAATAATCGCCGGGA

TACCTGGAGCGATCATATCCGCCGAACAATCATTAAAGAAAGCCGCCCATTTATTCTGGATTA
TTTGCACAAACAGGGTTGGGCGACGCGCTAA-3′

Table A.1: Sequences of the genes inserted in the nfsA site of MG1655 bacteria. Blue denotes the nfsA gene fragment. Red denotes the mRFP fluorescent protein sequence. Orange denotes the ribosome binding sites. Green denotes the sfGFP gene. Black denotes the linker or spacer sequences joining these components together.

Name	Sequences
Positive control	5′-aactttcagtttagcggtctgttttagagctagaaatagcaagttaa aataaggctagtccgttatcaacttgaaaaagtggcaccgagtcggtgc tttttttt-3′
Negative control	5′-gttttagagctagaaatagcaagttaaaataaggctagtccgttatc aacttgaaaaagtggcaccgagtcggtgctttttttt-3′

Table A.2: Sequences of the positive and negative control guide RNAs used to measure the full dynamic range of the *in vivo* silencing assay in bacteria. Red denotes the targeting domain sequence which silences mRFP. Blue denotes the Cas9 handle motif. Brown denotes the terminator stem.

Name	Sequences
cgRNA	5′-GTTAGTAGGGTGTTGCGGGATGATGGGATaactttcagttta gcggtctATCCCATCATCCCGATCCCATCATCCCGTAGCAAGTTA AAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTC GGTGCTTTTTTT-3′
trigger RNA	5′-ATCCCATCATCCCGCAACACCCTACTAAC-3′

Table A.3: Sequences of kissing loop cgRNA and trigger RNA used in Figure 2.4. Red denotes the targeting domain sequence which silences mRFP. Blue denotes the Cas9 handle motif. Green denotes the stem and loop. Brown denotes the terminator stem. Plasmid files are denoted xRFP_xD1-T15 in the data repository. See scRNA_Cas9_20160526ZC for cloning information.

Name	Sequences
cgRNA	5′-CATCTAATTCAACAAGAATTGTTTTAGAGCTAATCCCATCAT CCCGTAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAA AGTGGCACCGAGTCGGTGCTTTTTTT-3′
trigger RNA	5′-CGGGATGATGGGAT-3′

Table A.4: Sequences of the anti-loop cgRNA and trigger RNA used in Figure 2.5. Red denotes the targeting domain sequence which silences sfGFP. Blue denotes the Cas9 handle motif. Green denotes the stem and loop. Brown denotes the terminator stem. Plasmid files are denoted xGFP_xt1-D1 in the data repository. See scRNA_Cas9_20160622ZC for cloning information.

Name	Sequences
cgRNA[I] (strong stem)	5′-aactttcagtttagcggtctgttttagagctagaaatagcaa gttaaaataaggctagtccgACACAAggggAAATTAACAACACAA CACACACAACACAGGccccggcaccgagtcggtgcttttttt-3′
cgRNA[I] (weak stem)	5′-aactttcagtttagcggtctgttttagagctagaaatagcaa gttaaaataaggctagtccgACACAAtataAAATTAACAACACAA CACACACAACACAGGtataggcaccgagtcggtgcttttttt-3′
cgRNA[I] (no stem)	5′-aactttcagtttagcggtctgttttagagctagaaatagcaa gttaaaataaggctagtccgACACAAttccAAATTAACAACACAA CACACACAACACAGGttccggcaccgagtcggtgcttttttt-3′

Table A.5: Sequences of strong stem, weak, and no stem cgRNA used in Figure 2.7. Red denotes the targeting domain sequence which silencing mRFP. Blue denotes the Cas9 handle motif. Green denotes the stem and loop. Brown denotes the terminator stem. See scRNA_Cas9_20181214ZC for cloning information.

Name	Sequence
cgRNA[A]	5'-aactttcagtttagcggtctgtttttagagctagaaatagcaagttaaaataaggctagtccgAATATAGGGGAAGAGAAAGAAGAAGAGAAGAGAAAGATGTCCCCggcaccgagtcggtgcttttttt-3'
cgRNA[B]	5'-aactttcagtttagcggtctgtttttagagctagaaatagcaagttaaaataaggctagtccgTATCATGGGGTTGTGTGTTGTTGTAAGTGTGTGTGTGTTGCCCCggcaccgagtcggtgcttttttt-3'
cgRNA[C]	5'-aactttcagtttagcggtctgtttttagagctagaaatagcaagttaaaataaggctagtccgAATAATGGTGATAAATACCTAATAAAGTGACGATGAATAGCACCggcaccgagtcggtgcttttttt-3'
cgRNA[D]	5'-aactttcagtttagcggtctgtttttagagctagaaatagcaagttaaaataaggctagtccgATCAAACGGGTAAACAAACAGGATAATTAAGGAGGCAGTACCCGggcaccgagtcggtgcttttttt-3'
cgRNA[G]	5'-aactttcagtttagcggtctgtttttagagctagaaatagcaagttaaaataaggctagtccgATCATTGCACATTCATCTTTCTTTCTTTCTTCTTCTTTCCgtgcggcaccgagtcggtgcttttttt-3'
cgRNA[I]	5'-aactttcagtttagcggtctgtttttagagctagaaatagcaagttaaaataaggctagtccgACACAAGGGGAAATTAACAACACAACACACACAACACAGGccccggcaccgagtcggtgcttttttt-3'
trigger[A]	5'-ACATCTTTCTCTTCTCTTCTTCTTTCTCTTCCCCTATATT-3'
trigger[B]	5'-CAACACACACACACTTACAACAACACACAACCCCATGATA-3'
trigger[C]	5'-CTATTCATCGTCACTTTATTAGGTATTTATCACCATTATT-3'
trigger[D]	5'-TACTGCCTCCTTAATTATCCTGTTTGTTTACCCGTTTGAT-3'
trigger[G]	5'-GGAAAGAAGAAGAAAGAAAGAAAGATGAATGTGCAATGAT-3'
trigger[I]	5'-CCTGTGTTGTGTGTGTTGTGTTGTTAATTTCCCCTTGTGT-3'

Table A.6: Sequences of orthogonal cgRNA and trigger RNA pairs used in orthogonality test (Figure 2.6), dose response test (Figure 2.10), and synthetic mRNA detection (see Figure 4.7). Red denotes the targeting domain sequence which silences mRFP. Blue denotes the Cas9 handle motif. Green denotes the terminator stem and loop. Brown denotes the terminator stem. See scRNA_Cas9_20180325ZC for cloning information. NUPACK analysis of these sequences are shown in Figure A.3.

mRNA-trigger sequences for the terminator switch system

mRNA-trig[D]

5'-TGGCTAAAGaaagaggagaaaAGGTTTATGGTAGCAGGTCATGCCTCTGGCAGCCCCGCATTCGGGACCGCCTCTCATTCGAATTGCGAACATGAAGAGATCCACCTCGCCGGCTCGATCCAGCCGCATGGCGCGCTTCTGGTCGTCAGCGAACATGATCATCGCGTCATCCAGGCCAGCGCCAACGCCGCGGAATTTCTGAATCTCGGAAGCGTACTCGGCGTTCCGCTCGCCGAGATCGACGGCGATCTGTTGATCAAGATCCTGCCGCATCTCGATCCCACCGCCGAAGGCATGCCGGTCGCGGTGCGCTGCCGGATCGGCAATCCCTCTACGGAGTACTGCGGTCTGATGCATCGGCCTCCGGAAGGCGGGCTGATCATCGAACTCGAACGTGCCGGCCCGTCGATCGATCTGTCAGGCACGCTGGCGCCGGCGCTGGAGCGGATCCGCACGGCGGGTTCACTGCGCGCGCTGTGCGATGACACCGTGCTGCTGTTT

CAGCAGTGCACCGGCTACGACCGGGTGATGGTGTATCGTTTCGATGAGCAAGGCCACGGCCTG
GTATTCTCCGAGTGCCATGTGCCTGGGCTCGAATCCTATTTCGGCAACCGCTATCCGTCGTCG
ACTGTCCCGCAGATGGCGCGGCAGCTGTACGTGCGGCAGCGCGTCCGCGTGCTGGTCGACGTC
ACCTATCAGCCGGTGCCGCTGGAGCCGCGGCTGTCGCCGCTGACCGGGCGCGATCTCGACATG
TCGGGCTGCTTCCTGCGCTCGATGTCGCCGTGCCATCTGCAGTTCCTGAAGGACATGGGCGTG
CGCGCCACCCTGGCGGTGTCGCTGGTGGTCGGCGGCAAGCTGTGGGGCCTGGTTGTCTGTCAC
CATTATCTGCCGCGCTTCATCCGTTTCGAGCTGCGGGCGATCTGCAAACGGCTCGCCGAAAGG
ATCGCGACGCGGATCACCGCGCTTGAGAGCgaattcggtggtggtggttctggtggtggtggt
tctATGAGTGTCAACTTAGCTTCCCAGTTGCGGGAAGGGACGAAAAAATCCCACTCCATGGCG
GAGAACGTCGGCTTTGTCAAATGCTTCCTCAAGGGCGTTGTCGAGAAAAATTCCTACCGTAAG
CTGGTTGGCAATCTCTACTTTGTCTACAGTGCCATGGAAGAGGAAATGGCAAAATTTAAGGAC
CATCCCATCCTCAGCCACATTTACTTCCCCGAACTCAACCGCAAACAAAGCCTAGAGCAAGAC
CTGCAATTCTATTACGGCTCCAACTGGCGGCAAGAAGTGAAAATTTCTGCCGCTGGCCAAGCC
TATGTGGACCGAGTCCGGCAAGTGGCCGCTACGGCCCCTGAATTGTTGGTGGCCCATTCCTAC
ACCCGTTACCTGGGGGATCTTTCCGGCGGTCAAATTCTCAAGAAAATTGCCCAAAATGCCATG
AATCTCCACGATGGTGGCACAGCTTTCTATGAATTTGCCGACATTGATGACGAAAAGGCTTTT
AAAAATACCTACCGTCAAGCTATGAATGATCTGCCCATTGACCAAGCCACCGCCGAACGGATT
GTGGATGAAGCCAATGACGCCTTTGCCATGAACATGAAAATGTTCAACGAACTTGAAGGCAAC
CTGATCAAGGCGATCGGCATTATGGTGTTCAACAGCCTCACCCGTCGCCGCAGTCAAGGCAGC
ACCGAAGTTGGCCTCGCCACCTCCGAAGGCTAGtaaacgtcgactctcgagtgagattgttga
cggtaccgtattttTACTGCCTCCTTAATTATCCTGTTTGTTTACCCGTTTGAT–3′

mRNA-trig[I]
5′–TGGCTAAAGaaagaggagaaaAGGTTTATGGTAGCAGGTCATGCCTCTGGCAGCCCCGCA
TTCGGGACCGCCTCTCATTCGAATTGCGAACATGAAGAGATCCACCTCGCCGGCTCGATCCAG
CCGCATGGCGCGCTTCTGGTCGTCAGCGAACATGATCATCGCGTCATCCAGGCCAGCGCCAAC
GCCGCGGAATTTCTGAATCTCGGAAGCGTACTCGGCGTTCCGCTCGCCGAGATCGACGGCGAT
CTGTTGATCAAGATCCTGCCGCATCTCGATCCCACCGCCGAAGGCATGCCGGTCGCGGTGCGC
TGCCGGATCGGCAATCCCTCTACGGAGTACTGCGGTCTGATGCATCGGCCTCCGGAAGGCGGG
CTGATCATCGAACTCGAACGTGCCGGCCCGTCGATCGATCTGTCAGGCACGCTGGCGCCGGCG
CTGGAGCGGATCCGCACGGCGGGTTCACTGCGCGCGCTGTGCGATGACACCGTGCTGCTGTTT
CAGCAGTGCACCGGCTACGACCGGGTGATGGTGTATCGTTTCGATGAGCAAGGCCACGGCCTG
GTATTCTCCGAGTGCCATGTGCCTGGGCTCGAATCCTATTTCGGCAACCGCTATCCGTCGTCG
ACTGTCCCGCAGATGGCGCGGCAGCTGTACGTGCGGCAGCGCGTCCGCGTGCTGGTCGACGTC
ACCTATCAGCCGGTGCCGCTGGAGCCGCGGCTGTCGCCGCTGACCGGGCGCGATCTCGACATG
TCGGGCTGCTTCCTGCGCTCGATGTCGCCGTGCCATCTGCAGTTCCTGAAGGACATGGGCGTG

CGCGCCACCCTGGCGGTGTCGCTGGTGGTCGGCGGCAAGCTGTGGGGCCTGGTTGTCTGTCAC
CATTATCTGCCGCGCTTCATCCGTTTCGAGCTGCGGGCGATCTGCAAACGGCTCGCCGAAAGG
ATCGCGACGCGGATCACCGCGCTTGAGAGCgaattcggtggtggtggttctggtggtggtggt
tctATGAGTGTCAACTTAGCTTCCCAGTTGCGGGAAGGGACGAAAAAATCCCACTCCATGGCG
GAGAACGTCGGCTTTGTCAAATGCTTCCTCAAGGGCGTTGTCGAGAAAAATTCCTACCGTAAG
CTGGTTGGCAATCTCTACTTTGTCTACAGTGCCATGGAAGAGGAAATGGCAAAATTTAAGGAC
CATCCCATCCTCAGCCACATTTACTTCCCCGAACTCAACCGCAAACAAAGCCTAGAGCAAGAC
CTGCAATTCTATTACGGCTCCAACTGGCGGCAAGAAGTGAAAATTTCTGCCGCTGGCCAAGCC
TATGTGGACCGAGTCCGGCAAGTGGCCGCTACGGCCCCTGAATTGTTGGTGGCCCATTCCTAC
ACCCGTTACCTGGGGGATCTTTCCGGCGGTCAAATTCTCAAGAAAATTGCCCAAAATGCCATG
AATCTCCACGATGGTGGCACAGCTTTCTATGAATTTGCCGACATTGATGACGAAAAGGCTTTT
AAAAATACCTACCGTCAAGCTATGAATGATCTGCCCATTGACCAAGCCACCGCCGAACGGATT
GTGGATGAAGCCAATGACGCCTTTGCCATGAACATGAAAATGTTCAACGAACTTGAAGGCAAC
CTGATCAAGGCGATCGGCATTATGGTGTTCAACAGCCTCACCCGTCGCCGCAGTCAAGGCAGC
ACCGAAGTTGGCCTCGCCACCTCCGAAGGCTAGtaaacgtcgactctcgagtgagattgttga
cggtaccgtatttttCCTGTGTTGTGTGTGTTGTGTTGTTAATTTCCCCTTGTGT-3′

Table A.7: Sequences of orthogonal mRNA triggers used in the synthetic mRNA detection experiment for the terminator switch system (Figure 4.7). Blue denotes the ribosome binding site. Red denotes the miRFP670 fluorescent protein sequence. Orange denotes heme oxygenase protein. Green denotes unstructured trigger RNA sequence. Black denotes the linker or spacer sequences joining these components together.

Name	RNA Sequence
mtrig0	5′-ATGAAACCTAGAACTATGAAACAATGCTCAAATGCCAGATACAAGAA TACAGACGTTACCAAGATACAAGATACCAAGATACGCTTAAATGCACTA TAG-3′
mtrig1	5′-ATGTCAAAGTTCTCAGTATCATCGTTCAAGTCAAGTAGTCAATCGTC AATCAGTAATCAGTCTCAGTCCCGTAAATCAGTTTCAAAGTATCCAAGT TAG-3′
mtrig2	5′-ATGCAAAGATACAACGGTCACCCGCTAGACCTCCAGAACTCCGATAG AACAATGACATACGATACGACAATAGCAAATAGACTCAACGACTCAGAT TAA-3′

Table A.8: Table of universal trigger RNA sequences. NUPACK analysis of these sequences are shown in Figures 3.20 and 3.21. mRNA versions of these sequences are shown in Table A.9.

Name	RNA Sequence	Amino Acid Sequence
mtrig0	5′-TGGCTAAAGtcacacaggaaagAGGTTTATGAAACCTAGAACTATGAAACAATGCTCAAATGCCAGATACAAGAATACAGACGTTACCAAGATACAAGATACCAAGATACGCTTAAATGCACTATAG-3′	MKPRTMKQCSNARYKNTDVTKIQDTKIRLNAL*
mtrig1	5′-TGGCTAAAGtcacacaggaaagAGGTTTATGTCAAAGTTCTCAGTATCATCGTTCAAGTCAAGTAGTCAATCGTCAATCAGTAATCAGTCTCAGTCCCGTAAATCAGTTTCAAAGTATCCAAGTTAG-3′	MSKFSVSSFKSSSQSSISNQSQSRKSVSKYPS*
mtrig2	5′-TGGCTAAAGtcacacaggaaagAGGTTTATGCAAAGATACAACGGTCACCCGCTAGACCTCCAGAACTCCGATAGAACAATGACATACGATACGACAATAGCAAATAGACTCAACGACTCAGATTAA-3′	MQRYNGHPLDLQNSDRTMTYDTTIANRLNDSD*

Table A.9: Table of universal trigger mRNA sequences. Blue denotes the ribosome binding site. Green denote the universal trigger sequences. Non-coding RNA versions of these sequences are shown in Table A.8.

Name	Sequences
ts10n4	5′-aactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgagagaatgctggagaaattttttctccagcatcaaagtacattagtgtcattaatgtacgagagaa-3′
ts10n5	5′-aactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgtcccagttcagtgtgtaattacacactgaacgcttatcgcgagatttcccgtagggagatttcatgtccc-3′
ts10n6	5′-aactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgaaaccaccccgatcccctaagggatcggggatgcacctgtaggttacaggtgcac-3′

Table A.10: Sequences of synthetic guide RNAs with non-canonical terminator stems. Red denotes the targeting domain. Blue denotes the Cas9 handle. Green denotes the variable region containing linker A to linker C. Fluorescence measurements for constructs up above are shown in Figure 3.2. Domain dimensions for the gRNAs shown above are listed in Table 3.1.

Name	Sequence
ts45m0-1	5′-gttaccaagatacaagatacaactttcagtttagcggtctgt tttagagctagaaatagcaagttaaaataaggctagtccgaataa tggtcgtatacgcagtatgcgaccgatagtgacctctggtcactg tccaggtgtcttgtgtcttggtgacgtttgtattt-3′
ts45m0-3	5′-accaagatacaactttcagtttagcggtctgttttagagcta gaaatagcaagttaaaataaggctagtccgaatccacatcattga cggggtcaatggtggagtccatcctctggatggactcatggtgtc ttggtgtcttgtatt-3′
ts45m1-1	5′-ctcagtatcatcgttcaagtaactttcagtttagcggtctgt tttagagctagaaatagcaagttaaaataaggctagtccgatccc agcttatcgccgtaggtgatgagcctaacccgccctaggtgggtt ggcacgcttgagcgatgatgctgggaactttgata-3′
ts45m2-3	5′-gacatacgataactttcagtttagcggtctgttttagagcta gaaatagcaagttaaaataaggctagtccgtcaggagcctgccta cacggtaggcaggcctatacccgcagagtgggtgtagccagtcgt atgtcattgttctat-3′

Table A.11: Sequences of a set of orthogonal reverse toehold switch systems used in Figure 3.4 and 4.8. Red denotes the targeting domain sequence which silencing mRFP. Blue denotes the Cas9 handle motif. Green denotes variable domain containing the terminator stem, inhibitory duplex, and toehold.

Name	Sequence
2Comp-g1-1A	5′-aactttcagtttagcggtctgttttagagctaGAAAta gcaagttaaaataaggctagtccgCAATCTCAAGAGGAGG TAAGGGAATC-3′
2Comp-g2-1A	5′-ccttacctcctcttgggcaccgagtcggtgcttttttt t-3′
trigger RNA	5′-GATTCCCTTACCTCCTCTTGAGATTG-3′

Table A.12: Sequences of a set of reverse split terminator switch cgRNAs used in Figure 3.7 and 3.9. Red denotes the targeting domain sequence which silences mRFP. Blue denotes the Cas9 handle motif. Green denotes variable domain containing the terminator stem. *g*1 and *g*2 denote two halves of a full guide RNA complex.

Name	Sequence
ts12uni-g1	5′-gttagttttggtatcagtttcagtttcggtaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgtcccgtggcaccacacactgaaactgaaactgatac-3′
ts12uni-g2	5′-actgaaactgaaactgatactgaaactgatgtatcagtttcagtttcggtgtgtggtgccggcaccgagtcggtgcttttttt-3′
ts25uni-g1	5′-aactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgtcggtatcagtttcggtatcggtttcagtttcagttcttt-3′
ts25uni-g2	5′-gaaactgatactgaaactgaccccgccaagggcggggacgcaaaaaccccgcttcggcggggttttttcgc-3′
ts25uni-g1	5′-aactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgtcggtatcagtttcggtatcggtttcagtttcagttcttt-3′
unitrig	5′-GCCTGAAACTGATACTGAAACTGAACTGAAACTGAAACTGATACTGAAACTGATACTGAACTGAAACTGAAACTGAAACTGAAACTGACTGAAACTGACC-3′

Table A.13: Sequences of cgRNAs targeted against an unstructured trigger RNA called unitrig. Red denotes the targeting domain sequence which silencing mRFP. Blue denotes the Cas9 handle motif. Green denotes variable domain containing the terminator stem, inhibitory duplex, and toehold. unitrig is unstructured trigger RNA for the ts12uni and ts25uni designs. g1 and g2 denote two halves of a full guide RNA complex.

Name	Sequence
0ts18b-g1	5′-gtgtgtcttggtgttttgtgtcttggtgacaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgtcccaatccccccgttaccaagatacaagatac-3′
0ts18b-g2	5′-gggggggaggcggggtccctaggaccccgccgaa-3′
28ts18b-g1	5′-ggtcgttgagtctgtttgttgttgttgtgtaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccggtcccacccccccatacgacaatagcaaataga-3′
28ts18b-g2	5′-gggggggccgtcggggctttgccccgatggaaa-3′
29ts18b-g1	5′-tcggagttctggaggtctagcgggtaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgtatccacccccccacccgctagacctccagaac-3′
29ts18b-g2	5′-ggggggggtgggcagcgcacagcgctgcctgtaa-3′

Table A.14: Sequences of a set of orthogonal single hairpin split guide RNAs used in Figure 3.13. Red denotes the targeting domain sequence which silencing mRFP. Blue denotes the Cas9 handle motif. Green denotes variable domain containing the terminator stem, inhibitory duplex, and toehold. g1 and g2 denote two halves of a full guide RNA complex.

Name	Sequence
3ts23b-m0-g1	5′-atgtttgtattcttgtatctggtatttgagaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgacgccatccaacccccctcaaatgccagatacaaga-3′
3ts23b-m0-g2	5′-ctcaaatgccagatacaagaatatcttgtatctggtatttgaggggggttggagcgttacgaggtaacgca-3′
19ts23b-m1-g1	5′-gtattttgaagctggtttacgggacaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccggtccatcccccgtcccgtaaatcagt-3′
19ts23b-m1-g2	5′-gtcccgtaaatcagtgaggctggtttacgggacgggggccagcgccatgcgctgga-3′
30ts23b-m2-g1	5′-gtctgagttgttgagtctatttgttgttgtaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccgtcatcaaccccacaatagcaaatagactcaa-3′
30ts23b-m2-g2	5′-acaatagcaaatagactcaaaggttgagtctatttgttgttgtggggtggacgcccaaggcgtcct-3′
31ts23b-m2-g1	5′-tgttgttgtatcgtatgtcattgttctgtcaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccggcacccccctccacccgatagaacaatgacatacga-3′
31ts23b-m2-g2	5′-gatagaacaatgacatacgaaaatcgtatgtcattgttctgtcgggtggagggcaccggctttgccggtgg-3′
33ts23b-m2-g1	5′-ggtcgttgagtctatttgctattgtaactttcagtttagcggtctgttttagagctagaaatagcaagttaaaataaggctagtccggtcccaccccacaatagcaaataga-3′
33ts23b-m2-g2	5′-acaatagcaaatagagagtctatttgctattgtgggggccgccccaccggggcgga-3′

Table A.15: Table of sequences for the two hairpin guide RNA tested in Figure 3.15. Trigger RNA sequences are listed in Table A.8.

Name	stem a	duplex b	toehold c	trigger RNA
3ts23b	10	20	10	mtrig0
19ts23b	5	15	10	mtrig1
30ts23b	5	20	10	mtrig2
31ts23b	10	20	10	mtrig2
33ts23b	5	15	10	mtrig2
0ts12uni	10	20	10	unitrig

Table A.16: Table of dimensions for the two hairpin guide RNA sequences tested in 3.15. cgRNA sequences are listed in Table A.15. Trigger RNA sequences are listed in Table A.8.

A.6 MFE structures

MFE structure at 37.0 C

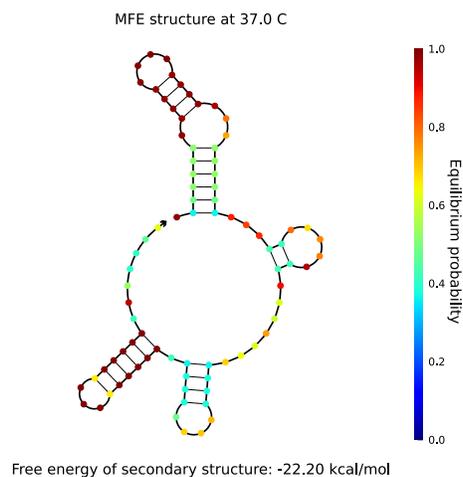Free energy of secondary structure: -22.20 kcal/mol

Figure A.2: NUPACK analysis of a canonical guide RNA without the targeting domain (see Negative Control sequence in Table A.2). The minimum free energy (MFE) structure represents the most probable RNA structure for the given sequence in the thermodynamic ensemble at equilibrium. Red means base pairs are highly likely to form. Blue means the base pairs are unlikely to form. The targeting domain was removed for the MFE analysis because it interfered with formation of the Cas9 handle structure (hairpin formed from the first 30nt). The base of the Cas9 handle is composed of G-U wobble pairs, which NUPACK predicts will form at around 0.5 probability (shaded green). These are not very stable interactions by themselves. Binding of this motif by Cas9 helps stabilize this structure.
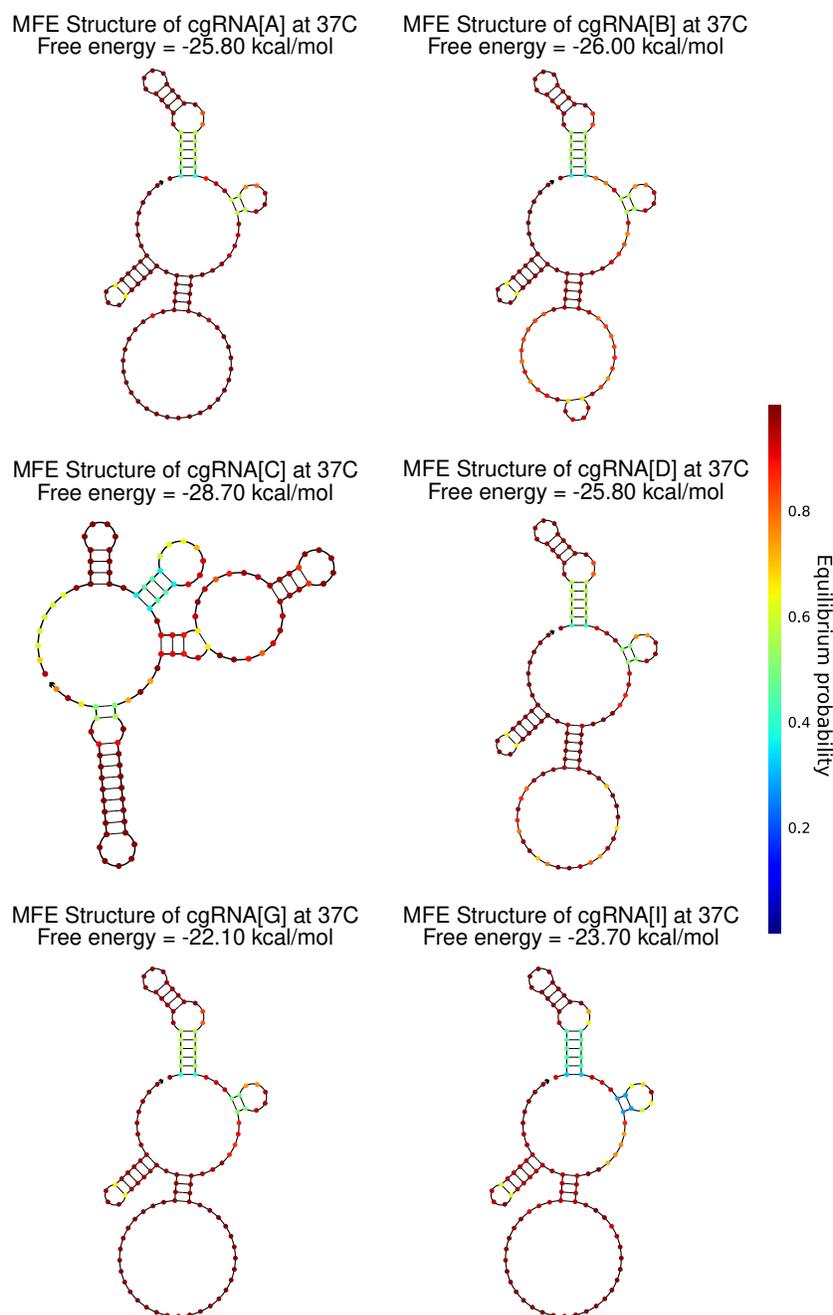
Figure A.3: NUPACK analysis of orthogonal terminator switch cgRNA sequences from Table A.6. These sequences have the targeting domain removed to enable visualization of the Cas9 handle and stem loop structures. The minimum free energy (MFE) structure represents the most probable RNA structure for the given sequence in the thermodynamic ensemble at equilibrium. Red means base pairs are highly likely to form. Blue means the base pairs are unlikely to form. All cgRNA sequences were active except cgRNA C. For cgRNA C, the MFE structure deviates significantly from the canonical guide RNA structure (see Figure A.2), resulting in loss of activity (see Figure 2.10).
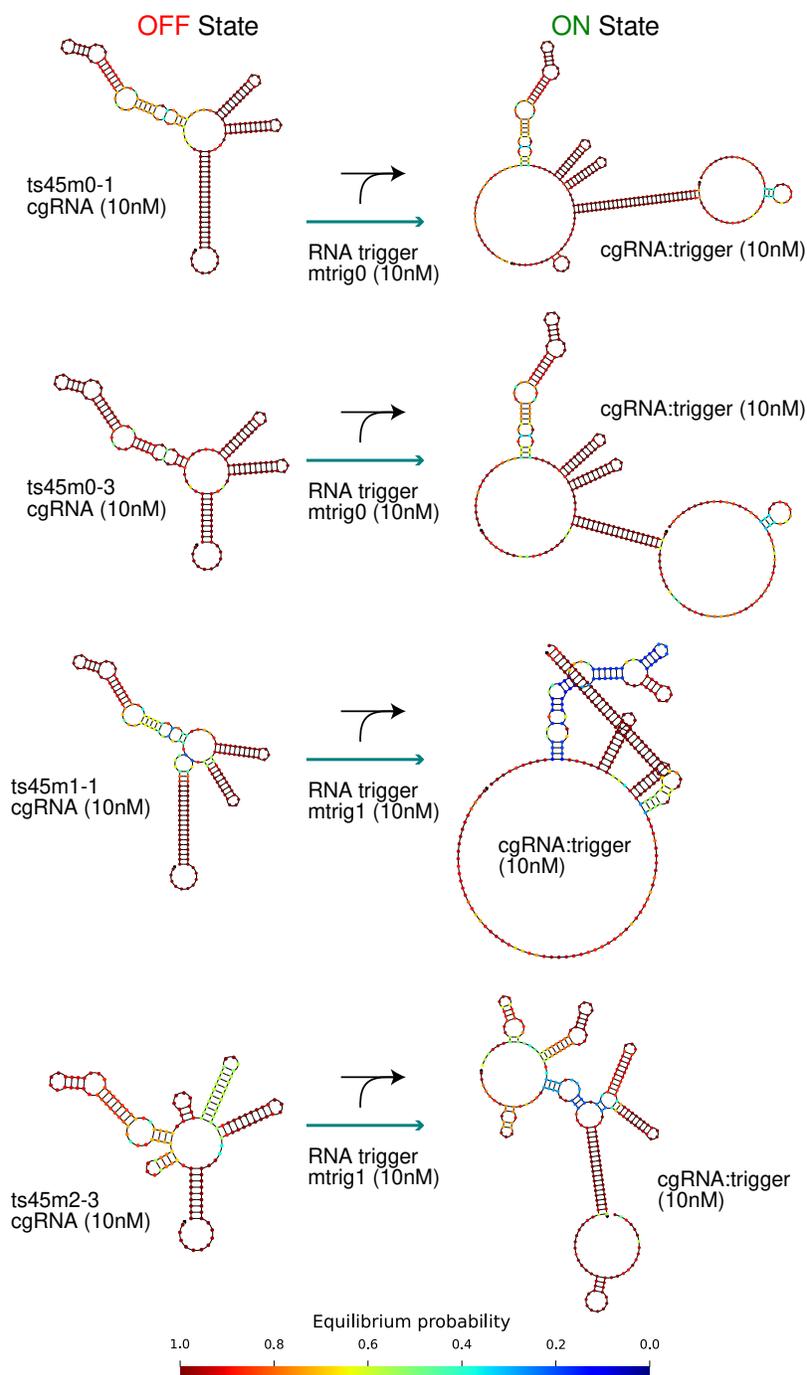
Figure A.4: NUPACK test tube analysis of reverse toehold switch cgRNA sequences from Table A.11. Experimental result shown in Figure 3.5. The minimum free energy (MFE) structure represents the most probable RNA structure for the given sequence in the thermo-dynamic ensemble at equilibrium. Red means base pairs are highly likely to form. Blue means the base pairs are unlikely to form. Concentrations of reactants and products for each step of the reaction are denoted next the structure label. RNA was simulated at 37°C in 1M Na⁺.
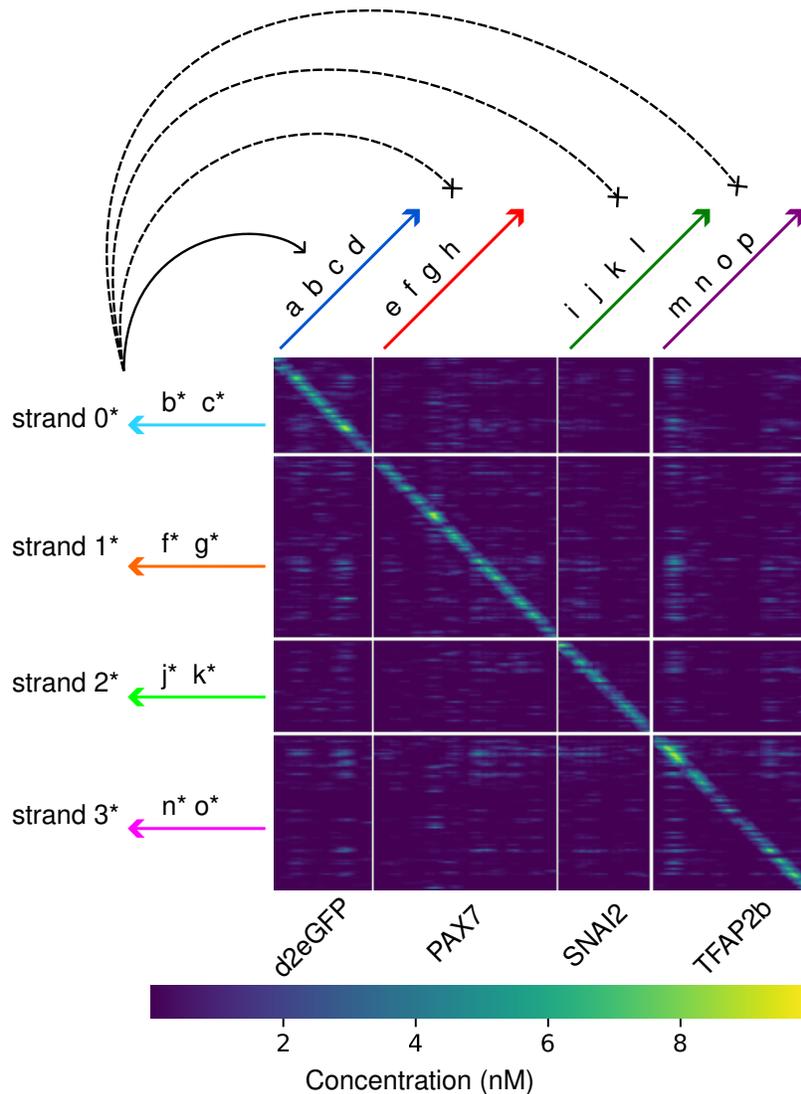
A.7   Supplemental data

Figure A.5: Prediction of sub-sequence specificity of a 10nt anti-sense RNA strand using mRNA scanner. The above heat map shows the NUPACK predicted crosstalk between pairs of mRNA sub-sequences (x axis) and their sub-sequence reverse-complements (y axis) in a crosstalk tube analysis at 37C. 10nM was used as the reactant concentration. Yellow denotes regions of high interaction, and blue denotes regions of low interaction. The yellow diagonal represents on-target interactions. Off diagonal values represent off-target interactions. A 10nt long strand of RNA was used as the anti-sense strand, which binds with a moving window of mRNA sub-sequences (100nt long). Lengthening this anti-sense strand increases crosstalk (see Figure 4.4). Sequestering the anti-sense strand in an RNA hairpin improves the specificity of sub-sequence detection (see Figure A.6).
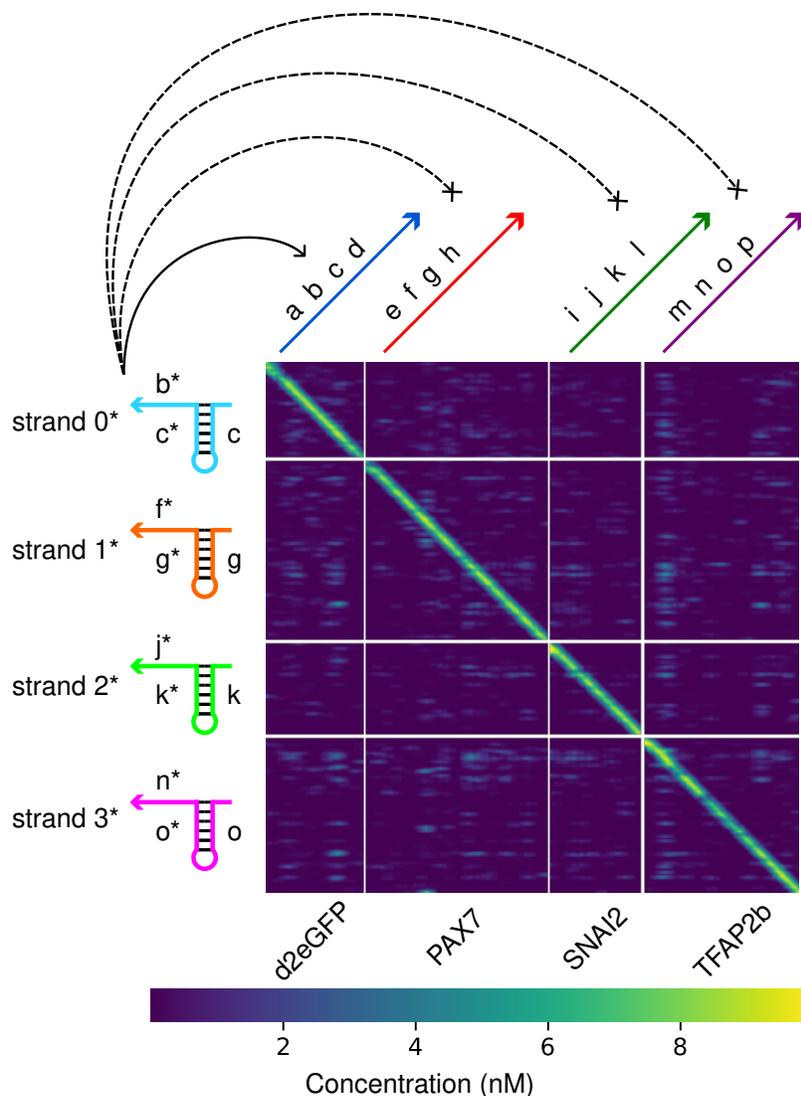
Figure A.6: Prediction of sub-sequence specificity of an RNA hairpin using mRNA scanner. The above heat map shows the NUPACK predicted crosstalk between pairs of mRNA sub-sequences (x axis) and their sub-sequence reverse-complements (y axis) in a crosstalk tube analysis at 37C. 10nM was used as the reactant concentration. Yellow denotes regions of high interaction, and blue denotes regions of low interaction. The yellow diagonal represents on-target interactions. Off diagonal values represent off-target interactions. An RNA hairpin with a toehold of 10nt and duplex of 20nt was used as the anti-sense strand. This hairpin was used to sense a moving window of mRNA sub-sequences (100nt long). Shortening the toehold reduces off-target interactions (see Figure A.7).
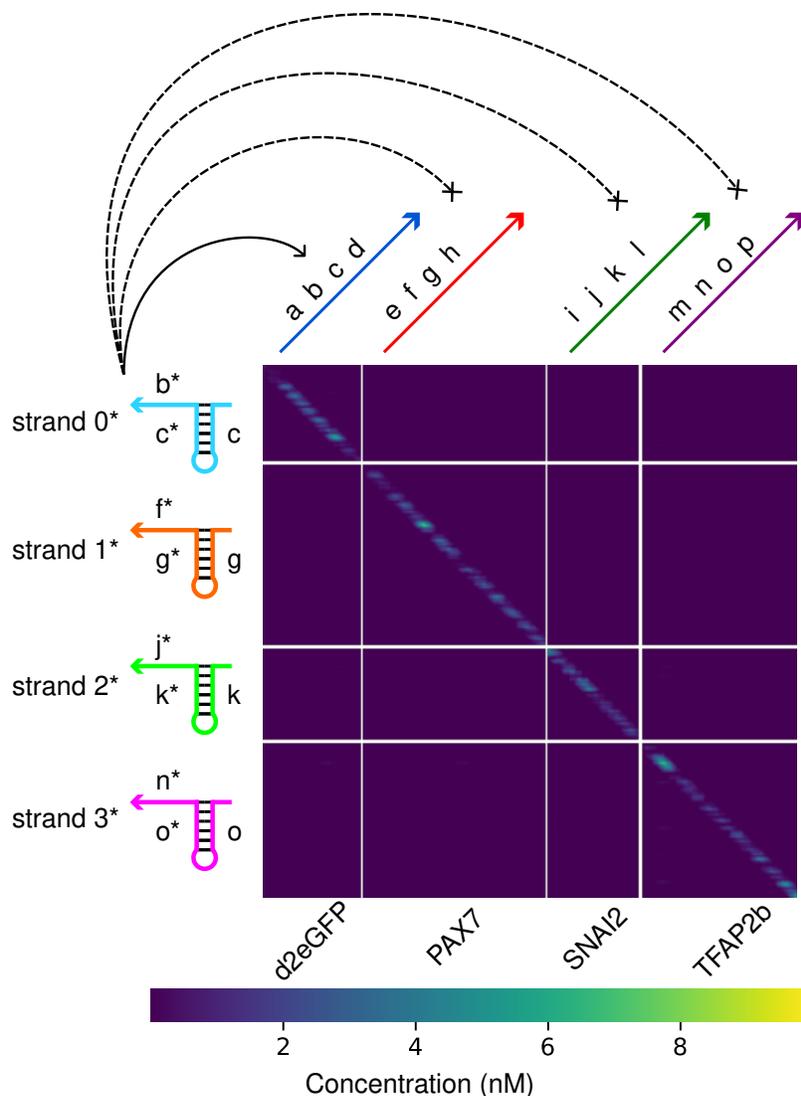
Figure A.7: Prediction of sub-sequence specificity an RNA hairpin with a 5nt toehold using mRNA scanner. The above heat map shows the NUPACK predicted crosstalk between pairs of mRNA sub-sequences (x axis) and their sub-sequence reverse-complements (y axis) in a crosstalk tube analysis at 37C. 10nM was used as the reactant concentration. Yellow denotes regions of high interaction, and blue denotes low interaction. The yellow diagonal represents on-target interactions. Off diagonal values represent off-target interactions. An RNA hairpin with a toehold of 5nt and duplex of 20nt was used as the anti-sense strand. This hairpin was used to sense a moving window of mRNA sub-sequences (100nt long). Compared to anti-sense strand of 20nt (see Figure 4.4), this hairpin has almost no off-target interactions. The row sum of this heat map was used to compute the bottom row plot in Figure 4.3
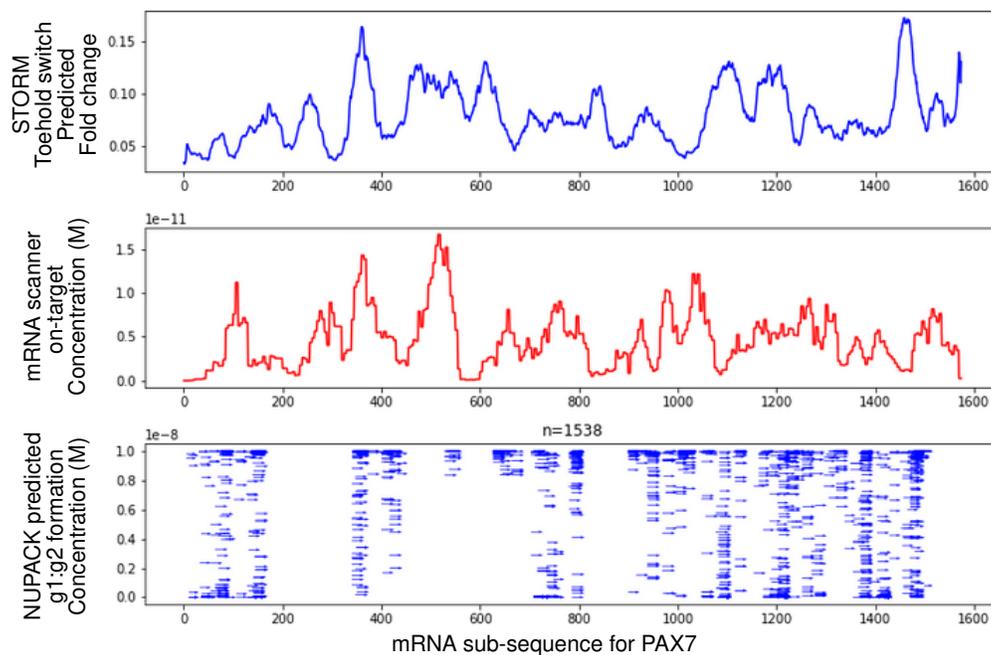
Figure A.8: Scatter plot of trigger mRNA sub-sequences for split terminator switch cgRNA overlayed against predictions from STORM and mRNA scanner for Pax7.
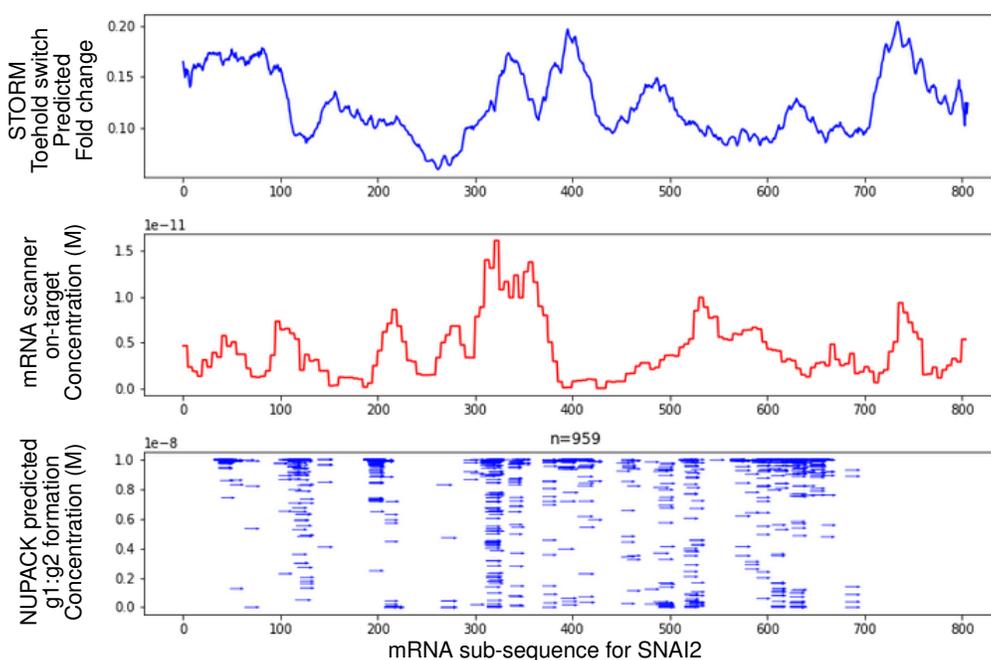

Figure A.9: Scatter plot of trigger mRNA sub-sequences for split terminator switch cgRNA overlayed against predictions from STORM and mRNA scanner for SNAI2
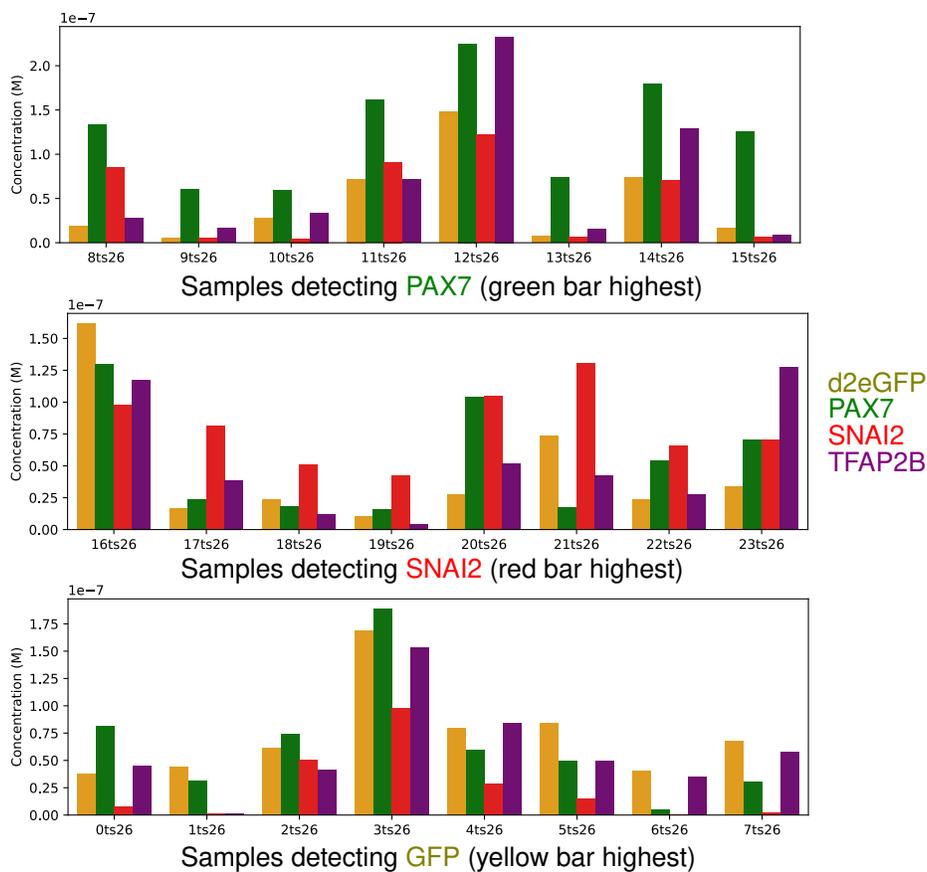
Figure A.10: Predicted mRNA orthogonality of the split terminator switch. cgRNAs and mRNAs were provided as input to the mRNA scanner tool. Test tube analysis was run at 37C with each reactant at 10nM. The concentration of cgRNA:trigger complexes for cognate and non-cognate sequences were summed up and plotted as the bar plot above.
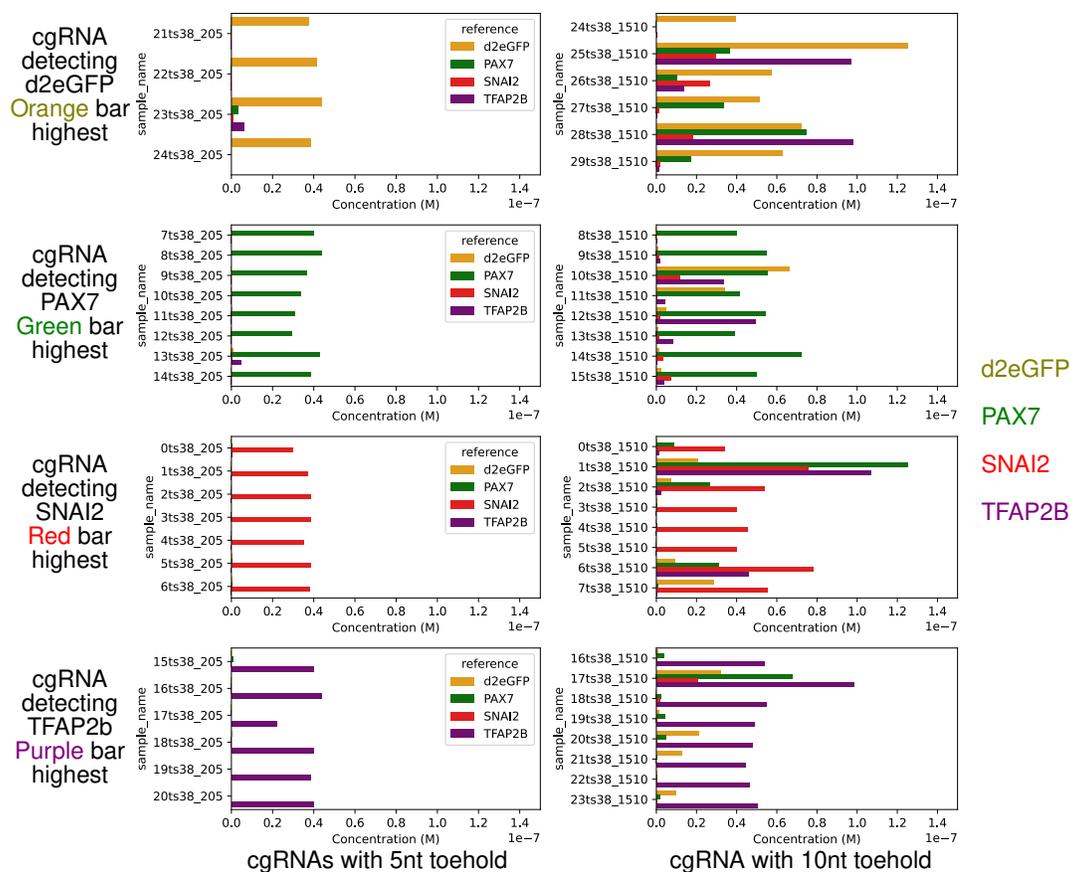
Figure A.11: Predicted mRNA orthogonality of the inhibited split terminator switch. Left plots show cgRNAs with 5nt toeholds. These have almost no crosstalk. Right plots show cgRNAs with 10nt toeholds, which have increased predicted crosstalk. cgRNAs and mRNAs were provided as input to the mRNA scanner tool. Test tube analysis was run at 37C with each reactant at 10nM. The concentration of cgRNA:trigger complexes for cognate and non-cognate sequences were summed up and plotted as the bar plot above.
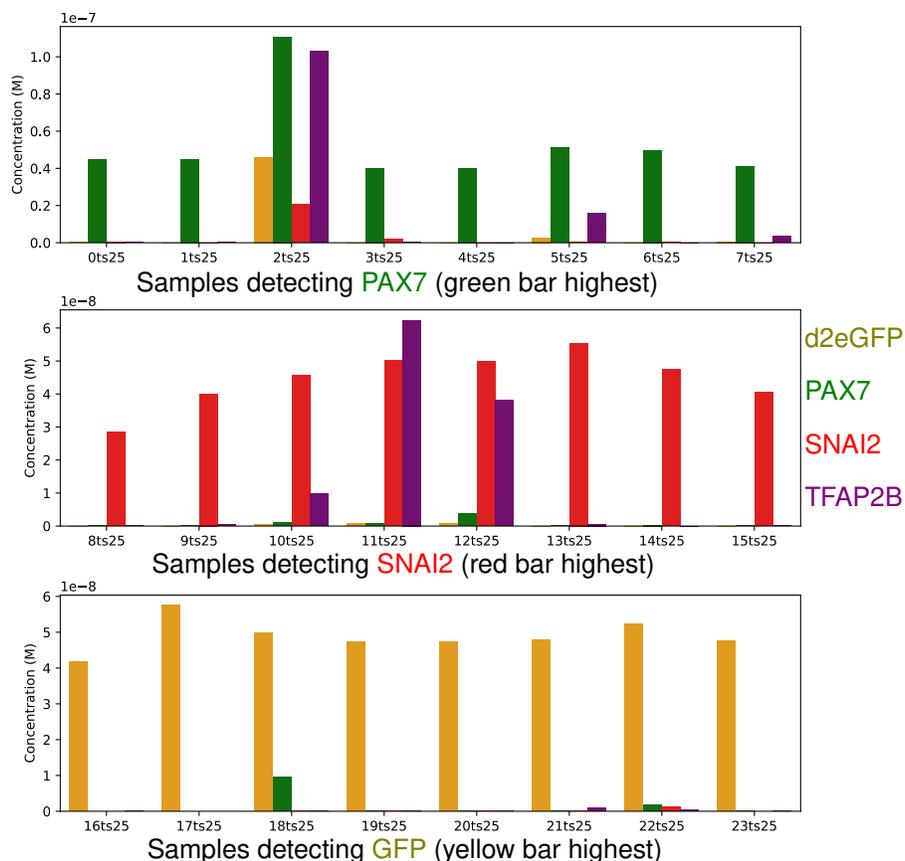
Figure A.12: Predicted mRNA orthogonality of the reverse split terminator switch. cgRNAs and mRNAs were provided as input to the mRNA scanner tool. Test tube analysis was run at 37C with each reactant at 10nM. The concentration of cgRNA:trigger complexes for cognate and non-cognate sequences were summed up and plotted as the bar plot above.

A.8   Genbank files of plasmid constructs and source code

The section details the organization of the data repository containing genbank files for plasmid constructs and source code for data analysis. This data repository is accessible from the Caltech Library at doi.org/10.22002/hktey-pzp98.

Table A.17: Table of genbank, data, and source file descriptions.

Begin of Table		
folder	**name**	**description**
./genbank/assay/	dCas9_LacI.gb	CamR plasmid containing dCas9 and LacI used in Chapter 2

Continuation of Table A.17

folder	name	description
./genbank/assay/	nfsA_seq.gb	Fluorescent reporter sequences inserted in the nfsA gene of MG1655 to generate the reporter cell line (see Table A.1)
./genbank/assay/	cgI-pLI.gb	cgRNA and trigger RNA plasmid map for low throughput setup used in Chapter 2
./genbank/assay/	xRFP.gb	Plasmid map for positive control gRNA used in the medium throughput setup in Chapter 3
./genbank/assay/	notar.gb	Plasmid map for negative control gRNA used in the medium throughput setup in Chapter 3
./genbank/assay/	cgRNA.gb	Plasmid map of a cgRNA used in the medium throughput setup in Chapter 3
./genbank/assay/	dCas9_RFP.gb	CamR plasmid containing dCas9 and mRFP used in the medium throughput setup in Chapter 3
./genbank/assay/	trigL5.gb	Carb plasmid expression miRFP670-HO1 fluorescent reporter gene used in the medium throughput setup in Chapter 3
./genbank/assay/	trigS2.gb	Carb plasmid used as the no trigger condition in the medium throughput setup in Chapter 3
./genbank/assay/	cgRNA_trig.gb	Carb plasmid used as the cloning vector for the high-throughput screening setup in Chapter 5

Continuation of Table A.17

folder	name	description
./genbank/mtrig/	all files	Carb plasmid containing mtrig0, mtrig1, mtrig2 used in the medium throughput setup in Chapter 3
./genbank/kissing_loop/	all files	Carb plasmid containing kissing loop and anti-loop cgRNAs from Chapter 2
./genbank/triggerRNA/	all files	Carb plasmid containing trigger mRNAs used in Chapter 4
./genbank/unitrig/	all files	Spectinomycin plasmid containing ts25 (reverse split terminator switch) and ts12 (catalytic two hairpin switch) cgRNAs used in Chapter 3
./genbank/2Comp/	all files	Spectinomycin plasmid containing split terminator switch cgRNA used in Chapter 3
./genbank/ts18_mtrig/	all files	Spectinomycin plasmid containing single hairpin split cgRNA detecting mtrigs used in Chapter 3
./genbank/ts23_mtrig/	all files	Spectinomycin plasmid containing two hairpin split cgRNA detecting mtrigs used in Chapter 3
./genbank/ts45_mtrig/	all files	Spectinomycin plasmid containing reverse toehold switch cgRNA detecting mtrigs used in Chapter 3
./genbank/ts23_mRNA/	all files	Spectinomycin plasmid containing two hairpin split guide cgRNA detecting mRNA used in Chapter 4

Continuation of Table A.17

folder	name	description
./genbank/ts25_mir/	all files	Spectinomycin plasmid containing reverse split terminator switch cgRNA detecting microRNA used in Chapter 4
./genbank/ts25_mRNA/	all files	Spectinomycin plasmid containing reverse split terminator switch cgRNA detecting mRNA used in Chapter 4
./genbank/ts26_mRNA/	all files	Spectinomycin plasmid containing split terminator switch cgRNA detecting mRNA used in Chapter 4
./genbank/ts38_dim15_10/	all files	Spectinomycin plasmid containing inhibited split cgRNA detecting mRNA used in Chapter 3
./genbank/ts38_dim20_5/	all files	Spectinomycin plasmid containing inhibited split cgRNA detecting mRNA used in Chapter 3
./genbank/ts38_short_trig/	all files	Carb plasmid containing trigger RNA of inhibited split cgRNA used in Chapter 3
./src/	all files	Scripts and tools for mRNA scanner, test tube analysis, and cgRNA design