The Pennsylvania State University

The Graduate School

DEVELOPING NUCLEIC ACID-BASED BIOSENSOR PLATFORMS

A Dissertation in

Agricultural and Biological Engineering

by

Grace E. Vezeau

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2021

The dissertation of Grace E. Vezeau was reviewed and approved by the following:

Howard M. Salis
Associate Professor of Chemical Engineering & Agricultural and Biological Engineering
Dissertation Advisor
Chair of Committee

Jeffrey Catchmark
Professor of Agricultural and Biological Engineering
Director, Graduate Studies
Affiliated Faculty of the The Rock Ethics Institute

Philip Bevilacqua
Distinguished Professor of Chemistry & Biochemistry and Molecular Biology

Andrew Zydney
Distinguished Professor of Chemical Engineering

Paul Heinemann
Professor of Agricultural and Biological Engineering
Head, Department of Agricultural and Biological Engineering

ABSTRACT

Biological systems are excellent sensing platforms, able to specifically detect a diverse array of chemical and biomolecular stimuli in their environment and respond rapidly and appropriately. Cells accomplish this by using genetically-encoded sensing systems, where a gene product such as protein or nucleic acid binds to a specific compound and subsequently directs a change in the expression of a relevant gene. While these genetically-encoded sensing systems let cells respond to their environments and maintain homeostasis, these responses are rarely human-usable. The process of re-engineering biological sensors is further complicated by the need for recognition elements for novel inputs and the fact that biological recognition elements are rarely modular. While re-engineering protein-based, genetically-encoded sensing elements is simple in a few cases, it is largely a complex and time-consuming process. Developing new nucleic acid-based recognition elements and mechanisms of signal transduction, however, is enabled by their relatively smaller sequence space and more predictable sequence-structure relationship.

Here, I develop new classes of nucleic-acid based sensors via computational sequence design, expanding the capabilities of nucleic acid-based sensing and developing sensors with the potential for real-world utility. I first focus on developing protein-detecting RNA switches that operate in cell-free expression systems. To enable further development of the cell-free platform, I examine how cell-free reaction formulation affects the processes of gene expression, and overall output characteristics. I next examine sequence-function relationships of mutated single-guide RNAs (sgRNAs) in directing Cas9 activity. Finally, I use the learned design rules to develop Cas9-based sensors that directly detect viral RNA sequences.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Making the jump to biological engineering from my undergraduate background in biogeochemistry has been a simultaneously challenging and rewarding experience. I am first grateful to Dr. Howard Salis for providing me with the opportunity to explore many diverse aspects of the field of synthetic biology. His support and guidance have been instrumental in my growth as a scientist. I would next like to thank my doctoral committee, Drs. Jeffrey Catchmark, Philip Bevilacqua, Andrew Zydney, and Manish Kumar. Their advice over the years has been invaluable in the direction of my Ph.D. work. I would next like to thank my labmates, who have been an essential part of my Ph.D. journey. Thanks to Drs. Amin Espah Borujeni, Iman Farasat, Tian Tian, and Chiam Yu Ng for teaching me the nuts and bolts of genetic engineering and life in the lab. Thanks also to Drs. Sean Halper and Alex Reis, as well as Dan Cetnar, Ayaan Hossain, David Sampson, Travis LaFleur, Lipika Gadila, Philip Clauer, Becca O'Toole, and Erin Essington. I have learned so much from working with all of you, and with any luck, you might even be able to say the same of me. Finally, I would like to thank my friends and family, most especially my husband, Dr. Allan Brooks, my parents, Debra and Patrick Vezeau, and my siblings, Neil and Anne Vezeau. Thank you all for your love and support over the years, for your guidance through the challenges of obtaining a doctoral degree, your encouragement to keep going (or take a break!), and help making sure I always, or at least mostly, stay on target.

The work in this dissertation was supported by funds from the Air Force Office of Scientific Research (FA9550-14-1-0089), an NSF Career Award to H.M.S. (CBET-1253641), the Defense Advanced Research Projects Agency (FA8750-17-C-0254), and the Department of Energy (DE-SC0019090). Findings and conclusions presented here do not necessarily reflect the views of these funding agencies.

Chapter 1

Introduction

Effective detection and quantitation of analyte compounds is a cornerstone of the modern technological landscape. From pathogen detection to analyzing the content of a soil sample, the diverse array of techniques that have been developed rely on the sensitive and specific detection of specific ions, small molecules, proteins, and nucleic acids[1, 2]. The past century and a half of chemical and biological research has led to the development of an incredible number of analytical techniques, from the earliest spectrophotometers for chemical analysis to current state-of-the-art molecular detection using highly precise instrumentation, such as LC-MS/MS[3, 4, 5]. The vector of technological development over the decades of research has pointed towards assays that are more sensitive, more specific, and have a broader range of potentially-detectable target compounds, while simultaneously becoming more affordable and easier to use[6, 7, 8, 9].

While the capabilities of human-designed analysis technologies are impressive, those found in biological systems are comparable in their abilities. Natural biological, genetically-encoded sensing systems are capable of responding to an impressive array of inputs while retaining sensitivity and specificity. For example, the human immune system is thought to be able to respond to over 10^7 antigens, while sensitivity is enabled by the extremely tight and specific binding of antibodies to their respective epitopes[10, 11, 12]. Even organisms as relatively simple as prokaryotes in total express hundreds of thousands of different sensors capable of responding to distinct stimuli[13, 14, 15, 16]. While genetically-encoded sensing systems are inherently highly multiplexed, they also have other benefits relative to established analytical techniques. The production of genetically-encoded sensors only requires, at the base level, gene expression machinery and energy substrates to produce sensing gene products, making expression systems

relatively inexpensive to operate[17]. Biological systems are also capable of producing output signals that are easily coupled to mature sensing platforms, such as light, fluorescence, or electrical current[18, 19, 20].

The challenge of developing sensor platforms that use natural biological sensing systems is that, while they are sensitive, specific, multiplex, and fast-acting, the output of natural systems is coupled to the regulation of gene expression processes. These responses are useful for the organism, but rarely usable by human researchers. In order to understand how to repurpose these systems, we first need to understand how they function. Biological sensing systems couple a biomolecular recognition event, where a biomolecule such as a protein binds its cognate ligand, to signal transduction process that results in a change in gene regulation. This can occur by changing the transcription or translation rates, by post-translationally modifying proteins, or even by changing the primary DNA sequence[21, 22, 23, 24]. Some natural systems, such as bacterial chemical-sensing transcription factors, offer some degree of modularity. However, re-engineering protein-based systems to respond to different inputs, or transduce the signal via a different regulatory mechanism, can be challenging. Accurate *de novo* computational protein design is still an open problem, and re-engineering an existing protein requires the screening of thousands of protein variants, which is still a small fraction of the total sequence space[25, 26].

While protein-based sensing systems predominate in biological systems, RNA-based recognition also plays a key role. Like proteins, RNA recognition elements are capable of detecting diverse ligands. RNA elements, called aptamers, are capable of folding into specific structures and binding to their target ligands with a high degree of affinity[27, 28]. Riboswitches, which incorporate aptamers into the regulatory sequences of an mRNA, are widespread among prokaryotes, and can turn on or off gene expression in response to ions and small-molecule metabolites[29, 30]. RNA-based nucleic acid recognition is biologically ubiquitous, and is used in

systems like CRISPR and RNA interference to identify target nucleic acids and degrade them[15, 31, 32].

Crucially, re-engineering nucleic acid-based detection is a much more tractable problem than the corresponding process with proteins. Nucleic acid-detecting elements rely on base-pairing for recognition, and so can easily be re-engineered to respond to different target sequences. While engineering novel chemical-detecting RNA recognition elements faces some of the same conceptual challenges as for proteins, the smaller sequence space (4^N versus 20^N) and dominance of predictable intramolecular base-pairing on overall structure simplifies the discovery of new aptamers, either computationally or experimentally[33, 34]. Of equal importance, well-developed models of nucleic acid structures and energies allow for the facile design of shape-changing RNAs, whose induced structural changes cause a change in the biological function of the molecule[35]. These *in silico*-designed RNA-based sensors can then be tested in their native biological context, for example, to see if they switch on the translation of a reporter gene when a ligand is added[36].

In this dissertation, I seek to expand the capabilities of nucleic acid-based sensing by coupling computational nucleic acid sensor design with fundamental investigation into the biochemistry and biophysics of the biological contexts in which these engineered sensors operate. I first implement RNA-based sensors of biomarker proteins in cell-free protein expression, and examine how gene expression processes are affected by the formulation of these systems. I next examine the sequence-function relationships of mutated sgRNA handles in directing Cas9 activity, and subsequently use these design rules to engineer Cas9-based, potentially highly multiplexable, viral RNA sensors. Overall, I demonstrate the utility of computational nucleic acid design for the development of highly-functional sensor sequences, and develop sensor elements with the potential for real-world utility.

In the next chapter, I discuss coupled transcription-translation (TX-TL) cell-free expression systems, which are increasingly used as sensor platforms. The composition of TX-TL cell-free expression systems are adjusted by adding macromolecular crowding agents and salts. However, the effects of these cosolutes on the dynamics of individual gene expression processes have not been systematically quantified. Here, I carry out kinetic mRNA and protein level measurements on libraries of genetic constructs using the common cosolutes PEG-8000, Ficoll-400, and magnesium glutamate. By combining these measurements with biophysical modeling, I show that cosolutes have differing effects on transcription initiation, translation initiation, and translation elongation rates with trade-offs between time-delays, expression tunability, and maximum expression productivity. I also confirm that biophysical models can predict translation initiation rates in TX-TL using *E. coli* lysate. I discuss how cosolute composition can be tuned to maximize performance across different cell-free applications, including biosensing, diagnostics, and biomanufacturing.

In the third chapter, I discuss the subsequent development of cell-free expression systems as protein detecting platforms. A diverse array of genetically-encoded sensor devices have been implemented in cell-free systems, enabling the detection of toxins like fluoride and pathogens like Zika virus in a portable and cost-effective platform. While sensitive and specific small-molecule and nucleic acid detection have been demonstrated in cell-free systems, protein detection has lagged behind. Here, I use aptamer-based RNA sensing elements that detect specific proteins to design riboswitches, using a previously-developed model of riboswitch translational regulation, the Riboswitch Calculator. I designed 35 riboswitches, detecting the phage MS2 coat protein, human monomeric C-reactive protein, and human IL32γ, and used a TXTL cell-free expression system as a test platform to determine riboswitch function. I discovered that some protein ligands, like CRP, are capable of repressing translation without inducing structural change in the RNA, and in the absence of obvious steric conflict with ribosome. I further demonstrate the importance of accurate

structural knowledge of aptamers for the precise design of highly-functional riboswitches. Overall, I expand the genetically-encoded sensor toolbox, and demonstrate the validity of automated computational design of protein-detecting RNA structural switches.

In the fourth chapter, I discuss the characterization of highly-mutated sgRNAs operating in nuclease-deficient and nuclease-competent Cas9 systems. Engineering cellular phenotypes often requires the regulation of many genes. When using CRISPR interference (CRISPRi), co-expressing many single-guide RNAs (sgRNAs) triggers genetic instability and phenotype loss, due to the presence of repetitive DNA sequences. Using CRISPRi, we stably co-expressed 22 highly-mutated, nonrepetitive sgRNAs within non-repetitive extra-long sgRNA arrays (ELSAs) to simultaneously repress up to 13 genes by up to 3500-fold. However, next-generation CRISPR applications like multiplex genome editing require the use of nuclease-competent Cas9, which differs in mechanism from the catalytically-dead Cas9 (dCas9) used in our CRISPRi characterization by performing a final DNA cleavage step. I sought to validate the function of our highly-mutated sgRNAs in Cas9 by testing sgRNAs with either specific mutations in key functional regions, or an overall high degree of mutation, in in vitro Cas9 cleavage assays. I confirmed that mutated sgRNA function in directing CRISPRi correlates to their function in directing Cas9 cleavage of a DNA target. I further investigated the mechanisms of failure of non-functional mutated sgRNAs by examining their ability to bind to Cas9 in electrophoretic mobility shift assays (EMSAs). Finally, I compared the degree of off-target activity of the highest-functioning mutated sgRNAs with the wild-type sequence, in both CRISPRi and Cas9 cleavage assays. Overall, I validate the application of these highly-mutated sgRNA sequences for use in diverse Cas9 applications.

In the fifth chapter, I prototyped a new CRISPR-based nucleic acid detection platform. CRISPR-based diagnostics combine the rapid signal generation of isothermal amplification-based assays with the additional sequence specificity inherent to CRISPR-based nucleic acid recognition. However, the current generation of CRISPR diagnostics are hindered by their nonspecific collateral

cleavage-based output, limiting the potential for multiplex pathogen detection and multiplex sample processing. Here, I demonstrate a new CRISPR nucleic acid detection platform utilizing the single-turnover type II CRISPR system from S. pyogenes, and re-engineering it as a direct RNA detection system with an inherent degree of multiplex activity. Using previously-determined sequence-structure-function rules for sgRNA design, I engineered tracrRNAs to bind to any desired target RNA sequence and direct Cas9 DNA cleavage. I designed and screened 20 tracrRNAs to bind to and detect the SARS-CoV2 N-gene in an isothermal strand annealing, RNA processing, and DNA cleavage workflow. I further optimized the isothermal workflow to reduce workflow time and discover what RNases are required to prepare the Cas9:N-gene RNA:tracrRNA complex for DNA cleavage. Finally, I prototyped a Cas9-triggered isothermal amplification method, based on the existing CRISDA method, towards the goal of ultrasensitive target RNA detection.

Chapter 2

Tuning Cell-free Composition Controls the Time-delay, Dynamics, and Productivity of TX-TL Expression

2.1 Introduction

Cell-free expression systems combine *in vitro* transcription and translation (TX-TL) within a reconstituted cellular environment, enabling the expression of RNAs and proteins in an open biochemical system. Due to the ease of introducing novel components, and subsequent shortening of the design-build-test cycle, these systems have been harnessed across many synthetic biology applications. New genetic parts are rapidly characterized in organism-specific environments[37, 38, 39, 40, 41]. Enzymes can quickly be expressed or combined to prototype synthetic metabolic pathways with maximal productivity[42, 43]. Engineered state-switching RNAs and proteins are expressed, stored, and activated to detect viral nucleic acids, pollutants, and biomarkers of interest[44, 45, 46, 47]. While the cell-free application space has grown considerably over the past few years, our understanding of how the cell-free environment affects genetic circuit and pathway function has not kept pace. This technical debt could limit future application development by reducing our ability to engineer more complex genetic systems inside cell-free environments.

Notably, the compositions of cell-free systems are chemically distinct from any *in vivo* environment. Overall, lysate-based cell-free systems are 20 to 30-fold more dilute than their corresponding cellular systems, though several salts, small molecules, and macromolecular crowding agents are added at much higher concentrations than typically found inside cells[17, 48, 49, 50]. There are now several recipes for different cell-free systems, optimized for different purposes, making it difficult to interpret quantitative measurements and compare results across systems[17, 49].

Changes in solute composition have a poorly understood effect on the many steps in gene expression, particularly on the quantitative activities of promoters, ribosome binding sites, and other genetic parts[51, 52, 53, 54]. Currently, it remains challenging to predict how these physio-chemical solute effects alter genetic part activities. For example, existing biophysical models of translation initiation rate do not take into account differences in solute composition[55, 56]. Consequently, the concentrations of these components are often empirically tuned to maximize a desired functionality, for example, *in vitro* protein expression titers or genetic circuit signal amplification[54, 57, 58].

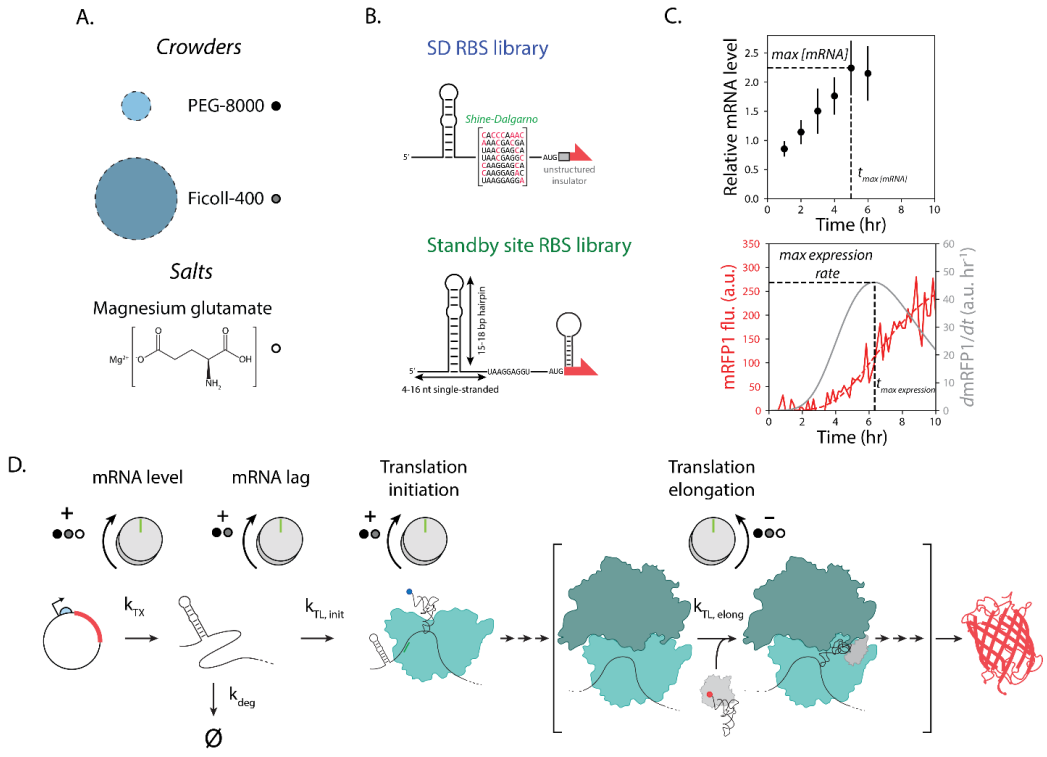

Figure 2-1: Measurements and models to characterize the effects of cosolutes on cell-free TX-TL assays. (A) Crowders and salts used in this study. (B) Sequence and structural differences for two ribosome binding site libraries. The SD RBS library varies translation initiation rates through changes to the Shine-Dalgarno sequence, while the Standby site RBS library varies translation initiation rates by altering the structural geometry of the upstream standby site. (C) The dynamics of mRNA and protein levels were measured using RT-qPCR and spectrophotometry, across several genetic constructs and cosolute compositions, during cell-free TX-TL assays. *mRFP1* expression dynamics were fitted to a 4-parameter logistic equation (red solid line). The first derivative of this equation (gray solid line) was used to determine the maximum expression rate. Error bars indicate

95% confidence intervals. (D) Biophysical modeling is used to distinguish the effects of cosolutes on mRNA dynamics, translation initiation rates, and translation elongation rates.

Here, as part of an effort to develop a more comprehensive mechanistic understanding of cell-free solvent effects, I systematically characterize how cell-free composition controls transcription, translation initiation, and translation elongation rates, as experimentally verified by dynamic mRNA level and protein level measurements. I developed a Markov model of translation that combines a statistical thermodynamic model of translation initiation (the RBS Calculator) with a thermodynamic model of solute-RNA interactions. I show that changes in cell-free composition can have differing effects on translation initiation versus elongation, leading to translation elongation becoming a rate-limiting step to protein expression. The developed model explains how changing the concentrations of commonly added solutes and crowding agents collectively control cell-free protein expression levels. Finally, I suggest how TX-TL reaction compositions could be tuned for various cell-free applications, including genetic system prototyping, biomanufacturing, and sensing.

2.1.1 Hypotheses

1. The cosolutes PEG-8000, Ficoll-400, and magnesium glutamate have distinct effects on mRNA levels and dynamics, protein production rates and dynamics, and maximum translation rates.

2. The cosolutes PEG-8000, Ficoll-400, and magnesium glutamate affect translation rate tunability.

3. Cosolute effects on translation rate tunability can be explained by modeling their effects on both translation initiation and post-initiation processes, especially translation elongation.

4. Biophysical mechanisms, such as volume exclusion and diffusion, can be used to describe model-derived cosolute effects on translation initiation and translation elongation rates.

2.2 Results & Discussion

2.2.1 Kinetic Characterization of Genetic Systems in TX-TL with Varied Cosolute Compositions

I selected PEG-8000 (PEG), Ficoll-400 (Ficoll), and magnesium glutamate (Mg-glut) as three cosolutes commonly added to cell-free expression systems (**Figure 1A**). PEG is a polymer of ethylene glycol with an average molecular weight of 8 kDa and a hydrodynamic radius of about 2.6 nm[50]. Ficoll is a branched polysaccharide polymer with an average molecular weight of 400 kDa and a hydrodynamic radius of about 10 nm. Both PEG and Ficoll are crowding agents that reduce the total volume available to other macromolecules, greatly increasing effective concentrations inside cell-free expression reactions. Mg-glut is a commonly added salt that increases Mg^{2+} concentration, which has significant effects on nucleic acid interactions and mRNA folding.

I then constructed a series of genetic systems expressing mRFP1 utilizing rationally designed ribosome binding sites (RBSs) with varied translation initiation rates. Two types of synthetic RBSs were designed, each varying a distinct interaction responsible for ribosome recruitment (**Figure 1B**). In the first set, the synthetic RBSs utilize different Shine-Dalgarno sequences with systematically varied hybridization energies to the 3’ end of the 16S rRNA, though they all contain an upstream insulating mRNA structure and an unstructured region at the beginning of the protein coding sequence (CDS). These SD RBS library variants were previously characterized using *in vivo E. coli* cultures , where they varied mRFP1 expression levels by 1649-

fold with well-predicted translation initiation rates (RBS Calculator model v2.1; $R^2 = 0.99$, $p = 9x10^{-7}$)[59]. In the second set, the synthetic RBSs utilize different standby site sequences with varied structural geometries, followed by a canonical Shine-Dalgarno sequence (9 nucleotides long), an optimal spacer, and a small mRNA structure inside the beginning of the CDS. These Standby site RBS library variants were also previously characterized using *in vivo E. coli* cultures, varying mRFP1 expression by 40-fold with similarly well-predicted translation initiation rates (RBS Calculator model v2.1; $R^2 = 0.94$, $p = 4x10^{-4}$)[55]. All genetic systems utilized the J23100 promoter, which has a moderate transcription initiation rate.

The purpose of the Standby site RBS library variants is to vary how fast a 30S ribosomal subunit can initially bind to the mRNAs' standby sites. Once a 30S ribosomal subunit is bound, these RBSs all have strong 16S rRNA binding sites, facilitating a rapid transition to forming a 30S pre-initiation complex (PIC) and initiating translation. In contrast, the SD RBS library variants all have a highly accessible upstream standby site, but the differences in their 16S rRNA binding sites lead to different transition rates in 30S PIC formation. Overall, the Standby site RBS library varies the rate of the first key step of translation initiation (a 2^{nd} order binding event) whereas the SD RBS library varies the rate of the second key step in translation initiation (a 1^{st} order sliding event).

For each of these genetic systems, I then carried out kinetic mRNA level and mRFP1 fluorescence level measurements during 10-hour cell-free expression assays, adding 2 nM circular plasmid template, and using baseline or systematically varied cosolute compositions (**Methods**) (**Figure 1C**). For each cosolute composition, mRNA level measurements were performed (N = 3 biological replicates) using RT-qPCR with a temporal resolution of 1 hour. mRFP1 fluorescence level measurements were performed (N = 6 biological replicates) using spectrophotometry with a temporal resolution of 10 minutes. Notably, during the RT-qPCR measurements, I found that endogenous 16S rRNA degrades over time. Therefore, as the internal control, I instead added a synthetic spike-in RNA at fixed concentration to each reaction immediately before RNA extraction

(**Methods**). From these measurements, I determined the dynamics of the mRNA and protein levels, including the time to reach maximum mRNA levels, the first derivative (slope) of the mRFP1 fluorescence levels, and the apparent translation rates of the mRNAs. For the purpose of slope calculations, I found that mRFP1 fluorescence levels fit well to a generalized logistic growth curve, which accounts for background autofluorescence and time delays (**Methods**). From these measurements, I determined how the concentrations of PEG, Ficoll, and Mg-glut differentially controlled the genetic system variants' expression dynamics, including delays in mRNA synthesis, maximum mRNA levels, translation initiation rates, and translation elongation rates (**Figure 1D**).

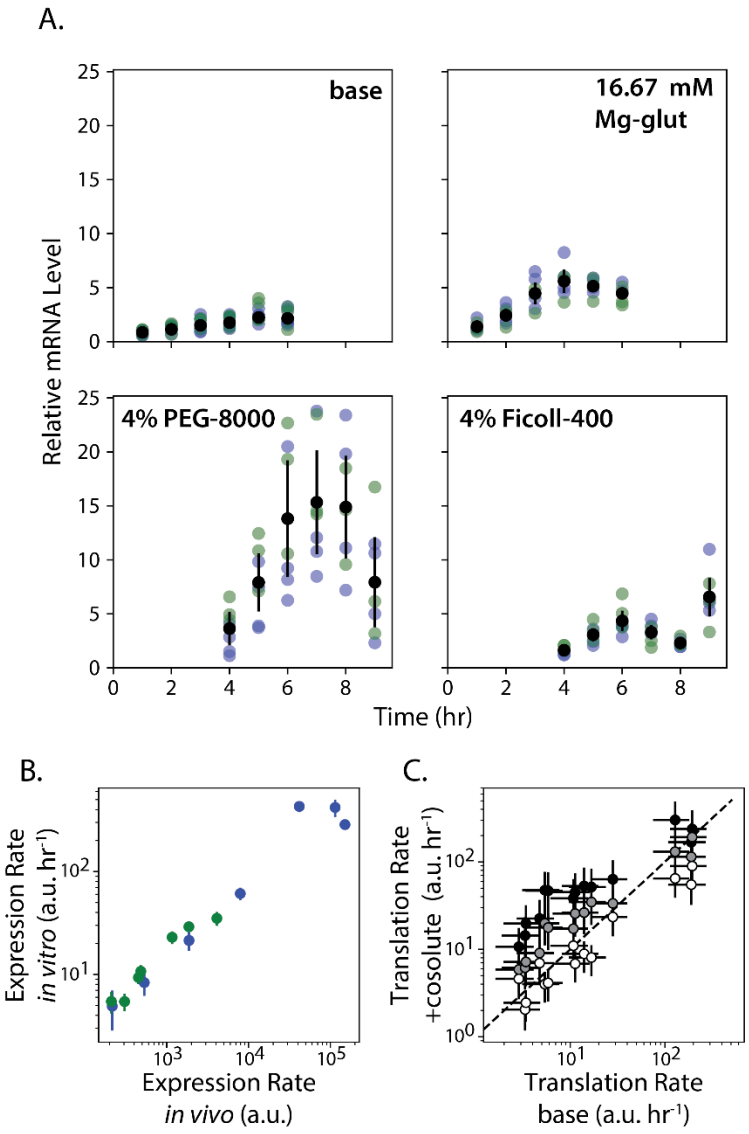

Figure 2-2: Cosolute effects on mRNA levels and translation rates. (A) Relative mRNA levels were measured when using 4% PEG-8000, 4% Ficoll-400, or 16.67 mM Mg-glut, as compared to a baseline composition. Data points are replicate mRNA level measurements from a SD RBS library variant (blue), a Standby Site RBS library variant (green), or the average across all replicates from both RBS library variants (black circles). Error bars represent the 95% confidence interval of at least 7 biological replicates. (B) Measured *in vivo* expression levels were compared to cell-free expression levels across SD RBS library variants (blue) and Standby site RBS library variants (green). (C) Translation rates for all RBS library variants at various cosolute compositions were compared to the translation rates at the baseline composition. Data points include 4% PEG-8000 (black circles), 4% Ficoll-400 (gray circles), and 16.67 mM Mg-glut (white circles). The dotted line indicates no differences. For (B) and (C), data points and error bars represent the mean and 95% confidence interval of at least 6 biological replicates.

2.2.2 Cosolute Composition Differentially Controls Dynamics of mRNA and Protein Levels

I first measured the mRNA level dynamics of two selected genetic system variants in TX-TL reactions with modified compositions of 4% w/v PEG, 4% w/v Ficoll, or 16.67 mM Mg-glut, as compared to the baseline composition. The baseline solution contains 8.67 mM Mg-glut, as TX-TL reactions containing less Mg-glut were unable to support robust expression from the weaker RBSs, but does not contain any PEG or Ficoll. The selected genetic system variants included a Standby Site RBS library variant and a SD RBS library variant with similar predicted translation initiation rates. Notably, I did not find any appreciable difference in mRNA dynamics across these genetic system variants. However, I found that changing the cosolute composition had distinct effects on transcription delays and mRNA maximum levels (**Figure 2A**). Overall, PEG had the highest impact on altering the mRNA maximum level, increasing it by about 6.8-fold, followed by Mg-glut (2.5-fold) and Ficoll (1.9-fold). PEG also had the highest impact on variability in maximum mRNA levels across replicates; its average coefficient of variation across all time points was 0.53, which is about 2-fold higher than the baseline composition and other tested cosolutes. Interestingly, the addition of PEG or Ficoll increased the time needed to reach maximum mRNA levels, due to an apparent delay in mRNA synthesis. The time delay was about 2.5 hours for PEG and 1.3 hours for Ficoll as compared to the baseline composition. In contrast, the addition of Mg-glut did not cause any appreciable difference in time delay.

I then measured the maximum synthesis rates of *mRFP1* mRNA and protein for all genetic system variants – 7 Standby Site RBS library variants and 7 SD RBS library variants – during TX-TL reactions, using modified compositions of either 4% w/v PEG, 4% w/v Ficoll, or 16.67 mM Mg-glut, as compared to the baseline composition. I also measured these genetic system variants' *in vivo* mRFP1 expression levels in *E. coli* DH10B during steady-state cultures maintained in the exponential growth phase (**Methods**). Notably, at the baseline TX-TL composition, the genetic

system variants' protein synthesis rates (expression rates) were highly correlated to their *in vivo* expression levels (R^2 = 0.96, p = $5x10^{-10}$, **Figure 2B**), though the dynamic ranges were starkly different. Changing the RBS sequence varied *in vivo* expression levels by 717-fold, while the same RBS sequences varied *in vitro* expression levels by only 67.9-fold.

I then examined how cosolute composition affected the genetic system variants' apparent translation rates. To calculate the apparent translation rates, I divided the measured protein synthesis rates by the measured mRNA levels. Surprisingly, I found that adding PEG or Ficoll greatly distorted the tunability of apparent translation rates as compared to the baseline composition or *in vivo* measurements. RBS library variants with the lowest measured translation rates (at the baseline composition) had the highest increases in protein synthesis rates (8.8-fold when adding 4% w/v PEG or 3.7-fold when adding 4% w/v Ficoll) (**Figure 2C**). However, this distortion was diminished when using RBS library variants with high measured translation rates. In contrast, adding additional Mg-glut had no appreciable effect on the RBSs' translation rates, which remained correlated with their *in vivo* expression levels. Overall, depending on the cosolute added to TX-TL, there are extremely large changes in maximum mRNA levels, delays in mRNA synthesis, apparent translation rates, and expression tunability.

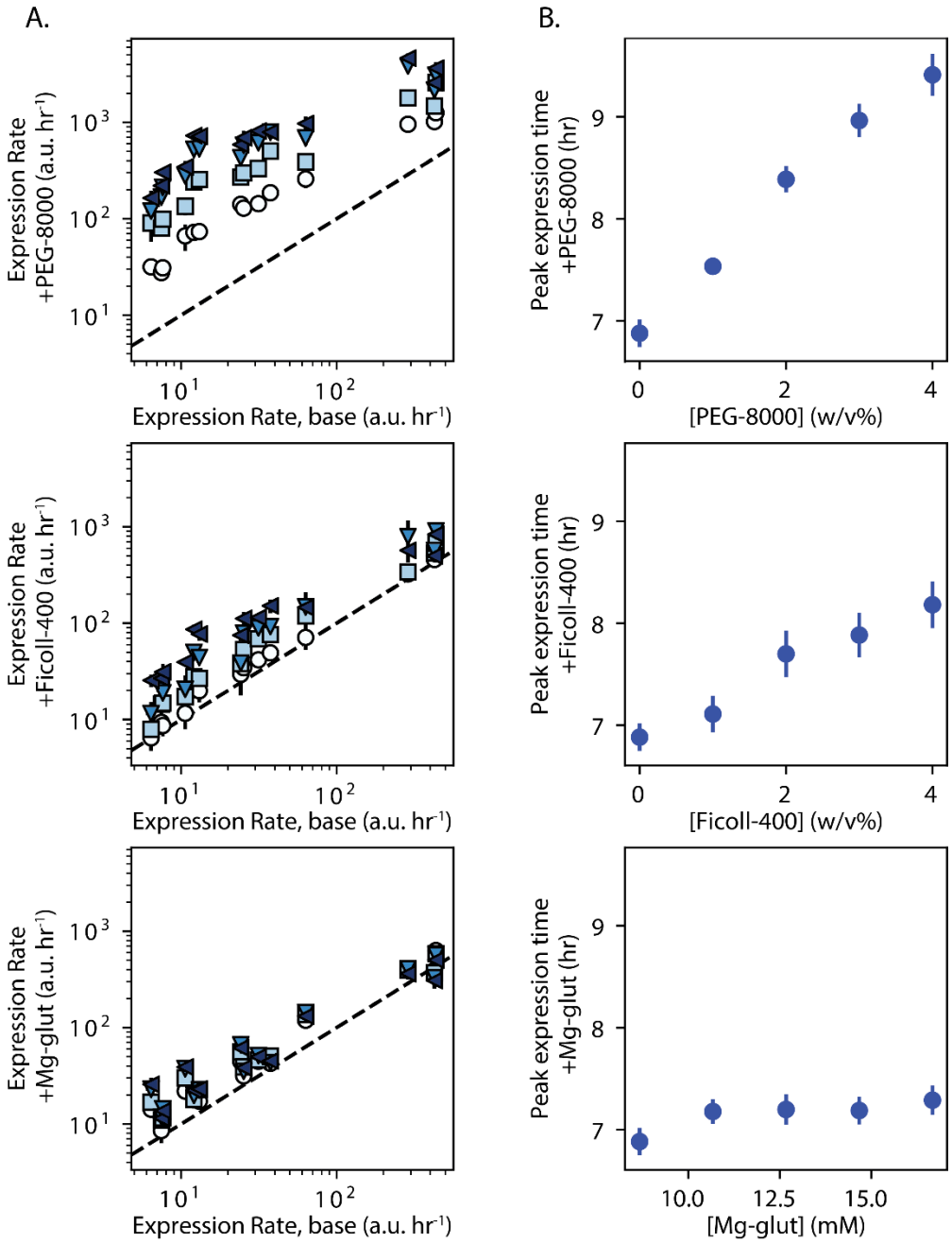

Figure 2-3: The effects of tuning cosolute composition on the maximum expression rate and peak expression time in cell-free assays. (A) The maximum expression rate was determined for all RBS library variants using either 1%, 2%, or 3%, or 4% PEG-8000 (top); 1%, 2%, 3%, or 4% Ficoll-400 (middle); or 10.67, 12.67, 14.67, or 16.67 mM Mg-glut (bottom). Data points of increasing composition are white circles, light blue squares, blue down-triangles, or dark blue up-triangles. Data points and error bars represent the mean and 95% confidence interval of at least 6 biological replicates. (B) The average peak expression time for all tested RBSs was determined when varying cosolute compositions. Data points and error bars represent the mean and 95% confidence interval for 14 genetic systems with 6 biological replicates each.

2.2.3 Cosolute Composition Controls the Magnitude and Timing of Protein Expression Levels

My next objective was to systematically vary cosolute composition and quantify their effects on the magnitude, timing, and tunability of protein expression levels across all 14 genetic system variants with varied RBS sequences. I first found that increasing PEG from 0 to 4% w/v greatly increased maximum protein expression levels by an average of 27.6-fold (**Figure 3A**, top). Similar to my previous measurements of the RBS variants' apparent translation rates, the cosolute effect was most pronounced on the RBS library variants with the lowest translation rates. As a result, the addition of PEG also greatly reduced the dynamic range of expression tunability by 3.1-fold. This effect was significant even at 1% w/v PEG and was further enhanced at higher PEG concentrations. As before, the addition of Ficoll yielded a similar effect with a smaller magnitude; 4% w/v Ficoll increased maximum protein expression levels by an average of 3.6-fold and reduced the dynamic range of tunability by 2.1-fold (**Figure 3A**, middle). Interestingly, Mg-glut only increased maximum protein expression levels by 1.9-fold, but decreased expression tunability by a larger amount (2.8-fold) (**Figure 3A**, bottom). The timing of expression was also significantly affected by cosolute composition. Systematic increases in PEG concentration increased the time needed to reach maximum protein expression by 2.5 hours (**Figure 3B**, top). Similarly, increasing Ficoll concentration resulted in a delay of 1.3 hours to reach maximum protein expression levels (**Figure 3B**, middle). In contrast, changing the Mg-glut concentration had little appreciable effect on overall expression timing (**Figure 3B**, bottom).

2.2.4 Biophysical Modeling Explains Changes in Translation Initiation and Elongation Rates

I next investigated how biophysical modeling can explain both the sequence-dependent and cosolute-dependent effects on translation initiation and elongation rates. As a baseline, I found that the mRFP1 expression levels from the 14 genetic system variants, as measured *in vivo* within *E. coli* DH10B cells, were highly proportional to the RBS Calculator v2.1 model's predicted translation initiation rates, suggesting that translation initiation was a key rate-limiting to protein production (**Figure 4A**, R^2 = 0.813, p = $1x10^{-5}$, N = 14). However, when expressing the same genetic systems in TX-TL, I found that the level of proportionality was reduced when adjusting the composition to either baseline, 4% w/v PEG, 4% w/v Ficoll, or 16.67 mM Mg-glut (overall R^2 = 0.63, p = $2x10^{-13}$, N = 64 conditions, **Figure 4E**). Overall, I found that the cosolute composition had a large impact on the proportionality constant relating model-predicted translation initiation rates to measured expression levels. I also observed a plateau effect whereby higher translation initiation rates did not yield appreciably higher mRFP1 expression levels (**Figure 4E**). Together with my prior measurements (**Figure 2**), these observations suggested that the composition of the TX-TL reaction has distinct effects on the magnitudes of both the translation initiation and translation elongation steps, potentially making translation elongation a rate-limiting step during protein expression.

I therefore augmented the RBS Calculator model to explicitly include the apparent translation elongation rate of the protein's coding sequence. As the starting point, the RBS Calculator calculates the ribosome's binding free energy (ΔG_{total}) using a 5-term free energy model[55, 56, 60, 61] and then predicts a protein coding sequence's translation initiation rate (r_{init}) according to:

$$r_{init} = c_1 \exp(-\beta \Delta G_{total}) \qquad \textbf{Equation 1}$$

where the 30S ribosomal subunit's binding free energy (ΔG_{total}) depends only on the mRNA's sequence and c_1 is a proportionality constant that accounts for extrinsic differences here influenced by the cosolute composition. After the 30S ribosomal subunit binds to the mRNA, it recruits the 50S ribosomal subunit, forms the 70S initiation complex, and initiates translation. Translation continues with a highly processive cycle with elongation rates that depend on codon identities, charged tRNA availabilities, and the cosolute composition. As soon as an elongating ribosome clears the ribosome binding site, a new 30S ribosomal subunit may bind to initiate a new cycle of translation, leading to polysome multi-ribosome dynamics.

Here, my objective is to determine how the cosolute composition affects the translation elongation rate, averaged over the codons in the mRFP1 protein coding sequence. I designated this averaged translation elongation rate as r_{elong} and formulated the simplest possible two-state model (**Figure 4B**) that accounts for how r_{init} and r_{elong} together control the translation rate (r_{TL}), according to the equation:

$$r_{TL} = \frac{r_{init}}{1+\frac{r_{init}}{r_{elong}}} \qquad \textbf{Equation 2}$$

where r_{elong} is c_2, a single cosolute-dependent coefficient that quantifies how the cosolute composition affects the average translation elongation rate of the mRFP1 coding sequence.

I then carried out a model identification procedure to determine the coefficients c_1 and c_2 for each cosolute condition (**Methods**), yielding a mean and confidence interval for each coefficient value. The parameterized two-state model (**Equation 2**) clearly shows how cosolute composition can negatively affect translation elongation rates, resulting in lower translation rates than otherwise predicted by the RBS Calculator model (**Figure 4C**). Overall, by including translation elongation, the two-state model is now able to accurately predict mRFP1 expression levels across all 14 genetic systems and 4 cosolute compositions ($R^2 = 0.81$, $p = 4x10^{-21}$, N = 64, **Figure 4F**).

More importantly, the fitted model shows us how each cosolute affected each translation step, providing an explanation for the observed phenomenon. For example, 4% w/v PEG and 4% w/v Ficoll increased the apparent translation initiation rates by 7.94-fold and 2.85-fold, respectively (**Figure 4D**). In contrast, 16.67 mM Mg-glut did not appreciably change the initiation rate (1.18-fold increase), consistent with my prior measurements (**Figure 3**). Surprisingly, all of the cosolutes had a negative impact on translation elongation rates. 16.67 mM Mg-glut had the largest effect; it lowered the apparent translation elongation rate by 5.6-fold whereas 4% w/v PEG and 4% w/v Ficoll lowered it by 1.82-fold and 2.23-fold, respectively. As a result, the model shows why adding a cosolute decreases the overall expression tunability when utilizing different RBS sequences. For 4% w/v PEG, the model shows that the increase in initiation rate and decrease in elongation rate yields the observed plateau effect whereby RBSs that bind better to ribosomes (strong RBSs) do not yield appreciably more protein than weak RBSs. For 4% w/v Ficoll, a similar plateau effect is predicted, though with lower overall amounts of expressed protein. For 16.67 mM Mg-glut, the model shows that expression tunability is limited by primarily making translation elongation a rate-limiting step in protein production.

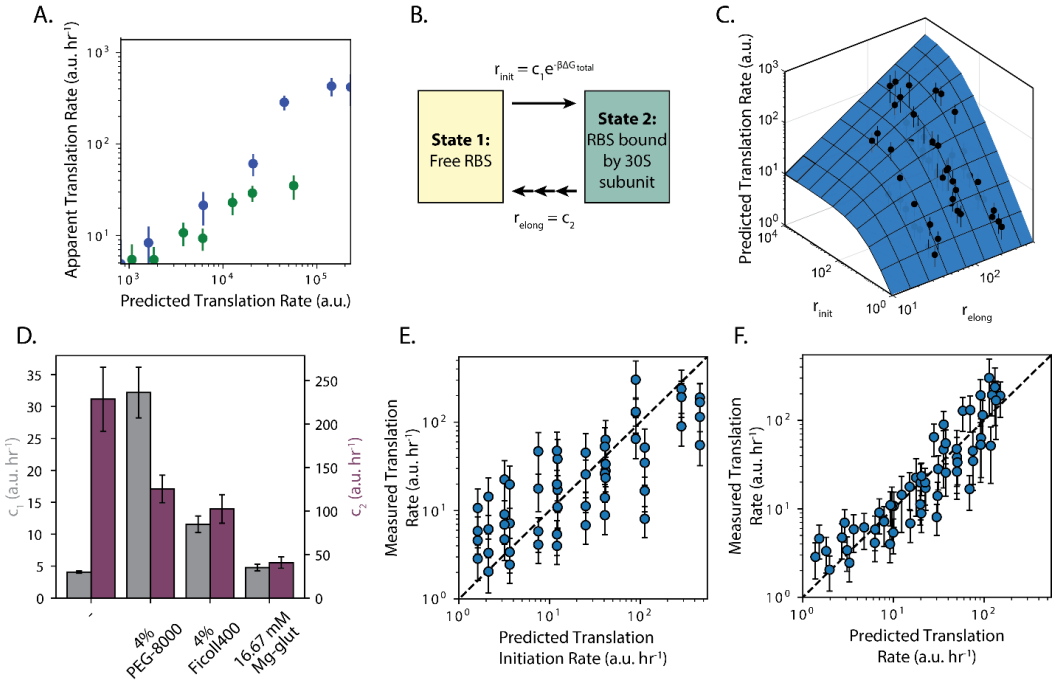

Figure 2-4: Biophysical modeling quantifies effects of cosolutes on gene expression processes. (A) Predicted translation initiation rates were compared to measured in vivo translation rates for SD RBS library variants (blue) and Standby Site RBS library variants (green). (B) A simplified two-state model is used to distinguish cosolute effects on translation initiation and translation elongation steps. (C) The parameterized model shows how cosolute-mediated changes in translation initiation and translation elongation control protein expression levels. (D) Measurements were used to identify parameter values c1 (gray bars) and c2 (purple bars), which quantify cosolute effects on translation initiation and translation elongation. Error bars are 95% confidence intervals for the fitted parameters. (E) Measured translation rates are insufficiently predicted by translation initiation rate alone in vitro. (F) Accounting for cosolute effects on both translation initiation and translation elongation rates increases predictive accuracy of model. Circles and error bars for (A), (C), (E), and (F) represent the mean and 95% confidence interval of at least 6 measurements.

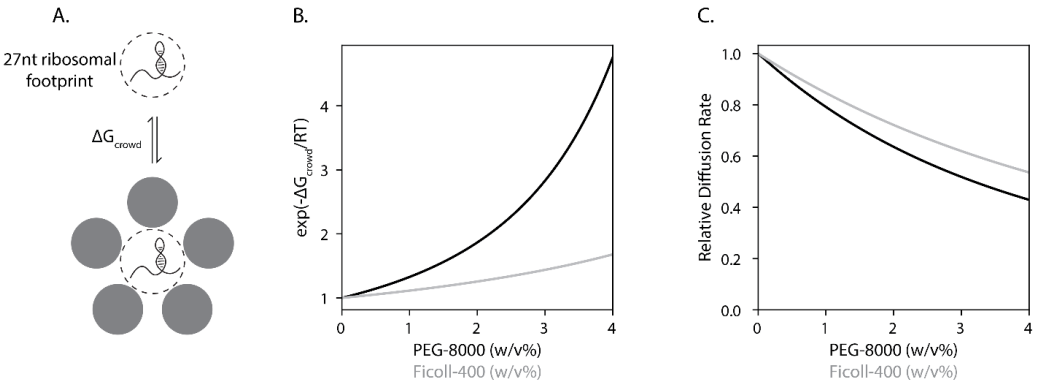

Figure 2-5: Crowding affects both translation initiation and molecular diffusion. (A) PEG and Ficoll bind to the 27-nt ribosomal footprint region of an mRNA with free energy ΔG¬crowd, which

increases a ribosome's ability to bind and initiate translation. (B) Scaled Particle Theory is used to calculate the enhancement of translation rates by varying PEG-8000 (black) and Ficoll-400 (gray) concentrations. (C) The Stokes-Einstein-Huggins equation is used to calculate the decrease in relative diffusion rates when varying PEG-8000 (black) and Ficoll-400 (gray) concentrations.

2.2.5 Physical Modeling Connects Cosolute Intrinsic Characteristics to Extrinsic Effects on Translation

To better understand why PEG, Ficoll, and Mg-glut can have such distinct effects on gene expression, I next applied physical modeling to calculate how cosolute composition affects the kinetic and thermodynamic properties of these non-ideal liquids (**Figure 5**). I first investigated how PEG and Ficoll composition affects the rate of diffusion inside TX-TL reactions. The rate of diffusion affects all binding interactions inside TX-TL, but will particularly affect the charging and loading of tRNAs during translation elongation, which are the slowest and most diffusion-limited steps in protein expression[62, 63] .

To do this, I combined the Stokes-Einstein equation with the Huggins equation to calculate how a cosolute's intrinsic viscosity and its concentration control the solution's diffusion coefficient, yielding:

$$\frac{D}{D_0} = (1 + [\eta]c + k_H[\eta]^2c^2)^{-1} \qquad \textbf{Equation 3}$$

where D is the diffusion coefficient of the TX-TL solution with cosolute present at concentration c, D_o is the diffusion coefficient in the absence of cosolute, $[\eta]$ is the intrinsic viscosity of the cosolute (in a dilute solution), and k_H is the Huggins coefficient of the cosolute. The intrinsic viscosities of PEG and Ficoll are 17 mL/g and 24 mL/g , respectively[64]. Typical Huggins coefficients vary from 0.3 to 0.5; here, in the absence of direct measurements, I assume that PEG and Ficoll both have a Huggins coefficient of 0.4. Using **Equation 3**, I calculated that 4% w/v PEG and 4% w/v Ficoll reduce the TX-TL diffusion coefficient by 2.2-fold and 1.8-fold,

respectively (**Figure 5C**). This reduction in the diffusion coefficient is quantitatively similar to the observed reduction in translation elongation rates for both PEG and Ficoll (**Figure 4D**).

Next, I considered how the divalent salt Mg-glut affects interactions during TX-TL expression. Notably, it has been determined that excess amounts of free Mg^{2+} will inhibit tRNA translocation through the 70S ribosome during translation elongation[65]. The amount of free Mg^{2+} is greatly determined by the concentration of other metabolites in TX-TL that act as chelators, such as glutamate (K_d = 15.1 mM)[66] and phosphoenolpyruvate (K_d = 11 mM)[67]. Taking into account the concentrations of these chelators, I calculate that the concentration of free Mg^{2+} is 0.79 mM in the baseline TX-TL composition (no added Mg-glut). The free Mg^{2+} then increases to 1.45 mM when adding additional Mg-glut to 16.67 mM. Doubling the free Mg^{2+} concentration results in at least a 2-fold reduction in ribosome translocation rate[65], suggesting a causal mechanism for the observed inhibition of translation elongation at high Mg-glut concentrations (**Figure 4D**).

Finally, I considered how PEG and Ficoll affect the thermodynamics of ribosome-mRNA interactions during translation initiation. As crowding agents, PEG and Ficoll interact with other chemical components in solution, reducing the amount of free volume available to them, and promoting the formation of more compact states that take up less space. In this way, the magnitude of the crowding effect depends on the size and shape of the interacting components, particularly larger macromolecules (e.g. mRNAs). Qualitatively, the addition of a crowding agent favors the formation of the 30S:mRNA complex as the bound state takes up less space in solution as compared to a free 30S subunit and a free mRNA. Here, I leveraged Scaled Particle Theory (SPT)[68] to quantitatively calculate the thermodynamic free energies between crowding agents and solution components to determine the magnitude of this crowding effect. To do this, I first assume that each particle is a hard body sphere and then leverage prior measurements to determine their sizes. The cosolutes each have a defined radius (PEG R_c = 2.6 nm, Ficoll R_c = 10 nm), and vary in volume fraction (ϕ_C, unitless) in a composition[69]. As a particle, the 30S ribosomal subunit has a radius of

about R_m = 11 nm[70, 71]. I then consider only the portion of the mRNA that binds to the 30S ribosomal subunit during translation initiation – a 27-nucleotide region called the ribosome footprint – and its defined radius (R_m = 1.75 nm)[72]. I then use SPT to calculate the positive free energy (ΔG_{SPT}) when crowding agent binds to either the 30S ribosomal subunit, the 27-nt mRNA region, or the bound 30S:mRNA complex, according to the equation:

$$\Delta G_{SPT}/k_BT = -\ln(1-\phi_c) + 3\frac{\phi_c}{1-\phi_c}\frac{R_m}{R_c} + 3\left(\frac{\phi_c}{1-\phi_c}\right)^2\frac{R_m}{R_c} + \frac{9}{2}\left(\frac{\phi_c}{1-\phi_c}\frac{R_m}{R_c}\right)^2 + \frac{\phi_c}{1-\phi_c}\left(\frac{R_m}{R_c}\right)^3 + 3\left(\frac{\phi_c}{1-\phi_c}\right)^2\left(\frac{R_m}{R_c}\right)^3 + 3\left(\frac{\phi_c}{1-\phi_c}\right)^3\left(\frac{R_m}{R_c}\right)^3 \quad \textbf{Equation 4}$$

where k_BT is the Boltzmann constant multiplied by the system temperature. ΔG_{SPT} is positive as it requires an input of energy to add a 30S or mRNA particle into a crowded solution with limited free space. I then calculate the difference in free energy when considering a 30S:mRNA particle added to a crowded solution versus a free 30S subunit and a free mRNA added to a crowded solution, according to Equation 5:

$$\Delta G_{crowd} = \Delta G_{SPT,30S:mRNA} - \Delta G_{SPT,30S} - \Delta G_{SPT,mRNA} \quad \textbf{Equation 5}$$

The quantity ΔG_{crowd} is the negative free energy that quantifies how the addition of crowding agent promotes the formation of the more compact 30S:mRNA complex. Because the ribosome is so much larger than the mRNA, here I assume that the 30S:mRNA particle has the same radius as the 30S particle. Therefore, ΔG_{crowd} simplifies to $-\Delta G_{SPT,mRNA}$. In Figure 5B, I show how ΔG_{crowd} becomes more negative as more crowding agent is added, leading to higher translation initiation rates. I calculate that 4% PEG and 4% Ficoll increase a mRNA's translation initiation rate by 4.8-fold and 1.7-fold, respectively. These calculations agree with the empirical measurements shown in **Figure 4D**.

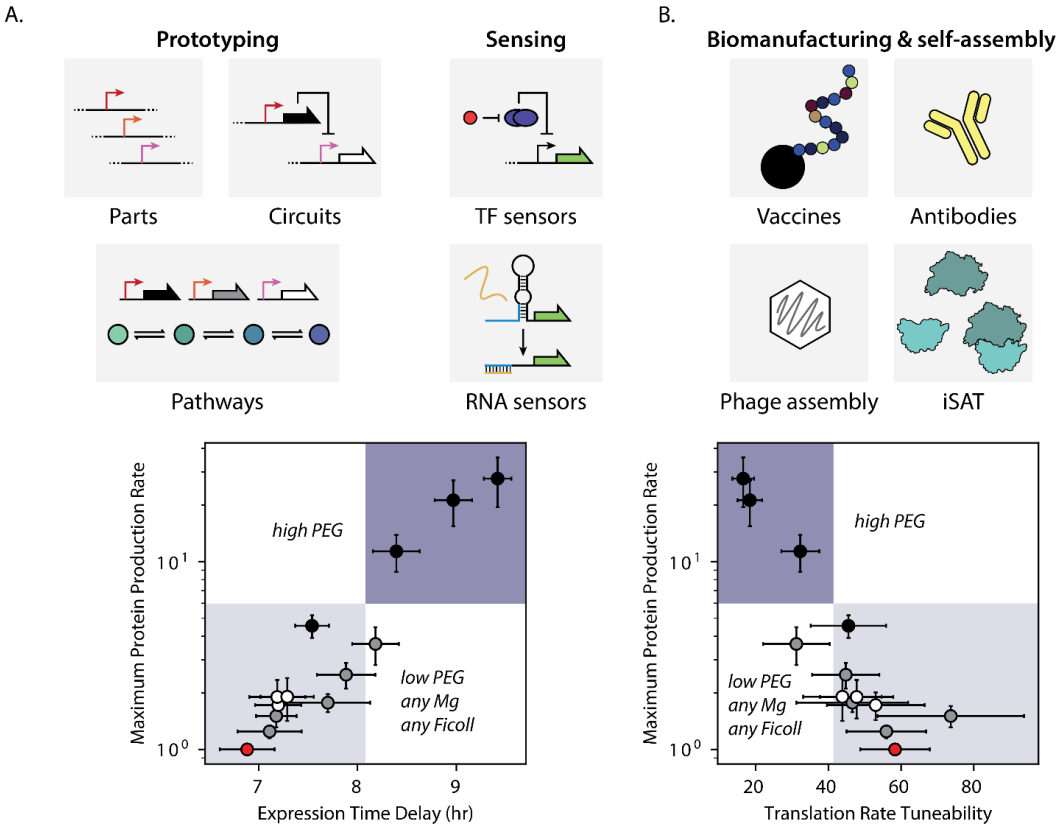

Figure 2-6: Trade-offs between cell-free expression timing, tunability, and magnitude. (A) Several cell-free applications require rapid turn-on times, high tunability, and high dynamic range in expression levels. When tuning cell-free composition, there is a trade-off between a rapid turn-on time and expression tunability. (B) Several cell-free applications require high production rates of protein. Here, cell-free composition can be tuned to maximize protein expression, though sacrificing expression tunability. Circles in (A) and (B) are base composition (red) or increasing concentrations of PEG (black), Ficoll (gray), or Mg-glut (white). Circles and error bars represent the mean and 95% confidence interval at least 6 biological replicates.

2.3 Discussion

Cell-free expression systems are now commonly used for biomanufacturing, diagnostic assays, and genetic part characterization[37, 38, 44, 73, 74, 75, 76, 77, 78, 79]. However, the amounts of cosolute added to prepared cell-free assays can vary considerably across protocols, without a clear understanding of how cosolute composition will affect the performance metrics for each of these

applications[49, 80]. For example, when using cell-free expression to manufacture proteins, it is highly desirable to adjust cosolute concentration to maximize the amount of protein produced, though it less important to activate protein production as rapidly as possible[81]. In contrast, when using cell-free expression as a diagnostic assay, it is more important to rapidly activate production of the observable output[77]. Expression tunability is another important performance metric when multiple proteins need to be expressed at different rates.

Here, I investigated how the concentrations of three cosolutes (PEG, Ficoll, and magnesium glutamate) affected the dynamics of transcription, translation initiation, and translation elongation across 14 genetic systems with varied genetic parts. Overall, I found that all cosolutes increased transcription rates, with PEG having the highest impact. All cosolutes decreased translation elongation rates, with Mg-glut having the highest impact. However, only PEG and Ficoll increased translation initiation rates with PEG having the highest impact. Altogether, the addition of cosolutes increased protein synthesis rates, though the magnitude of improvement depended on which genetic parts were used. When using weak ribosome binding sites, the addition of PEG or Ficoll increased the apparent translation initiation rate, resulting in higher protein expression. However, when using stronger ribosome binding sites, the addition of any cosolute lowered the translation elongation rate, causing it to become the rate-limiting step during protein production and creating a plateau in the protein synthesis rate. I then applied theory from physical chemistry to explain how the cosolutes' intrinsic differences in viscosity, size, and charge could be responsible for these effects through alteration of the cell-free assay's solvent properties.

From these results, I observe distinct trade-offs between timing, expression tunability, and maximum protein production that affect how cosolute composition directly impacts an application's performance metrics (**Figure 6**). Adding PEG or Ficoll will lead to much higher protein production rates, though at the cost of introducing a substantial time delay and limiting the ability to tune protein expression levels by varying translation rates. In contrast, Mg-glut has a

much smaller effect on all these performance criteria. With these trade-offs in mind, it becomes possible to rationally tune cosolute composition towards maximizing performance metrics for a particular application. In a biomanufacturing application where only a single protein needs to be expressed, adding a large amount of PEG will increase the overall protein production rate and titer. However, when multiple proteins need to be expressed at different levels, the PEG and Ficoll concentrations can be tuned to achieve the minimum level of expression tunability, while maximizing the overall protein production rates. In contrast, for a diagnostic application where the time delay becomes more important, the absence of any additional cosolute may instead be the optimal choice.

Finally, my results show that it can be highly misleading to utilize cell-free systems to characterize genetic parts for later re-use in *in vivo* systems. The cosolute composition of all cell-free systems are distinct from the *in vivo* environment, and these differences play an important role in controlling transcription and translation rates. First, differences in genetic part activities (e.g. the translation rates of ribosome binding sites) are often compressed inside cell-free expression systems, leading to lower changes in protein expression levels than otherwise expected. Second, because cosolutes decrease a mRNA's translation elongation rate, the role of synonymous codon usage in protein coding sequences could play an even greater role in cell-free expression systems, particularly when using strong ribosome binding sites and/or high PEG/Ficoll concentrations. While we can apply theory from physical chemistry to calculate and predict how cosolute composition controls genetic part activities in cell-free systems, we should never assume that *in vivo* and cell-free systems will yield quantitatively equivalent results.

2.4 Conclusions

The amounts of macromolecular crowding agents and salts are commonly tuned to optimize protein expression within cell-free expression systems. However, the steps of gene expression these cosolutes effect mechanisms of action, are poorly understood. To better understand how to optimize the composition of these systems for different cell-free applications, I characterized the effects of these cosolutes on mRNA and protein level dynamics. I further derived their effect on the maximum translation rate from each member of a library of ribosome binding site constructs. I extended the RBS Calculator model of translation initiation to incorporate the effects of post-initiation events, especially translation elongation, on the overall translation rate. I finally connected the observed cosolute effects on translation initiation and translation elongation to biophysical descriptions of the effects of macromolecular crowders on macromolecular association equilibria and diffusion rates. Finally, I considered the tradeoffs between maximizing protein expression, minimizing the time delay of expression, and maximizing translation rate tuneability that emerge when adding increasing amounts of crowding agent.

2.5 Methods

2.5.1 Crude cell lysate preparation

Crude cell lysate was prepared according to a previously published protocol, with the following modifications[17]. 20 L of Escherichia coli BL21 with the Rosetta2 plasmid encoding rare tRNAs was cultured in a Micros 30-liter fermentor (New Brunswick) in 2XYT+P medium until the cells reached an OD600 of 1.5-2.0. The cell pellet was then collected in a T-1-P Laboratory continuous flow centrifuge (Sharples), and resuspended in 1 mL S30A buffer per gram of cell

pellet. The resuspended cells were run through a M110-EH-30 microfluidizer (Microfluidics Corp.) at 20,000 PSI twice to ensure complete lysis. The lysate was clarified by centrifugation at 12,000xg for 30 minutes at 4C. The clarified lysate was then incubated for 80 minutes at 37 minutes while undergoing orbital shaking to perform the runoff reaction. After incubation, the lysate was centrifuged again for at 12,000xg for 30 minutes at 4C.

Following lysis, clarification, and the runoff reaction, the lysate was diafiltered with a Pellicon Biomax 10 kDa MWCO 0.005 m^2 ultrafiltration module. Six retentate volumes of buffer S30B were run against the lysate at 4C. After diafiltration, the retentate was centrifuged for 30 minutes at 12,000xg at 4C. The protein concentration of the retentate was quantified using a Bradford BSA Protein Assay Kit assay (Bio-Rad). The retentate was aliquoted and flash-frozen in liquid nitrogen, and stored at -80C.

2.5.2 Cell-free expression reactions

Cell-free expression reactions were assembled on ice according to previously published protcols, with the following modifications[17, 80]. Amino acid and energy solutions were prepared separately, and combined with crude cell extract to reach the following final concentrations: 7.4 mg/mL protein (1/3rd total reaction volume), 1.5 mM each amino acid (except for leucine at 1.25 mM), 50 mM HEPES, 1.5 mM ATP and GTP, 0.9 mM CTP and UTP, 0.2 mg/ml tRNA, 0.26 mM CoA, 0.33 mM NAD, 0.75 mM cAMP, 0.068 mM folinic acid, 1 mM putrescine, and 30 mM PEP. Unless otherwise indicated, 4 mM additional magnesium glutamate *(8.67 mM total), 80 mM additional potassium glutamate (100 mM total), and no crowding agents were added to each reaction. Plasmid DNA containing an RBS variant controlling the expression of an mRFP1 reporter was either miniprepped and ethanol precipitated, or midiprepped and isopropanol precipitated, and added to the reaction to a final concentration of 2 nM. 5 uL reactions were incubated at 29°C for

12 hours in a 96-well polypropylene conical bottom plate sealed with a plate storage mat (Corning) in a TECAN Spark microplate reader. mRFP1 fluorescence was measured every 10 minutes, using 584nm/60nm ex/em with a 5 nm bandwidth.

2.5.3 RT-qPCR

5 uL TXTL reactions, assembled as above, were incubated at 29C in 96-well Costar conical-bottom plates in a TECAN Spark.After incubating for the given amount of time, reactions in microcentrifuge tubes were flash-frozen in liquid nitrogen, while reactions in 96-well plates were directly processed. To each reaction, 500 pM iclR normalization control RNA was. Total RNA, including the normalization control, was extracted using a Norgen Total RNA Extraction kit, after which any remaining plasmid DNA was removed via digestion with TurboDNase (Invitrogen). RNA integrity was verified via agarose gel electrophoresis. RNA was then diluted to 100 ng/uL, and 25 ng/uL yeast tRNA was added. First-strand cDNA synthesis was performed using a High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems). qPCR reactions were assembled using PowerUp SYBR Green Master Mix (Applied Biosystems) and primer sets for mRFP1 and iclR, and run in a StepOnePlus Real-Time PCR System (Applied Biosystems). Melting curve analysis was performed to confirm product homogeneity. RNA levels were calculated using a variant on the ddCt method, accounting for differences in probe efficiency.

2.5.4 Fluorescence time-course data analysis

To determine the time-course rates of gene expression, I fit the fluorescence production timecourse for each TXTL reaction to a generalized logistic growth model of the form[82]:

$$mRFP1\ flu = A\left(1 + (d-1)\exp\left(\frac{-k_U(t-T_i)}{d^{\frac{d}{1-d}}}\right)\right)^{\frac{1}{1-d}}$$

Where A is the upper asymptote, T_i is the time at inflection, k_U is the relative maximum growth rate, and d is a parameter controlling the inflection value. The generalized logistic curve allows for a variable amount of asymmetry about the inflection point, which better describes the behavior of fluorescent protein production over time in the cell-free expression system used here. Timecourses were corrected for background fluorescence by subtracting out the average fluorescence from 30-80 minutes. Fitting was performed using scipy's curve_fit module. The rate of fluorescence production was calculated as the gradient of the fit fluorescence timecourse.

Chapter 3

Automated design of genetically-encoded protein-sensing riboswitches in a cell-free expression system

3.1 Introduction

Synthetic biologists have developed a diverse set of genetically-encoded sensor systems for the facile detection of many kinds of analytes, from small molecules to nucleic acids, by regulating the expression of downstream genes[83, 84, 85, 86, 87, 88, 89, 90, 91]. Although many genetically-encoded sensing elements and systems have been implemented in cells, their utility for detecting compounds of environmental and biomedical relevance is limited by the requirement for low toxicity, and few sensing modalities allow for the detection of exogenously-added macromolecules[92, 93, 94, 95]. Cell-free expression systems, where the expression machinery is removed from the cellular context and reconstituted in cell-free protein expression reactions, lack these cell-based sensor platform limitations, and have been increasingly used as genetically-encoded sensor platforms[77, 96, 97]. In cell-free expression, there is no barrier between the expression machinery and exogenously-added analyte, making the detection of bulky molecules like nucleic acids or toxic compounds like fluoride using gene-expression based sensors simple and feasible. This combination of attributes has made them attractive as sensor platforms for diverse chemicals and nucleic acids. Cell-free sensors for environmental toxins and novel pathogens have already been tested and deployed in the field[45, 77]. In contrast, few protein-detecting sensing platforms have been implemented in cell-free systems[98, 99].

Protein detection is a cornerstone of both modern medical diagnostics and biological research in general[100]. Currently, immunoassay-based methods, like ELISAs, or mass spectrometry,

are considered state-of-the-art, offering high sensitivity and specificity, and the ability to detect a diverse range of protein targets[101, 102]. However, these assays require expensive equipment and reagents, as well as cold chains from both the manufacturer and point-of-care to a core facility. Genetically encoded sensors operating in a cell-free context, in contrast, are inexpensive and can be lyophilized to become stable at ambient conditions, attractive traits for point-of-care diagnostics[96, 103, 104]. However, almost all genetically-encoded recognition elements currently used in cell-free systems, such as natural metabolite-sensing transcription factors or complementary sequences within a toehold switch, respond to small molecules or nucleic acids, not proteins. Another class of genetically-encoded recognition element, nucleic acid aptamers, have been engineered to bind to diverse molecules, including proteins[105]. Aptamers that bind small molecules have been incorporated into riboswitches, regulatory elements which switch gene expression on or off[106]. Small-molecule detecting riboswitches have already been incorporated into cell-free systems[45, 107, 108]. However, a generalizable approach to protein detection using aptamer-based sensing elements has not yet been demonstrated.

Here, I demonstrate the use of a previously-developed biophysical model, the Riboswitch Calculator, to design riboswitches which detect proteins and activate or repress the translation of mRFP1, a reporter protein[36]. I designed and characterized the function of 35 rationally designed riboswitches, detecting the MS2 phage coat protein, human modified C-reactive protein (mCRP), and human IL-32γ, in a TX-TL cell-free expression system[17]. The designed riboswitches have similar dose-responses to their respective ligands as the aptamers from which they were designed. I further demonstrate that these riboswitches can affect gene expression via two mechanisms: secondary structure switching, which can activate or repress gene expression; and steric repression from a protein bound in the standby site of the 5' untranslated region (5'UTR) which inhibits translation initiation. I then demonstrated the predictive accuracy of Riboswitch Calculator on the developed riboswitches, especially when the actual region of aptamer interaction with its protein

ligand is taken into account. Overall, I validate a strategy of *in silico* design and *in vitro* testing for the further development of riboswitch-based protein sensors which function in the cell-free expression environment.

3.1.1 Hypotheses

1. The Riboswitch Calculator model of aptamer-based translational regulation can be used to design functional riboswitch sequences using protein-detecting aptamers.

2. C-reactive protein can directly repress translation when bound to an aptamer placed in the 5'UTR.

3. The dose-response of each riboswitch to its respective ligand is governed by the equilibrium binding affinity of the full ribosome binding site to the ligand.

4. Model predictions of riboswitch activation or repression vary strongly based on the size of the supplied aptamer constraint structure.

3.2 Results

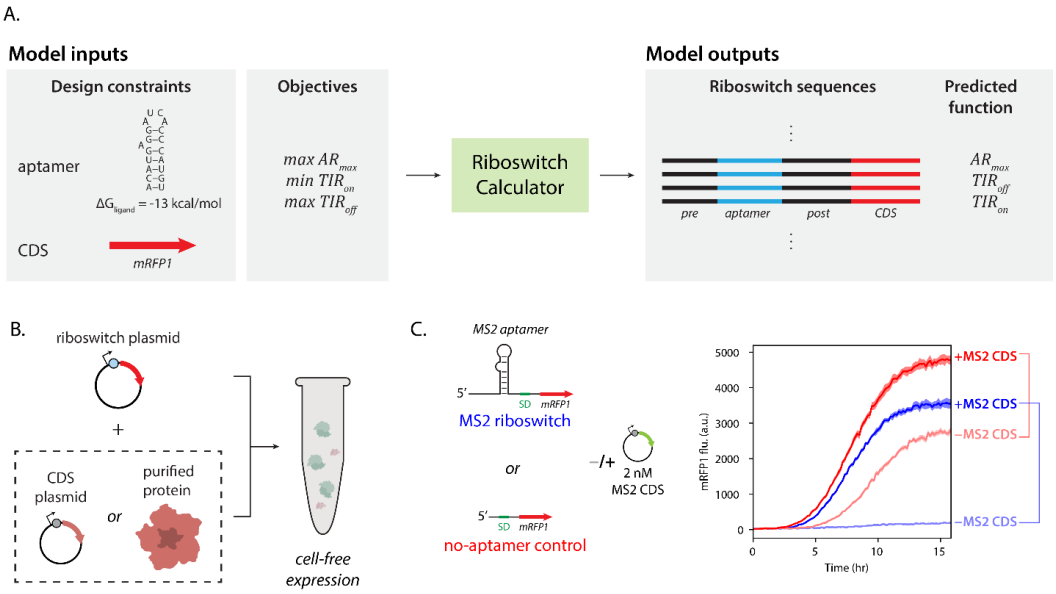

Figure 3-1: In silico protein-detecting riboswitch sequence design and in vitro riboswitch testing. A. The Riboswitch Calculator model of translational riboswitch regulation takes as inputs the aptamer sequence, secondary structure, binding free energy of ligand to aptamer (ΔG_{ligand}). For forward sequence design, the user selects objectives to be minimized or maximized. Forward sequence design will output a list of designed, optimal pre- and post-aptamer sequences, along with their predicted function. B. The cell-free test system consists of 2 nM riboswitch-expressing plasmid, combined with either a protein-CDS expressing plasmid or the purified protein itself. These two components are added to a TXTL cell-free expression system. C. Measuring riboswitch function. To correct for any nonspecific effect of the analyte on riboswitch expression, mRFP1 levels are measured from both a riboswitch and a no-aptamer control plasmid, with and without added analyte. The reported riboswitch mRFP1 levels and activation/repression ratios are corrected for this nonspecific effect. 2 nM riboswitch + 0 nM MS2 CDS plasmid, light blue; 2 nM ribositch + 8 nM MS2 CDS, blue; 2 nM no-aptamer control + 0 nM MS2 CDS, pink; 2 nM no-aptamer control + 2 nM MS2 CDS, red. Shaded regions around fluorescence timecourses represent the 95% confidence interval (n=6).

3.2.1 In silico riboswitch design and in vitro test platform

To design the protein-detecting riboswitches, I used the previously-developed Riboswitch Calculator model of translational riboswitch regulation. Translational riboswitches function by causing the structure of the 5'UTR to switch between two different structures in the presence and

absence of the ligand. This structural shift can affect the translation rate via two mechanisms: by affecting the secondary structure within the ribosomal footprint, affecting the rate at which the ribosome is able to unfold this structure and bind to the Shine-Dalgarno sequence, or by changing the amount of favorable single-stranded surface area in the standby site region of the 5'UTR, upstream of the ribosomal footprint. The model captures these effects by calculating the free energy of the 30S ribosomal subunit unfolding mRNA and hybridizing to the SD sequence ($\Delta\Delta G_{mRNA:rRNA}$) and the free energy of single-stranded surface area interacting with the platform domain of the 30S subunit ($\Delta G_{standby}$) (**Equation 1**). These changes affect the free energy of translation initiation (ΔG_{total}), which is related to the translation initiation rate (TIR) by a Boltzmann relationship (**Equation 2**)[109]. By calculating the TIR with and without a supplied aptamer constraint, the riboswitches' predicted activation or repression ratio (AR_{max}) is then calculated as the ratio of the predicted TIRs in the induced and uninduced (on and off) states (**Equation 3**).

Equation 1 $\Delta G_{total} = \Delta\Delta G_{mRNA:rRNA} + \Delta G_{standby} + \Delta G_{spacing} + \Delta G_{start}$

Equation 2 $TIR \propto e^{-\beta \Delta G_{total}}$

Equation 3 $AR_{max} = \frac{TIR_{on}}{TIR_{off}} = e^{-\beta(\Delta G_{on} - \Delta G_{off})}$

To design riboswitch sensors *de novo*, a known aptamer sequence, structure, and binding affinity must be provided, as well as the coding sequence of the protein whose expression will be regulated by the riboswitch. Given these sequence and structural constraints, as well as the desired riboswitch function (e.g. maximizing the activation ratio, or minimizing the off-state translation rate), the model will output optimized riboswitch 5'UTR sequences, along with their corresponding predicted function (**Figure 1A**).

The Riboswitch Calculator has previously been applied to design genetically-encoded sensors for diverse chemical inputs, including drugs like theophylline and environmental toxins like 2,4-dinitrotoluene and fluoride. However, it has not yet been applied for the rational design of

RNA-based protein biosensors. Protein-binding RNA aptamers are capable of extremely high-affinity binding, similar to antibody-based recognition elements, and are also specific enough to be capable of distinguishing between protein isoforms[110, 111]. These favorable properties indicate that protein-binding RNA aptamers could potentially be used to develop highly sensitive and specific riboswitch-based biosensing elements.

To test the designed riboswitches, I used the previously-developed TX-TL cell-free expression system[17]. I either coexpressed each riboswitch with the CDS encoding its cognate ligand, or added the purified protein ligand directly to the riboswitch expression reaction, and measured the endpoint production of the mRFP1 fluorescent reporter protein (**Figure 1B**). Prior work has demonstrated that gene expression in both the cellular and cell-free context is sensitive to added components, or additional genetic load[57, 80, 112, 113, 114]. I determined the nonspecific effect of each target ligand on mRFP1 expression by measuring the ligand-induced change in expression of a no-aptamer control sequence, and corrected each riboswitch's activation or repression ratio for this effect (**Figure 1C**). Further results discussing each riboswitch's on-state fluorescence, activation ratio, or repression ratio are all corrected for these nonspecific effects on gene expression.

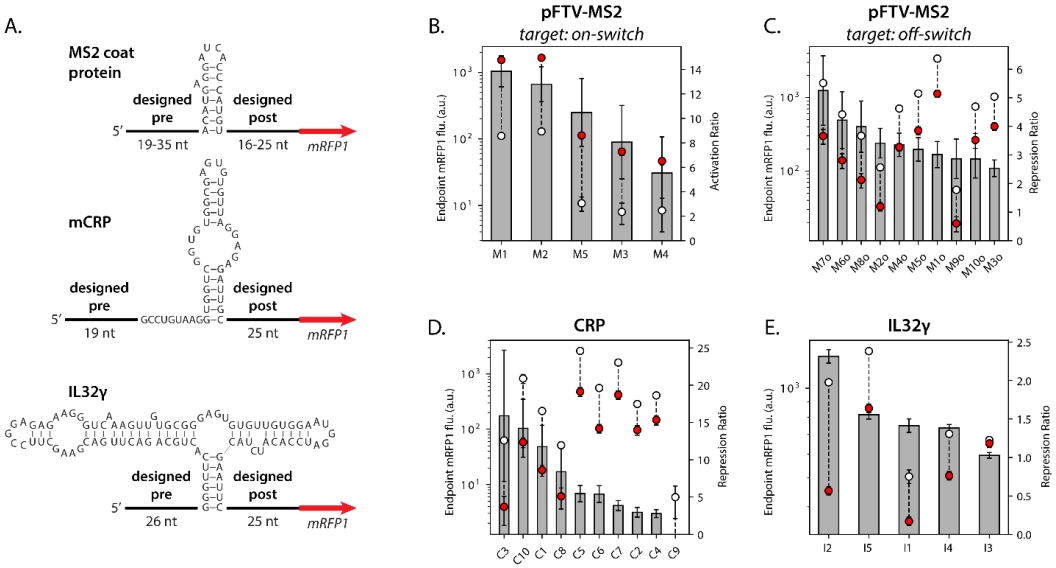

Figure 3-2: Design and function of protein-detecting riboswitches. A. Sequence design of protein-detecting riboswitches for MS2 coat protein, mCRP, and IL32γ. For B., C., D., E., gray bars represent activation or repression ratios for designed riboswitches, while white circles represent off-state endpoint mRFP1 fluorescence, and red circles represent on-state mRFP1 fluorescence. B. MS2 on-switches, induced with 8 nM MS2 CDS. C. MS2 off-switches, induced with 16 nM MS2 CDS. D. CRP on-switches, induced with 1.25 uM of CRP monomer. On-state expression from C9 was less from zero, so its repression ratio is not shown here. E. IL32γ on-switches, induced with 780 nM IL32γ. Error bars for endpoint mRFP1 fluorescence, activation ratios, and repression ratios represent the 95% confidence interval (n=6 for MS2 on-switches; for all other data, n=8).

3.2.2 Design and characterization of MS2, mCRP, and IL32γ riboswitches

To test the *in silico* riboswitch design and *in vitro* functional assay approach, I designed riboswitches using three different aptamers: the MS2 phage coat protein aptamer, a C-reactive protein (CRP) aptamer, and an interleukin-32 (IL32) aptamer (**Figure 2A**). The MS2 coat protein aptamer interaction with the coat protein is extremely high-affinity (K_d = 0.7 nM), and the motif that dictates binding is small (19 nt) and has a well-defined structure[115, 116]. Phage coat protein-responsive riboswitches, including the MS2 hairpin, have previously been successfully developed, making MS2 coat protein-responsive riboswitches a useful validation of the design and test

strategy[117]. The Riboswitch Calculator model also shows good predictive accuracy (R^2 = 0.35, p = $7x10^{-9}$, n = 78) on the maximally induced and uninduced translation rates of these previously-designed riboswitches, further validating the application of the Riboswitch Calculator model to protein riboswitch design (**Supplementary Figure 1, Supplementary Table 1**). C-reactive protein is a biomarker of inflammation, elevated levels of which indicate inflammatory conditions such as heart disease[118]. CRP occurs natively in both pentameric (pCRP) and monomeric, modified (mCRP) isoforms, but mCRP has been shown to be more highly correlated to inflammatory states[119, 120, 121, 122]. The CRP aptamer used here binds specifically to mCRP (K_d = 187.7 nM) , but not pCRP[111]. IL32 is a proinflammatory cytokine, overexpression of which has been linked to several pathologies[123, 124, 125]. I used an RNA aptamer generated against the full-length isomer, IL32γ (K_d = 78 nM), to design riboswitch-based sensors[126].

I initially designed and cloned 30 riboswitches using these three aptamers, designing 20 to activate and 10 to repress the expression of mRFP1 (**Supplementary Table 2, Supplementary Table 5**). I designed and tested 5 activating MS2 coat-protein responsive riboswitches, and screened the response of the 5 designed MS2 on-switches against 8 nM of MS2 coat protein-expressing plasmid (pFTV-MS2) (**Figure 2B**). Each designed riboswitch significantly increased mRFP1 expression upon induction with the coat protein-expressing plasmid, from 5.5-fold by up to 13.8-fold. I then designed and tested 10 repressing MS2 riboswitches, and screened them against 16 nM pFTV-MS2 to ensure more complete repression, as repressing riboswitches require higher-fold induction to achieve a similar signal as activating riboswitches (**Figure 2C**). Each repressing MS2 riboswitch significantly reduced the expression of mRFP1, from a minimum of 2.5-fold to a maximum of 5.2-fold. These results indicate that the Riboswitch Calculator can be used to effectively design both activating and repressing protein-responsive riboswitches, and is in agreement with previous results demonstrating the use of the MS2 coat protein aptamer to design translational activators and repressors.

I then designed and tested 10 mCRP-responsive activating riboswitches, and screened them against 1.25 uM of CRP monomer (**Figure 2D**). Although these riboswitches were designed to activate translation upon ligand binding, they all instead significantly repressed the expression of mRFP1 when induced, from 2.8-fold to up to 15.9-fold. Similarly, I designed and tested 5 IL32γ-responsive riboswitches against 780 nM IL32γ (**Figure 2E**). I observed that 4 out of the 5 significantly repressed expression, albeit to a smaller degree than the CRP riboswitches, from 1.4 to 2.5-fold. Added IL32γ had no significant effect on mRFP1 expression, so the no-aptamer control correction was not used when calculating the IL32 riboswitches' activity. Despite the fact that the riboswitches did not function as designed, they achieved a significant sensor response in response to inducing ligand.

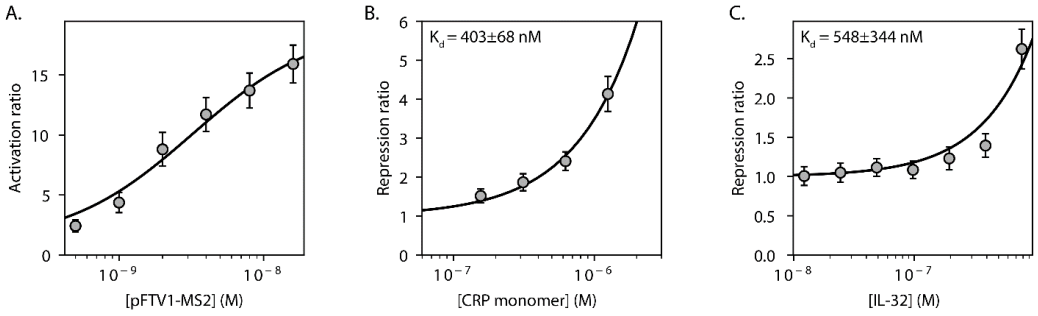

Figure 3-3: Dose-response of MS2, CRP, and IL32 riboswitches. The activation or repression ratios of 3 riboswitches were fit to a Hill equation to estimate the K_d of riboswitch-ligand association. Gray circles represent riboswitch activation ratios at different ligand concentrations, while the black line represents the fit Hill equation. A. Riboswitch M2 was titrated with increasing concentrations of the pFTV-MS2 plasmid. B. Riboswitch C5 was titrated with CRP. C. Riboswitch I2 was titrated with IL32. Error bars for all data represent the 95% confidence interval (M2, n=6; C5, n=6; I2, n=8).

3.2.3 Dose-response of designed riboswitches

For each ligand, I selected a high-functioning riboswitch (M2, C5, and I2 for MS2, CRP, and IL32, respectively), to observe its dose-response relationship, determine its K_d, and its

apparent maximum activation or repression ratio (**Supplementary Table 3**). For riboswitch M2, I varied [pFTV-MS2] between 0 and 16 nM (**Figure 3A**). I confirmed that the increase in protein expression was roughly linear in this DNA concentration regime by measuring mRFP1 from the no-aptamer control plasmid between 0 and 32 nM (**Supplementary Figure 2**). For C5, I varied [CRP monomer] between 0 and 1250 nM, and for I2, I varied [IL32γ] between 0 and 780 nM (**Figure 3B, Figure 3C**). For each riboswitch, I fit the activation ratio at each ligand concentration to the following modified Hill equation (**Equation 4**). Repression ratios were fit to the inverse (**Equation 5**).

Equation 4 $AR_{conc} = 1 + \frac{AR_{max}-1}{\frac{K_d}{[lig]}+1}$

Equation 5 $RR_{conc} = (AR_{conc})^{-1}$

M2 is able to detect at least 0.5 nM plasmid (p=0.0006, two-tailed t-test), with an apparent K_d of 3.2 nM (± 0.62 nM) added pFTV-MS2 plasmid. The M2 AR_{max} was determined to be 19.1 (± 1.2). C5 is able to detect down to at least 156 nM CRP monomer, the lowest concentration of mCRP that I tested (p=0.003, two-tailed t-test). This translates to 4 mg/L, near the high sensitivity-CRP test limit of 1 mg/L. Riboswitch C5's K_d is 403 nM (± 68 nM), similar to its literature value of 187.7 nM[111]. I2 is able to detect down to 195 nM IL32γ (p=0.026, two-tailed t-test). The I2 fit K_d is more variable, at 548 (± 344 nM), but qualitatively similar to literature value of 78 nM. AR_{max} for both C5 and I2 was determined to be extremely low ($4.9x10^{-14}$ ± 0.045 and $8x10^{-14}$± 0.27), although the standard deviations are relatively large, making an accurate maximum repression ratio estimate difficult. These results demonstrate that, while the MS2 coat protein riboswitches are almost fully induced, the CRP and IL32 riboswitches are likely not at full induction at the concentrations of protein assayed.

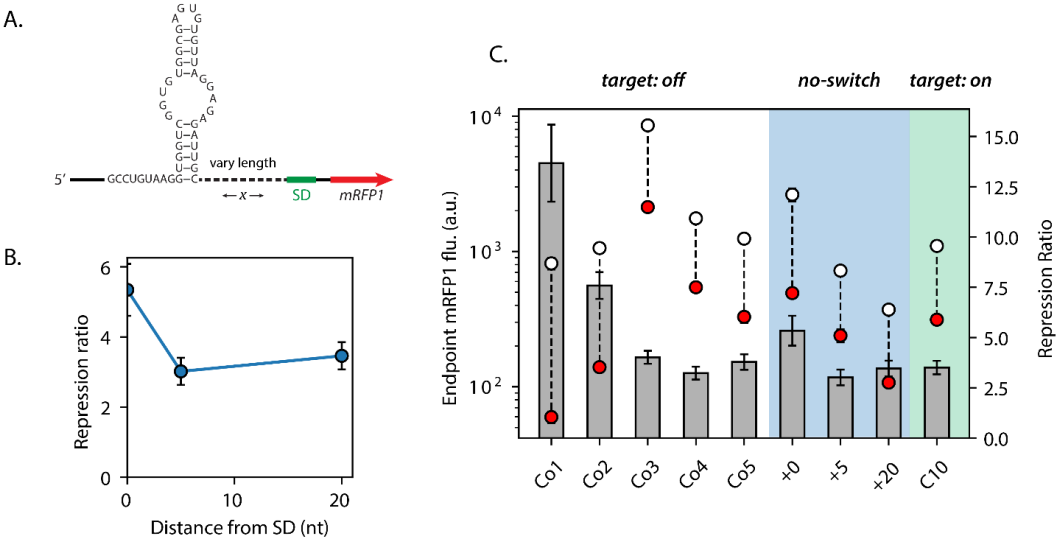

Figure 3-4: Placement of CRP aptamer in the standby site of the 5'UTR represses gene expression. A. Schematic of no-switch controls. The distance between the CRP aptamer and the Shine-Dalgarno sequence was varied between 0 and 20 nt. B. The effect of distance between the CRP aptamer and Shine-Dalgarno on repression of mRFP1 expression. C. Comparing repression between designed CRP off-switches, no-switch controls, and a CRP designed on-switch. Gray bars represent repression ratios, white circles represent off-state fluorescence, and red circles represent on-state fluorescence. Error bars for all figures represent the 95% confidence interval (n=8).

3.2.4 CRP inhibits translation initiation

I hypothesized that the CRP designed on-switches functioned as repressors instead of activators because CRP itself, when bound to the standby region of the 5'UTR, repressed translation initiation by sterically hindering 30S subunit association with the RBS. In order to examine this, I designed three no-switch CRP-binding RBSs, controlling the expression of mRFP1 (**Figure 4A**). These sequences were designed with the CRP aptamer sequence placed upstream of the Shine-Dalgarno sequence, with flanking sequences designed to ensure that the RBS's secondary structure was the same with or without CRP bound. I varied the distance between the Shine-Dalgarno (the 5' end of the ribosomal footprint) and the aptamer in order to determine the distance-dependence of the repressive interaction[127]. Surprisingly, although placing the aptamer directly

adjacent to the ribosomal footprint resulted in 5.3-fold repression of mRFP1 expression, placing it 5 or 20 nt upstream from the footprint resulted in smaller but similar degrees of repression, 3.0 and 3.4-fold, respectively (**Figure 4B**).

To see whether I could design more highly-active CRP riboswitches by designing them as translational repressors, rather than activators, I designed 5 more CRP off-switches using the Riboswitch Calculator (**Figure 4C, Supplementary Table 4**). Four of the five designed riboswitches repressed translation more than the +5 nt no-switch control (Co1, $p=4x10^{-8}$; Co2, $p=1x10^{-8}$; Co3, p=0.002; Co4, p=0.015; two-tailed t-test), with a maximum fold-repression of 13.7-fold. To compare the performance of the no-switch controls and off-switches with an on-switch, I re-characterized C10, one of the highest-repressing designed on-switches. C10 repressed by 3.5-fold, not significantly different (p=0.08) from the fold-repression generated by the no-switch controls. This was 4.1-fold lower than the previous round of characterization, which I attributed to a lower activity of the CRP batch used in this round of experiments. The performance of the designed on-switch relative to the no-switch controls suggests that the main effect dominating its effect on gene expression was the steric effect of CRP bound to the standby site, rather than any change in RBS secondary structure. As C10 was repressed to one of the greatest degrees of the designed on-switches, this suggests that CRP may have induced the other designed on-switches to switch to a structural state more accessible for the 30S ribosomal subunit, but the steric effect on expression dominated the overall sensor response. For further analysis of model predictive accuracy, I corrected the CRP riboswitch repression ratios for the apparent steric effect of repression on riboswitch C10.

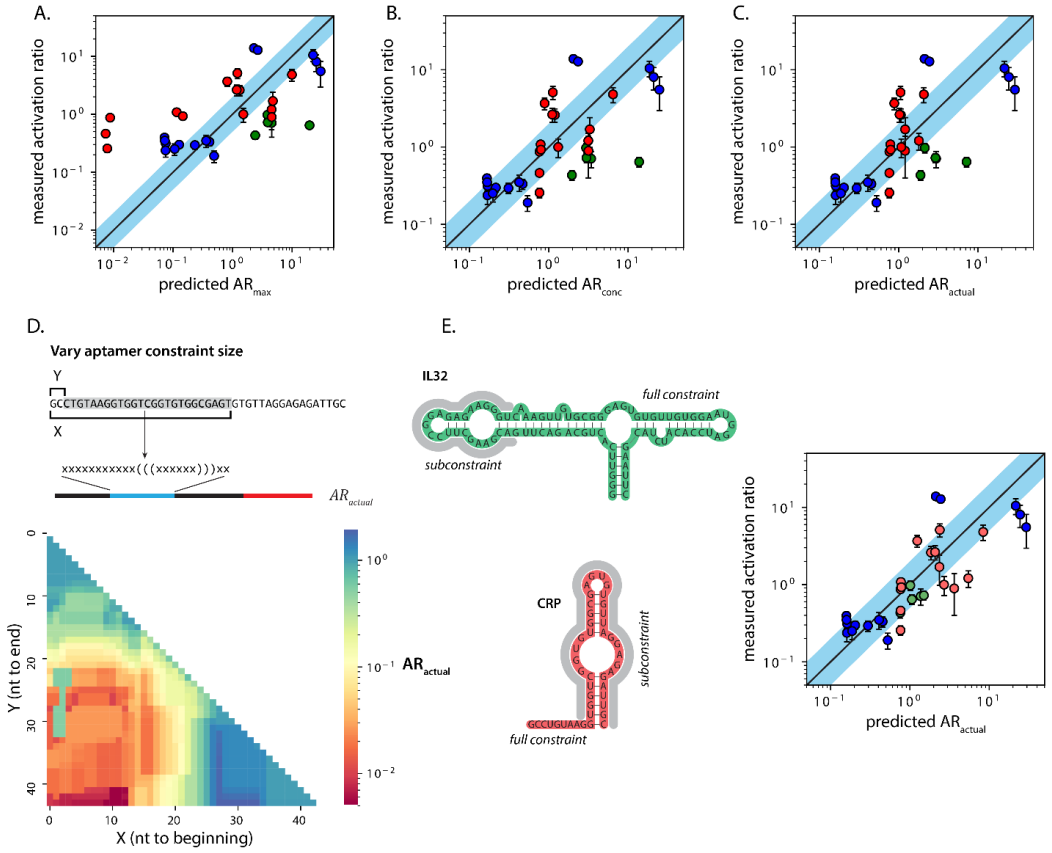

Figure 3-5: Optimizing Riboswitch Calculator predictions for *in vitro* protein riboswitches. A. Measured activation ratio vs. predicted AR_{max}. Pearson $R^2 = 0.38$, while only 14% of riboswitch ARs are within 2-fold of predicted. B. Measured activation ratio vs. predicted AR_{conc}. After calculating the concentration-dependence of riboswitch activation using a Hill function, $R^2 = 0.46$, while 38% of riboswitches activated within 2-fold of predicted. C. Measured activation ratio vs. predicted AR_{actual}. After accounting for the dissociation of mRNA:ligand complex for metastable riboswitches, $R^2 = 0.48$, while 44% of riboswitches activated within 2-fold of predicted. D. Varying aptamer constraint region changes predicted AR_{actual}. I varied the size of the constrained region for riboswitch Co1, to observe how the assumption of what sequence was interacting with CRP changed predicted riboswitch behavior. For each aptamer sub-constraint, I re-folded the sub-sequence *in silico*, and input the reduced aptamer constraint into the Riboswitch Calculator model. By varying the size of the constrained region, I observed 385-fold variation in the predicted AR_{actual}. E. Predictions of riboswitch function can be improved by rational selection of subconstraints within CRP and IL32 aptamers. I decomposed the CRP and IL32 aptamer constraints into subconstraints. The subconstraints which minimized the average log fold-change error in AR further improved model predictive accuracy, with $R^2 = 0.64$, and 55% of riboswitches predicted within 2-fold. A, B, C: Blue circles are MS2 riboswitches, red circles are CRP riboswitches, and green circles are IL32 riboswitches. Black line represents unity, while cyan shaded area represents 2-fold more or less than predicted AR. E: Blue circles are MS2 riboswitches, pink circles are CRP riboswitches with subconstraint applied, and light green circles are IL32 riboswitches with subconstraint applied. Error bars for A, B, C, and E represent the 95% confidence interval (n=6 for MS2 riboswitches, n=8 for CRP, IL32 riboswitches).

3.2.5 Improving Riboswitch Calculator model predictions

After correcting the on-state and activation ratio data for nonspecific protein-mediated repression, I next analyzed the predictive accuracy of the Riboswitch Calculator model on all of the protein riboswitches tested. I first analyzed the uninduced, off-state data, to determine whether mRFP1 expression from these sequences were well-predicted in the cell-free system, by fitting the observed mRFP1 endpoint fluorescence measurements in the induced and uninduced states to **Equation 2**. I observed that, while there was correlation between predicted ΔG_{total} and endpoint expression (R^2 = 0.47, p=5.8x10^{-6}, n=34), the range of translation, accounted for by the coefficient β, in the cell-free system was much lower than *in vivo* (0.23 vs. 0.45), a phenomenon that has previously been observed in cell-free systems (**Supplementary Figure 3**)[128, 129]. To account for this, further analysis of model predictive accuracy uses the lower β fit to the *in vitro* data. After correcting for the lower translational range in cell-free systems, the riboswitch predicted activation ratios (AR_{max}) were weakly correlated with their measured values (R^2 = 0.38, p=9.3x10^{-9}, n=34), but only 15% of the riboswitches activated expression within twofold of their predicted value (**Figure 5A, Equation 3, Supplementary Table 5**). I then calculated the concentration-dependent activation ratio (ARc_{conc}) using **Equation 4** and the previously-fit dissociation constants for each aptamer, approximating the active mCRP concentration of the second CRP batch as five-fold less than the previous batch (**Figure 5B**). This further improved model correlation (R^2 = 0.46, p=1x10^{-5}, n=34) and increased the percent performing within twofold of predicted function to 38%. Finally, I applied a correction for the dissociation of unstable mRNA:ligand complexes, further increasing the model correlation (R^2 = 0.48, p=4.7x10^{-6}, n=34), and the percent correctly predicted within twofold to 44% (**Figure 5C, Equation 6**)[36].

Equation 6 $$AR_{actual} = 1 + \frac{AR_{conc} - 1}{e^{\beta(\Delta\Delta G_{mRNA} + \Delta G_{ligand})} + 1}$$

I then turned my attention to the aptamer constraints used for CRP and IL32. Although macromolecules like proteins likely interact with more of the aptamer directly than small-molecule ligands, it is unlikely that the entirety of each aptamer (44 and 90 nt for the CRP and IL32 aptamers, respectively), interacts with the protein. For example, only 13 nt of the 19 nt MS2 operator hairpin are fully-resolved in the crystal structure of the aptamer in complex with an MS2 coat protein dimer[115]. However, the model assumes that the entirety of the supplied aptamer constraint is 'locked' into the full folded structure when the translation rate of the induced state is calculated. To examine how Riboswitch Calculator predictions varied based on what portion of the aptamer is assumed to be interacting with the protein ligand, and therefore 'locked' into its structure, I first examined the effect of systematically varying the structurally constrained region of the aptamer for several of the designed CRP and IL32 riboswitches (**Figure 5D, Supplementary Figure 4**). I then re-folded each sub-constraint, and used the sequence and structure of the subconstrained region to calculate AR_{actual} for each riboswitch so generated, without changing its sequence. I performed this analysis on several of the designed CRP and IL32 riboswitches, and observed up to a 385-fold variation in riboswitch activation by varying the subconstrained aptamer region. The predicted AR_{actual} response surface could be quite rugged, with a truncation or addition of several nucleotides to the constrained region changing the predicted riboswitch function from a translational repressor to an activator.

While the above analysis clearly demonstrates the sensitivity of predicted riboswitch function to the aptamer constraint, it does not demonstrate which constraint, of all the ones tested, is most likely to be correct for the CRP and IL32 aptamers. To examine this in more detail, I decomposed the CRP and Il32 aptamers into a series of hairpin loops, based on the full folded structure, to determine the effect of rationally selecting specific subconstraints on the predictive capacity of the model. I then examined how these rationally selected subconstraints affected the average fold-change AR (measured activation ratio/predicted AR_{actual}), $R^{2,}$ and percent of

riboswitches predicted within 2-fold (**Supplementary Table 6**). The best IL32 subconstraint reduced the fold-change AR error of the IL32 riboswitches from being underpredicted by 6.04-fold to 1.6-fold. It increased the overall R^2 for the entire dataset to 0.61 ($p=4.6x10^{-8}$, n = 34), and increased the percent predicted within 2-fold to 56%. The effect of varying the size of the CRP subconstraint was less dramatic, as the best subconstraint decreased model fold-change error from 1.31-fold overpredicted to 1.27-fold underpredicted. However, it increased the overall R^2 to 0.54 ($p=6.5x10^{-7}$, n=34), and increased the percent predicted within 2-fold to 44%. Combining the effects of these two subconstraints, the model correlation increased further ($R^2 = 0.64$, $p=1.2x10^{-8}$, n=34), and the percent predicted within 2-fold increased to 56% (**Figure 5E**). Although experimental validation would conclusively determine what region of the aptamer each protein ligand interacts with, these results suggest that these aptamer subregions could be responsible for determining riboswitch structural change and therefore function. Altogether, accounting for the decreased translational range in cell-free expression, the concentration-dependence of riboswitch activation, dissociation of ligand from unstable mRNA:ligand complexes, and the actual constrained region of each riboswitch increased R^2 from 0.32 to 0.64, and the percent predicted within 2-fold from 11% to 56%. Accounting for these factors is crucial for the accurate design of further highly-functional protein-detecting riboswitches.

3.3 Discussion

While synthetic protein sensor systems have successfully been developed from natural RNA-binding motifs, these existing methods of protein detection have been largely designed for cellular computing, and are almost entirely tied to the limited natural repertoire of protein-RNA interactions[117, 130, 131, 132, 133]. Additionally, few genetically-encoded protein sensing systems are capable of detecting exogenously-added protein[93]. By extending a previously-developed model of

translational regulation, the Riboswitch Calculator, to the development of protein sensors using synthetic RNA aptamers, and further implementing these sensors in a cell-free sensing platform, I overcome these issues to develop a highly-effective system for sensor sequence generation and validation. The further use of this method opens the door to the rapid development of inexpensive, portable protein sensors for diverse applications, including medical diagnostics or novel mechanisms of gene regulation in cellular computing.

In this study, I generated riboswitch sequences *in silico* which detect the MS2 coat protein, human modified C-reactive protein, and human IL32γ, and measured their regulation of the production of mRFP1 in a TXTL cell-free system. I identified 34 effective protein sensor sequences, 31 of which activated or repressed expression greater than 2-fold, out of 35 sequences so generated and tested. I were also able to predictably develop both translational activators and repressors, after accounting for the steric repression of expression and the extent of the aptamer constraint, with 77% of riboswitch activity correctly predicted as activating or repressing. Significantly, I extended the development of activating and repressing protein sensors past the use of natural protein recognition elements and model synthetic protein-binding aptamers, using the wealth of protein aptamer sequences that have been generated over the past several decades. At least 82 artificially generated protein-binding RNA aptamers have been developed, some for biomedically relevant protein targets like toxins, human pathogen-associated proteins, and other biomarkers[134, 135, 136, 137, 138, 139, 140, 141, 142, 143]. Aptamers for further protein targets can be generated *in vitro* via SELEX[27, 28]. While some of these aptamers have been used to detect protein *in vitro*, RNA aptamers used as sensing elements outside of a context of active gene expression are prone to chemical and biological degradation, limiting their utility in the detection of protein in complex media[144, 145, 146, 147].

I investigated some of the key determinants of *in vitro* protein riboswitch function. I demonstrated that, *in vitro*, the sensitivity of a translational riboswitch is similar to the sensitivity

of the aptamer used to design the riboswitch. Previously-designed synthetic riboswitches have often been assayed *in vivo*, making the determination of actual riboswitch sensitivity difficult, due to differentials in intracellular and extracellular concentration of many small molecules and ions[148, 149]. I also demonstrated that some protein ligands are capable of operating as translational repressors in the absence of any secondary structural change, even when located quite distally from the ribosomal footprint. While proteins like the global regulator csrA, or a number of phage proteins, are capable of modulating gene expression by either competing with the ribosome for binding directly or by altering the secondary structure of the ribosome binding site, it is unexpected that CRP binding far upstream of the ribosomal footprint would result in translational repression[150, 151]. While mCRP has a clear repressive effect, it is unclear why it represses while the MS2 coat protein, which binds as a dimer of similar molecular weight to mCRP (MS2 coat protein dimer MW = 27.4 kDa, mCRP MW = 23 kDa), does not[152, 153]. I also demonstrate that, in order to accurately design maximally responsive sensors, knowledge of the exact binding site is essential. This information is known for some aptamers, through crystallographic, mutagenic, or other sequence-structure-function studies, but not for all. However, I demonstrate that the estimated binding region can also be inferred from existing riboswitch data, informing future sensor design.

In addition to demonstrating the use of synthetic, nonmodel protein aptamers in genetically-encoded sensors, I show that the design of highly effective RNA secondary structural switches can be efficiently automated. Classes of RNA switch beyond the translational riboswitches I characterized here include allosteric ribozymes, turn-on fluorescent sensors, and ligand-inducible Cas9 sgRNAs[154, 155, 156, 157, 158, 159]. While several protein-detecting RNA switches have previously been developed, computationally designing these sensor sequences using an extended range of protein-binding aptamers could further extend their utility as sensors and signal transduction elements[160]. I demonstrate the utility of computational nucleic acid sensor design to generate a handful of sensor sequences from a combinatorically-large sequence space, further decreasing the

time and resources needed to generate effective protein sensors from previously-developed protein-binding aptamers.

3.4 Conclusions

Cell-free expression systems are powerful sensor platforms, but a generalizable approach to cell-free protein detection has not yet been demonstrated. I extended a previously-developed model of translational riboswitch regulation (the Riboswitch Calculator), previously validated only with small-molecule detection, to design 34 highly-functional riboswitch sensors for MS2 coat protein, human monomeric C-reactive protein, and human interleukin-32γ. Overall, I demonstrated the utility of *in silico* sensor sequence design and *in vitro* testing for the rapid and efficient design and test of potential sensor sequences. Finally, by incorporating the effects of reduced cell-free translational tunability, direct repression of some proteins, and the effect of supplied aptamer constraint on the model's predictions, I greatly improved the model's predictive accuracy for the further design of protein-detecting riboswitches in cell-free expression systems.

3.5 Methods

3.5.1 Construction of riboswitch and protein-expressing plasmids

To construct riboswitch and CDS plasmids for this study, I started with the pFTV vector backbone, which contains mRFP1 modified to contain an N-terminal SacI restriction site[109]. I used the Riboswitch Calculator to design candidate riboswitch sequences, and the Operon Calculator to codon-optimize the MS2 coat protein CDS and design an optimal RBS sequence (**Supplementary Table 7**)[36, 161]. I designed and ordered gBlocks, containing primer binding sites and additional

restriction sites, and PCR primers for both the riboswitches and MS2 coat protein CDS (Integrated DNA Technologies). I PCR amplified the gBlocks using Phusion or Q5 DNA polymerase (New England Biolabs). For the riboswitches, I digested the riboswitch amplicons and pFTV vector backbone with XbaI and SacI-HF (New England Biolabs). For the MS2 coat protein CDS, I digested the CDS amplicon and pFTV vector backbone with XbaI and NotI-HF (New England Biolabs). For both the riboswitches and CDS, I ligated the digested inserts with digested backbone using T4 DNA ligase (New England Biolabs) and heat-shock transformed the ligation product into chemically competent *E. coli* DH10B. I then performed Sanger sequencing to verify that the insert had been cloned correctly.

3.5.2 Crude cell lysate preparation

Crude cell lysate was prepared according to a previously published protocol, with the following modifications[17]. 20 L of Escherichia coli BL21 with the Rosetta2 plasmid encoding rare tRNAs was cultured in a Micros 30-liter fermentor (New Brunswick) in 2XYT+P medium until the cells reached an OD600 of 1.5-2.0. The cell pellet was then collected in a T-1-P Laboratory continuous flow centrifuge (Sharples), and resuspended in 1 mL S30A buffer per gram of cell pellet. The resuspended cells were run through a M110-EH-30 microfluidizer (Microfluidics Corp.) at 20,000 PSI twice to ensure complete lysis. The lysate was clarified by centrifugation at 12,000xg for 30 minutes at 4C. The clarified lysate was then incubated for 80 minutes at 37 minutes while undergoing orbital shaking to perform the runoff reaction. After incubation, the lysate was centrifuged again for at 12,000xg for 30 minutes at 4°C.

Following lysis, clarification, and the runoff reaction, the lysate was diafiltered with a Pellicon Biomax 10 kDa MWCO 0.005 m^2 ultrafiltration module. Six retentate volumes of buffer S30B were run against the lysate at 4°C. After diafiltration, the retentate was centrifuged for 30

minutes at 12,000xg at 4°C. The protein concentration of the retentate was quantified using a Bradford BSA Protein Assay Kit assay (Bio-Rad). The retentate was aliquoted and flash-frozen in liquid nitrogen, and stored at -80°C.

3.5.3 Cell-free expression reactions

Cell-free expression reactions were assembled on ice according to previously published protocols, with the following modifications[17, 80]. Amino acid and energy solutions were prepared separately, and combined with crude cell extract to reach the following final concentrations: 7.4 mg/mL protein (1/3rd total reaction volume), 1.5 mM each amino acid (except for leucine at 1.25 mM), 50 mM HEPES, 1.5 mM ATP and GTP, 0.9 mM CTP and UTP, 0.2 mg/ml tRNA, 0.26 mM CoA, 0.33 mM NAD, 0.75 mM cAMP, 0.068 mM folinic acid, 1 mM putrescine, and 30 mM PEP. 4 mM additional magnesium glutamate *(8.67 mM total), 80 mM additional potassium glutamate (100 mM total), and 2% w/v PEG-8000 were added to each reaction. Plasmid DNA was either miniprepped and ethanol precipitated, or midiprepped and isopropanol precipitated, and added to the reaction to a final concentration of 2 nM. 5 uL reactions were incubated at 29°C for 16 hours in a 96-well polypropylene conical bottom plate sealed with a plate storage mat (Corning) in a TECAN Spark microplate reader. mRFP1 fluorescence was measured every 10 minutes, using 584nm/60nm ex/em with a 5 nm bandwidth.

3.5.4 Endpoint mRFP1 data analysis

The endpoint was taken as the average of the last approximately 15 mRFP1 fluorescence data points of each reaction, during which active mRFP1 production had ceased. The following corrections were applied: the background for each reaction was calculated as the average of the

first 15 data points, approximately, of each reaction, prior to the onset of mature mRFP1 accumulation. I observed that, even in the absence of plasmid DNA, there was a slight increase in the fluorescence between the beginning and end of each reaction, so each reaction was also corrected for the nonspecific fluorescence increase by performing the same endpoint fluorescent calculation as above on a no-DNA reaction, and subtracting that nonspecific fluorescence increase.

3.5.5 Design of protein-detecting riboswitch sequences using the Riboswitch Calculator

I obtained the sequence and binding affinity for each aptamer used in this study from the literature. Aptamer secondary structures were determined using RNAfold, using the Turner 2004 nearest-neighbor parameter set, with no dangling end free energies[35]. To design the riboswitches, I used the design mode of the Riboswitch Calculator model of translation initiation regulation, which builds on the RBS Calculator v2.1 model to predict translation initiation states of each riboswitch in the uninduced and induced states[36, 60]. From the output list of riboswitch sequences, I subselected sequences based on their predicted maximum activation ratios, and on- and off-state expression.

3.5.6 Alternate constraint analysis

To perform the alternate constraint analysis on the selected riboswitches, I iterated through every possible aptamer substring and re-calculated the predicted AR_{actual} for each substring. Briefly, for subconstraints i (first subconstrained nucleotide position) and j (last subconstrained nucleotide position) I re-folded the aptamer subconstraint from position i to j using RNAfold as described above. I then appended the sequence from the beginning of the full aptamer constraint to i to the existing pre-aptamer sequence to generate a new pre-aptamer sequence, and prepended the sequence from the end of the subconstraint to the end of the full aptamer sequence to the existing

post-aptamer sequence to generate a new post-aptamer sequence. With the new pre-aptamer, post-aptamer, and aptamer sequences and structural constraint, I then calculated the predicted AR_{actual} at maximum induction using the Riboswitch Calculator.

3.5.7 Statistical analysis

For pairwise comparisons, I used 2-tailed, two-sample t-tests. For correlations, I used the Pearson R^2. For all tests, the significance level was set to $\alpha = 0.05$.

Chapter 4

Highly-mutated sgRNAs guide Cas9 cleavage similarly to dCas9 repression

4.1 Introduction

Engineered CRISPR-based systems have been harnessed to bind, edit, and cut genomic DNA at specified locations with increasing precision, versatility, and portability[162, 163, 164, 165]. Many CRISPR applications require targeting many, distinct genomic locations simultaneously, for example, to modify several SNPs at once to treat a genetic disease[166, 167], to regulate many endogenous genes at the same time to trigger changes in cell state[168], or to study complex gene regulatory networks or polygenic diseases[169, 170]. Targeting each distinct genomic location requires the expression of an additional crRNA or single-guide RNA (sgRNA)[169]. Multiple guide RNAs have been co-expressed for gene regulation, however these systems contained several long DNA repeats in both the sgRNAs and the genetic parts used to express them[171, 172, 173, 174]. Genetic systems with repetitive DNA sequences are more difficult to assemble *in vitro*[175] and create genetic instability *in vivo*, particularly in organisms and viral vectors with more frequent homologous recombination[176, 177]. Several studies have noted that repetitive DNA sequences can trigger spontaneous deletions that result in a loss of function[178, 179, 180, 181, 182], particularly when selective phenotypes are introduced. Enabling simultaneous and stable co-expression of many sgRNAs would therefore create a new class of CRISPR-based applications providing access to a much larger phenotypic space.

The sgRNA handle sequence has a defined structure, and previous work has demonstrated that mutations that preserve the broader structure have minimal impact on the sgRNA's ability to direct Cas9 activity[183, 184, 185, 186]. Following this design principle, we designed, built, and

characterized 54 nonrepetitive, highly-mutated sgRNA handle sequences, and successfully built three extra-long sgRNA arrays (ELSAs) which coexpressed up to 28 sgRNAs to overproduce succinic acid, regulate cellular amino acid biosynthesis, and regulate the cellular stress response. These nonrepetitive sgRNA sequences were highly mutated from the wild-type sequence, sharing down to 62% nucleotide identity with the original sequence, and we discovered additional nucleotide constraints required for the efficient direction of dCas9 activity in CRISPR transcriptional interference (CRISRPi) assays. While we demonstrated that these 28 of these nonrepetitive handles were effective at directing dCas9, the catalytically inactive version of the enzyme, it was unclear whether the mutations present in these sequences would prevent the final DNA cleavage step that the catalytically-active Cas9 performs.

To examine whether the sgRNA sequence and structural constraints we learned from dCas9 activity assays held for directing Cas9 activity, I re-characterized 25 sequences from the original nonrepetitive sgRNA handle toolbox in *in vitro* Cas9 DNA cleavage assays. The sequences were chosen to either vary in the degree of mutation to key functional regions, or lack mutations to those regions but have a very high degree of mutation overall. I then examined whether the non-functional mutated sgRNAs failed to bind to Cas9, using electrophoretic mobility shift assays, as a possible explanation for their lack of (d)Cas9 activity. I also investigated whether off-target activity differed between dCas9 and Cas9, and whether mutated sgRNA handles had a different degree of off-target activity to the wild-type sgRNA sequence. Altogether, I confirm that these nonrepetitive sgRNA handles are similarly able to direct dCas9 DNA binding and Cas9 DNA cleavage, validating their use in further applications requiring nuclease activity.

4.1.1 Hypotheses

1. sgRNA handle mutations in previously identified key locations will reduce the sgRNA's ability to direct Cas9 DNA cleavage, versus dCas9 DNA binding.

2. The overall extent of sgRNA handle mutation will reduce the sgRNA's ability to direct Cas9 cleavage, versus dCas9 DNA binding.

3. Nonfunctional sgRNAs are unable to form a specific sgRNA:Cas9 complex.

4. Highly-functional sgRNA handles will have similar off-target activity in both Cas9 and dCas9, versus the wild-type sgRNA handle sequence.

4.2 Results

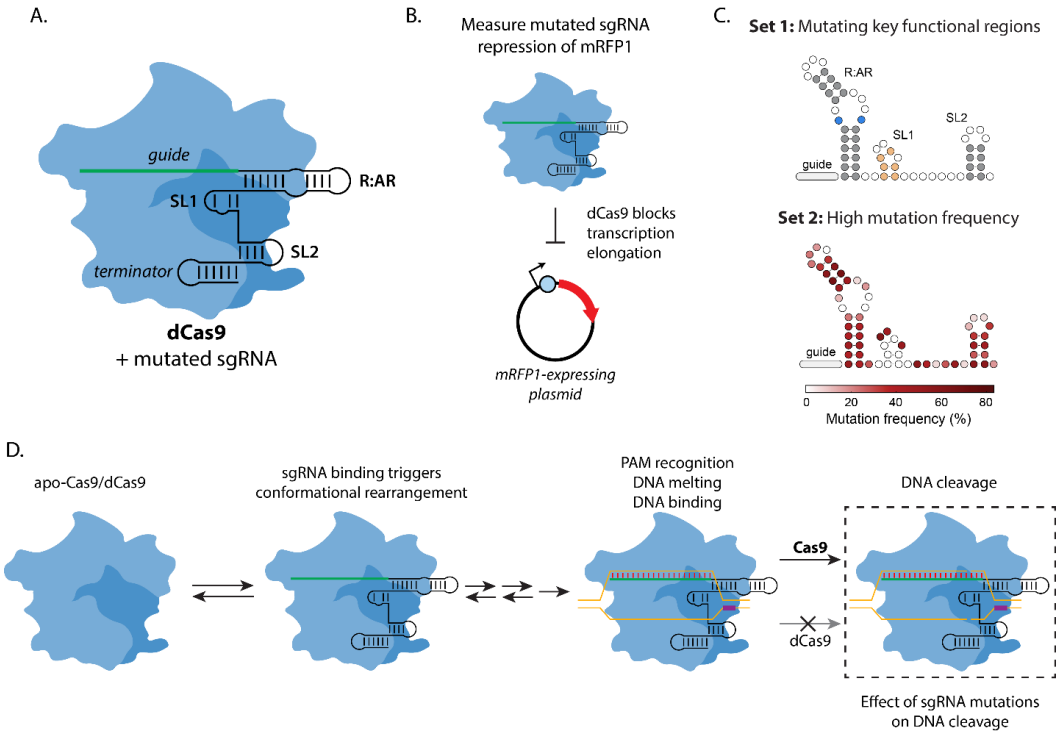

Figure 4-1: Effect of mutations on sgRNA's ability to direct (d)Cas9 activity. (A) Schematic of dCas9 in complex with sgRNA, including key functional regions of RNA, labeled as follows: guide, repeat:antirepeat (R:AR), stem loop 1 (SL1), stem loop 2 (SL2) , and terminator. (B) Schematic of genetic system measuring mutated sgRNA:dCas9 repression of mRFP1 expression. dCas9:sgRNA binds to the target DNA sequence and inhibits RNA polymerase transcription elongation. (C) Goals of mutations in selected sgRNAs. Set 1 was selected to have mutations in two key functional regions: the bulge region of the R:AR (key mutated nucleotides in blue), or SL1 (key mutated nucleotides in orange), or to mutate non-key regions (gray and white nucleotides). Set 2 was selected to not mutate key functional nucleotides, but to increase the overall degree of mutation by mutating base pairs and non-key nucleotides in loops. (D) Schematic of dCas9 vs. Cas9 function. sgRNA binding to (d)Cas9 triggers a conformational rearrangement of the protein, and (d)Cas9:sgRNA complex binding to DNA is a multi-step process involving recognition of an NGG PAM sequence, melting of the duplex DNA and hybridization of the guide sequence to the target strand, and a final irreversible DNA binding step. Following the irreversible DNA binding step, dCas9 remains bound, but Cas9 follows with an irreversible DNA cleavage step. I sought to understand whether sgRNA mutations that allowed dCas9 DNA binding had a differential effect on Cas9 binding and DNA cleavage.

4.2.1 Nonrepetitive, mutated sgRNAs direct dCas9 repression of gene expression with varying efficiency

CRISPR-Cas9 is guided by an sgRNA, which is composed of a 20nt guide sequence, a 61-nt handle sequence, and a transcriptional terminator[187]. The guide sequence is variable, designed by the end user to target a sequence of choice for Cas9 cleavage or dCas9 binding. The handle sequence folds into a defined structure, composed of several key structural motifs: the repeat:antirepeat region (R:AR), stem loop 1 (SL1), and stem loop 2 (SL2). This sequence is then recognized and bound by (d)Cas9 (**Figure 1A**). We initially hypothesized that the key determinant of sgRNA loading into Cas9 and direction of its binding and nuclease activity was determined by the sgRNA's structure. Through three rounds of design-build-test-learn cycles, we designed 54 nonrepetitive, highly-mutated sgRNA sequences. We screened them for activity in a dCas9-based repression assay, where dCas9 was directed by the mutated sgRNA handles to block the transcription of *mRFP1* RNA and repress its expression (**Figure 1B**)[188]. Through these assays, we determined that 28 of the designed handle sequences were highly functional, even with

mutation frequencies up to 38%. Surprisingly, we did not find any correlation between the average degree of mutation and binding activity. However, through linear discirminant analysis (LDA), we determined that several specific key nucleotides in the bulge region of the R:AR and the hairpin of SL1 had a significant effect on dCas9 function.

While these assays allowed us to determine the role of both key functional regions and the overall degree of mutation from the wild-type sequence on dCas9 activity, this did not necessarily imply that the same design rules applied to mutated sgRNA-directed Cas9 cleavage. dCas9, compared to the wild-type Cas9 enzyme, has a mutation in each of its two nuclease domains that completely abrogate its nuclease activity, leaving it to irreversibly bind DNA without cleaving it. Prior to DNA cleavage, dCas9 and Cas9 go through the same multi-step pathway of sgRNA binding, DNA recognition and melting, and irreversible DNA binding[189]. The conformational shifts of the enzyme in this process are controlled by key sgRNA and target DNA sequence recognition events[185, 190, 191]. It is therefore possible that, while dCas9 and Cas9 are similarly affected by sgRNA mutations in their DNA binding steps, Cas9 could be differentially effected by sgRNA mutations during the final cleavage step (**Figure 1D**).

To investigate the role of specific sgRNA functional regions and the role of overall sequence divergence from the wild-type sgRNA, I selected two subsets of mutated sgRNAs from our original set that had been screened in dCas9 (**Figure 1C**). Set 1 consisted of sgRNAs with a lower average degree of mutation (13%), but with mutations in two of the LDA-identified key functional regions. For each of the functional region mutations, I chose a highly-functional and a low-functioning sgRNA handle, to determine whether the functional region mutation was more or less detrimental in Cas9 versus dCas9. I also selected low-, medium-, and high-functioning sgRNA handles without any mutations to key functional regions to determine whether the rank-order of activity was preserved in Cas9. The second set (Set 2) of sgRNAs were highly mutated

(by an average of 28%), but lacked mutations in LDA-identified specific essential functional regions.

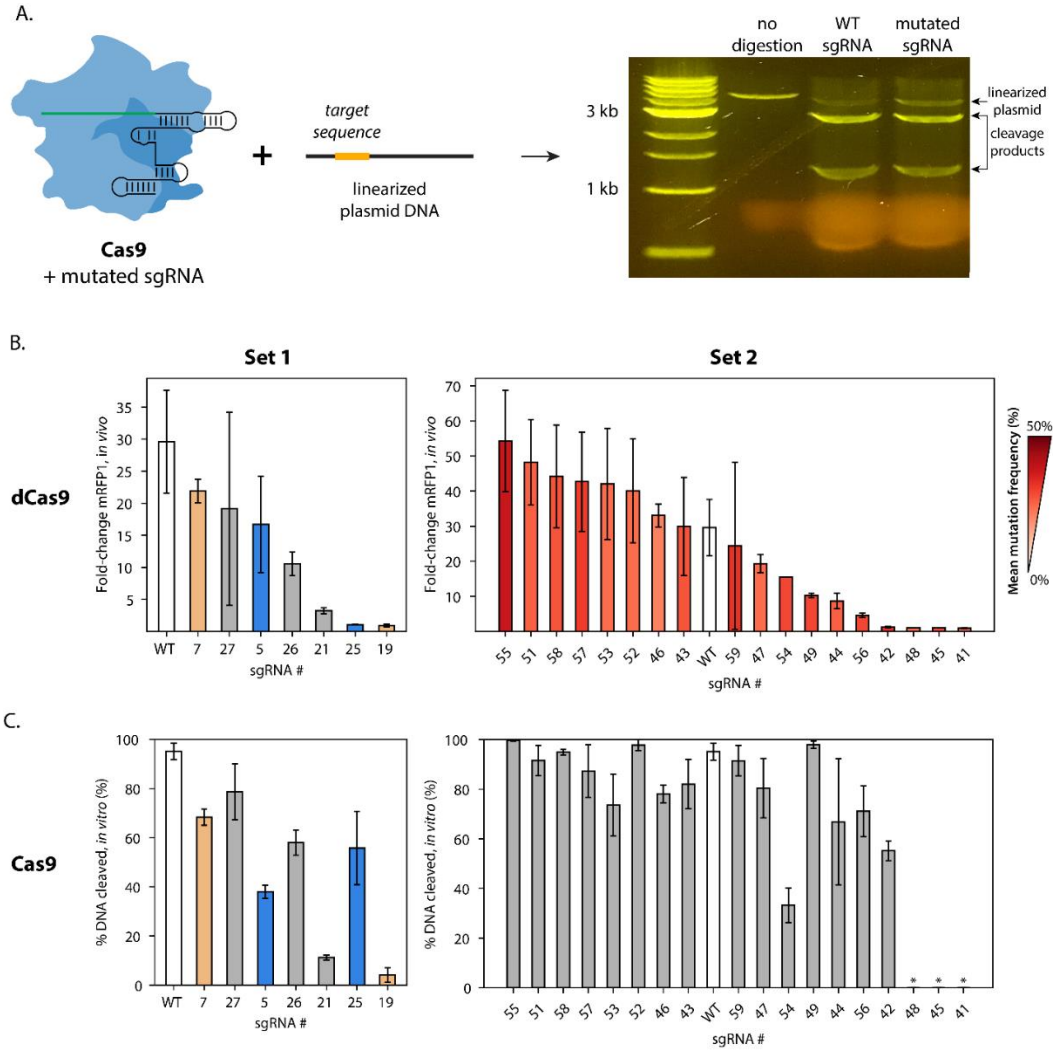

Figure 4-2: Effect of sgRNA mutations on Cas9-mediated DNA cleavage efficiency. (A) DNA cleavage assay schematic. Cas9 is loaded with a mutated sgRNA, and its cleavage efficiency against a linearized plasmid DNA target is measured. (B) Fold-change repression of Set 1 and Set 2 sgRNAs, with dCas9. For Set 1 (left), colors represent mutations in R:AR (blue), SL1 (orange), or other (gray). WTsgRNA repression is shown in white. Error bars represent the 95% confidence interval of N=3 biological replicates. For Set 2 (right), degree of mutation from WTsgRNA is shown by red colorbar (far right). WTsgRNA is shown in white. Error bars represent the 95% confidence interval of N=3 biological replicates. (B) Cleavage efficiency of Set 1 and Set 2

sgRNAs. For Set 1 (left), colors represent mutations in R:AR (blue), SL1 (orange), or other (gray). WTsgRNA repression is shown in white. Error bars represent the 95% confidence interval of N=2 biological replicates. For Set 2 (right), mutated sgRNAs are shown in gray, while the WTsgRNA is shown in white. Error bars represent the 95% confidence interval of N=2 biological replicates. (D) DNA cleavage kinetics of WT (white), highly-functional (sgRNA 55, blue), and moderately-functional (sgRNA 54, red) sgRNAs. Circles represent mean and 95% confidence interval of N=2 biological replicates. (E) Mutated sgRNA activity in dCas9 repression vs. Cas9 cleavage rate. Blue circles represent mean fold-change mRFP1 vs. mean DNA cleavage rate. Error bars represent the 95% confidence interval of N=3 biological replicates for fold-change repression, and N=2 biological replicates for DNA cleavage rate.

4.2.2 Mutated sgRNAs are similarly able to direct dCas9 repression of expression and Cas9 DNA cleavage

I next tested Set 1 and Set 2 sgRNAs in an *in vitro* Cas9 cleavage assay, to compare the performance of sgRNAs in directing dCas9 and Cas9 activity. Briefly, Cas9 and each mutated sgRNA were combined, and then a linearized target DNA sequence corresponding to each sgRNA guide sequence was added. The degree of cleavage was observed by performing fragment analysis on cleavage products, visualized via agarose gel electrophoresis (**Figure 2A**).

I found that the ability of mutated sgRNAs to direct dCas9 repression broadly corresponded with their ability to direct Cas9 cleavage, for both Set 1 and Set 2 sgRNAs (**Figure 2B, 2C**). Almost all of the sgRNAs that had previously been determined to be high-functioning in dCas9 also efficiently directed Cas9 cleavage of their DNA targets, with an average of 81% cleaved over a 15-minute digestion. However, the lower-functioning sgRNAs largely either inefficiently directed Cas9 cleavage or had no detectable activity in the assay, even after extended digestion (**Supplemental Figure 1**). Examining the specific sequence determinants selected in the Set 1 sgRNAs, I observed that the relative activity of the high- and low-functioning SL1 mutants, as well as the three nonspecifically-mutated sequences in Set 1, was preserved in the Cas9 cleavage assays. Unexpectedly, however, both of the bulge mutants had moderately high activity in Cas9, with the ‘high-functioning’ sgRNA 5 cleaving 38% of its substrate, and the

'low-functioning' sgRNA 25 cleaving 56% of its substrate. While this specific type of mutation appeared to have a differential effect in dCas9 and Cas9, I found that the Set 1 and Set 2 sgRNAs had no correlation between the degree of mutation and Cas9 DNA cleavage activity (**Supplemental Figure 2**).

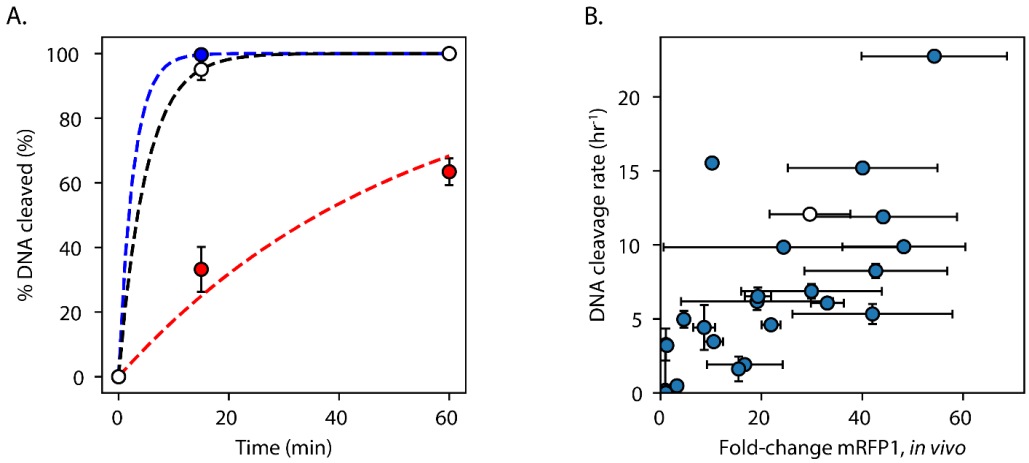

Figure 4-3: Cas9 DNA cleavage kinetics vs. dCas9 fold-change repression. (A) DNA cleavage kinetics of WT (white), highly-functional (sgRNA 55, blue), and moderately-functional (sgRNA 54, red) sgRNAs. Circles represent mean and 95% confidence interval of N=2 biological replicates. Dotted lines represent best-fit of first-order rate equations. (B) Mutated sgRNA activity in dCas9 repression vs. Cas9 cleavage rate. Blue circles represent mean fold-change mRFP1 vs. mean DNA cleavage rate. WTsgRNA is shown in white. Error bars represent the 95% confidence interval of N=3 biological replicates for fold-change repression, and N=2 biological replicates for DNA cleavage rate.

While the high-functioning sgRNAs were very active in both dCas9 and Cas9-based assays, I observed that, for both Set 1 and Set 2, the moderately-functioning sgRNAs had 2-fold higher activity, relative to the wild-type sgRNA, in the cleavage assays versus the repression assays. In order to better compare the overall performance of our mutated sgRNAs in dCas9 repression and in Cas9 cleavage assays, I analyzed the kinetics of Cas9 cleavage over the course of an hour for several of the tested sgRNA handles (**Figure 3A**). Under the conditions of excess Cas9:sgRNA relative to target DNA found in the assay, Cas9 cleavage of its DNA target can be

described by a pseudo first-order rate equation[192]. I fit this rate equation to time-course digestion reactions for both the wild-type sgRNA handle and a high- and moderately-functional mutated sgRNA handle. I next extracted rate constants for each sgRNA I tested, using the cleavage percent at 15 minutes:

$$k_{cleave} = \frac{\ln(1-\%cleaved)}{0.25\ hr} \qquad \textbf{Equation 1}$$

I found the cleavage rates to fall within the range of 0 to 23 hr^{-1}. Multiple studies have found a correlation between RNA:DNA hybridization energy and Cas9 activity, stemming from a dependence of the cleavage rate on the Cas9 DNA binding rate, while other studies have found that structure in the guide sequence of the RNA also effects Cas9 activity[193, 194, 195, 196] Extending this analysis to the favorability of Cas9:sgRNA activity, I assume that the formation of the Cas9:sgRNA:DNA ternary complex is the rate-limiting step. The rate constant for cleavage will then be proportional to the free energy of ternary complex formation (ΔG_{TC}):

$$k_{cleave} \propto e^{-\beta \Delta G_{TC}} \qquad \textbf{Equation 2}$$

Similarly, assuming that binding is the rate-limiting step in dCas9 transcriptional repression, then the maximum fold-change repression is also proportional to ΔG_{TC}:

$$FC_{repression} = 1 + Ke^{-\beta \Delta G_{TC}} \qquad \textbf{Equation 3}$$

Based on these assumptions, the fold-change repression with dCas9 and cleavage rate constant with Cas9 should therefore be proportional. I found that the rate constant and fold-change repression were well-correlated ($R^2 = 0.54$, $p = 2x10^{-5}$), further indicating that, although the magnitudes of some key sgRNA:Cas9 interactions may be different, the general controllers of mutated sgRNA activity in dCas9 and Cas9 are similar (**Figure 3B**).

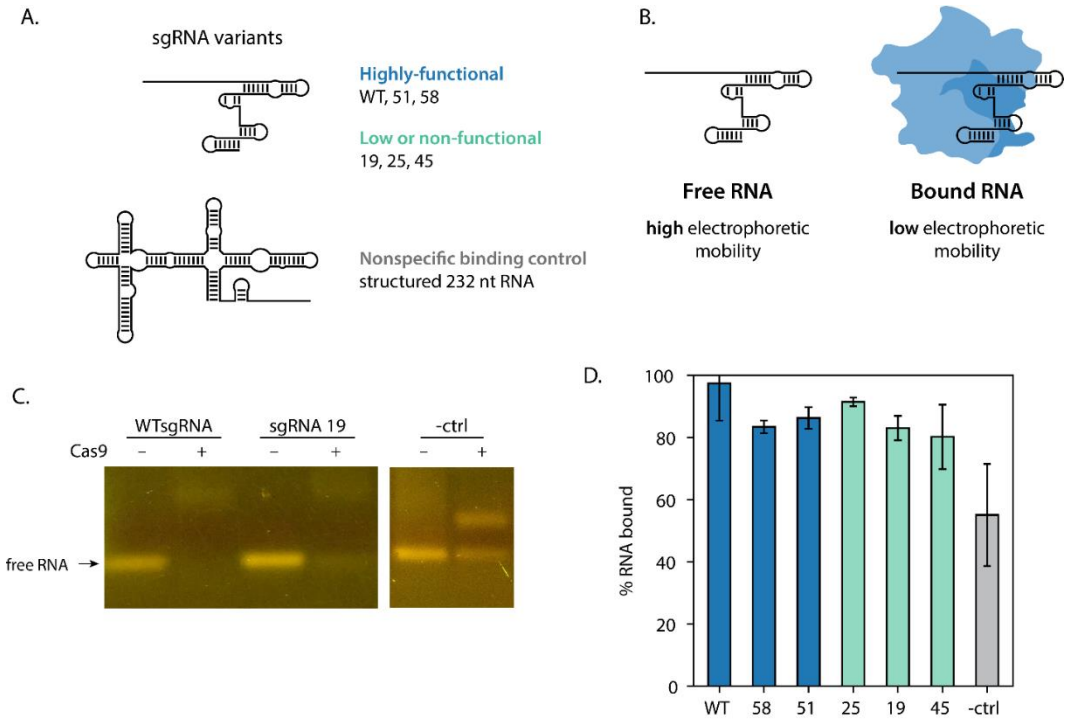

Figure 4-4: Electrophoretic mobility of RNA:Cas9 complexes. (A) Types of RNAs used. I tested three highly-functional sgRNAs (WT, 51, 58), three low-functioning sgRNAs (19, 25, 45), and one nonspecific binding control. (B) Electrophoretic mobility of free RNA is higher than Cas9-complexed RNA. (C) Representative gel results of high-functioning, low-functioning, and non-specific RNA, showing change in intensity of free RNA band. (D) Percent bound of each RNA. Highly functional sgRNAs are in blue, low-functional sgRNAs are in green, and the nonspecific binding control is in gray. Error bars represent the 95% confidence interval of N=2 biological replicates.

4.2.3 Highly-functional and non-functional sgRNAs bind Cas9 with similar efficiency

While the Cas9 and dCas9 assays revealed that the nonfunctional sgRNA handles failed to direct (d)Cas9 to bind DNA, it was unclear whether this was due to an inability of the sgRNA to bind to (d)Cas9, or a failure of the sgRNA to activate the required conformational changes to a DNA-binding competent structure. To investigate this question, I used electrophoretic mobility shift assays (EMSAs) to determine the degree of Cas9:sgRNA binary complex formation for

seven RNA sequences (**Figure 4B, 4C**). I determined the degree of binding between Cas9 and RNA for the wild-type sgRNA handle, two highly-functional sgRNA handles, two low-functioning sgRNA handles, and one handle with no observable activity in dCas9 or Cas9. I also screened a highly-structured 232-nt nonspecific binding control RNA (**Figure 4A**). I observed that each tested sgRNA was bound by Cas9 to a similar extent, with an average of 89% of the high-functioning sgRNAs bound to Cas9, and an average of 84% of the low functioning sgRNAs binding (**Figure 4D**). 55% of the nonspecific binding control also was bound by Cas9. Cas9 is a nucleic acid binding protein, with several surface regions that are highly positively charged, and we therefore might expect that even a nucleic acid that lacks specific binding motifs would bind to some extent[184]. However, each sgRNA, even the non-functional sgRNAs, bound with greater efficiency than the non-specific binding control, suggesting that each tested sgRNA is able to interact with the higher-affinity sgRNA binding site, versus nonspecifically associating with the surface of the protein.

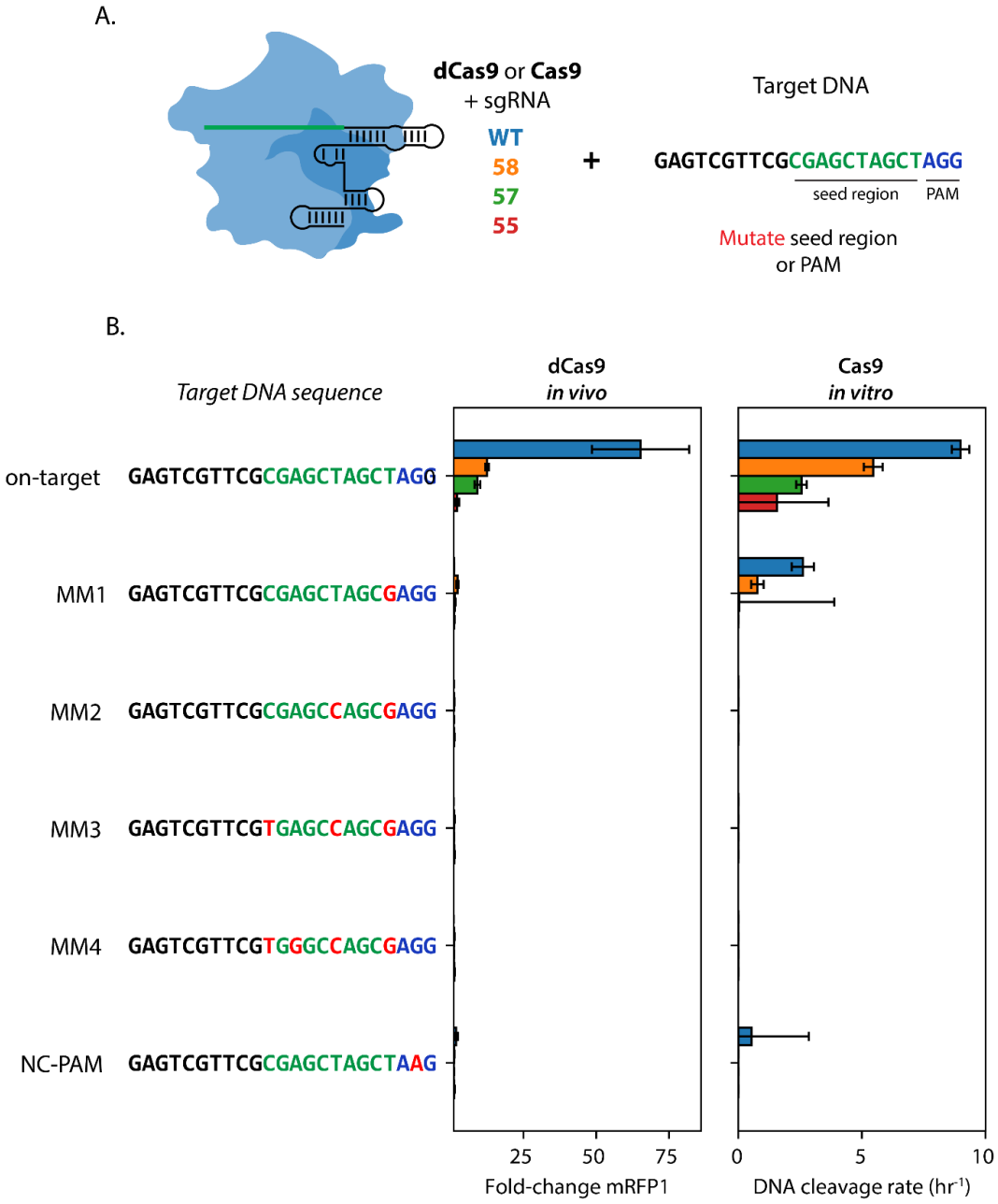

Figure 4-5: Off-target activity of highly-functional, mutated sgRNA handles in dCas9 and Cas9. (A) Test system, consisting of measuring WTsgRNA or mutated, highly-functional sgRNA handle performance against off-target sequences. Off-target activity of both dCas9 (fold-change repression) and Cas9 (cleavage efficiency) was measured. Off-target DNA sequences either had mutations in the seed region or the PAM sequence. (B) Activity of WT (blue bars) and mutated sgRNAs (58 – orange, 57 – green, 55 – red), against on-target and 5 off-target sequences (far left). dCas9 fold-change repression is shown on the left, while the corresponding Cas9 cleavage rates are shown on the right. Mutations to target sequence are shown in red. Error bars represent 95% confidence interval of N=4 biological replicates (dCas9) or N=2 replicates (Cas9).

4.2.4 Mutated sgRNAs have low off-target activity in dCas9 repression and Cas9 cleavage

Finally, I investigated whether the off-target activity of the mutated sgRNAs was significantly different from the wild-type handle sequence, in both Cas9 and dCas9. Off-target activity, due to partial DNA sequence complementarity with the sgRNA guide sequence, can lead to unintentional mutations, toxicity, and cell death[197, 198, 199]. In order to examine the off-target effects of the mutated sgRNAs, I constructed sgRNA sequences consisting of the same guide sequence with 4 different, highly-functional sgRNA handles: wild-type, sgRNA 58, sgRNA 55, and sgRNA 57 (**Figure 5A**). I then assayed their ability to direct dCas9 repression or Cas9 cleavage against an on-target or an off-target sequence. The five off-target sequences contain either increasing numbers of mutations in the seed region of the target, where sequence mismatches are known to have the greatest effect, or a non-canonical PAM sequence. The predicted cutting frequency determination (CFD) scores of the off-target activity against the one-mismatch (MM1) and non-canonical PAM (NC-PAM) sequences was relatively high, at 18% and 26%, respectively, while the 2, 3, and 4-nt seed sequence mismatches were predicted to have minimal off-target activity[200].

The on-target activity was much greater than the off-target activity for each sgRNA sequence, across both the dCas9 and Cas9 test systems (**Figure 5B, 5C**). Surprisingly, I observed that the wild-type sgRNA had the highest efficiency in directing on-target dCas9 repression, 65-fold, while the highly-functional sgRNAs were significantly less efficient, from 12-fold to 2.2-fold. A similar pattern was observed for activity directing Cas9 cleavage, where the cleavage rate with the wild-type sgRNA was 9.0 hr^{-1}, while the rates of the nonrepetitive sgRNAs ranged from 5.5 hr^{-1-} to 1.6 hr^{-1}. Significant off-target activity was only observed for the MM1 and NC-PAM sequences, in line with predictions, although only for wild-type sgRNA and sgRNA 58. There was less than 1.5-fold off-target activity against the off-target sequences with more than one

mismatch, for dCas9, and no detectable Cas9 cleavage activity. Notably, the degree of on-target and off-target activity appeared to correlate, with the two highest-activity sgRNAs also exhibiting the highest degree of off-target activity.

4.3 Discussion

CRISPR-Cas9 systems have revolutionized the engineering of both genomes and gene regulation. However, it is still difficult to build more complex systems where many sgRNAs are simultaneously expressed, due to repetitive sequence-induced difficulties in gene synthesis and *in vivo* genetic stability. To overcome this problem, we designed 54 nonrepetitive, highly-mutated sgRNA handles and validated their function with dCas9, finding 28 highly-active sequences. However, in order to maximize the utility of these handles, we needed to demonstrate that they retained their activity across the different contexts in which CRISPR:Cas9-systems have been applied, including the catalytically dead dCas9 and catalytically active Cas9. I validated that the nonrepetitive sgRNA handles previously screened in dCas9 similarly guide Cas9 function in *in vitro* cleavage assays.

I also investigated whether either mutations to key functional regions or the overall degree of mutation affected the ability of Cas9 to cleave its DNA substrate. I found no correlation between cleavage efficiency and degree of mutation, reflecting the *in vivo*, dCas9 results. We had previously found that mutations in several key regions negatively affected dCas9 function on average, although did not totally abrogate dCas9 activity for some sequences. While mutations to SL1 appear to have the same effect in both dCas9 and Cas9, mutations to the bulge region in the R:AR of the sgRNA handle, while having a strong effect in dCas9, appeared to not have a significant effect in Cas9. As we might expect that Cas9, which must cleave its substrate in addition to binding it, would have more stringent requirements for sgRNA handle composition

than dCas9, this could be due to the differences in reaction environment between the *in vivo* dCas9 repression assays and the *in vitro* Cas9 cleavage reactions. I further demonstrated that, by directly correlating the cleavage rates and fold-change repression for each sgRNA, the activity of the handle in Cas9 can be generally well-predicted by its activity in dCas9. Overall, I demonstrate that the sequence and structural constraints learned from dCas9-based sgRNA screening are sufficient to generate highly-active mutated sgRNA handles, allowing for the construction of large genetic systems utilizing either enzyme.

When performing the cellular dCas9 assays, it was unclear whether the failure mode of the nonfunctional sgRNAs was due to an inability of dCas9 to bind to the mutated sgRNA, or whether the dCas9:sgRNA binary complex was unable to effectively bind DNA. I confirmed that the nonfunctional sgRNAs are able to bind to Cas9 to a similar degree as the wild-type sgRNA, and to a greater degree than a noncognate, structured RNA sequence, suggesting that binary complex formation is taking place. Cas9 is able to strongly bind even sgRNA sequence motifs that are unable to direct Cas9 activity, such as truncated fragments of sgRNAs and sgRNAs with high degrees of guide structure, suggesting that as long as enough structure is preserved, strong and specific binding can take place[185, 196, 201]. However, while these nonfunctional sgRNAs strongly bind Cas9, they fail to activate the series of allosteric structural changes in Cas9 that control DNA binding and cleavage, giving a possible mechanism for the failure of the nonfunctional sgRNAs I tested here.

I finally investigated whether the highly-functional sgRNAs had similar degree of off-target activity to the wild-type sgRNA, in both dCas9 and Cas9. In the original *in vivo* testing, each sgRNA handle was paired with a unique guide, which can itself lead to differential activity, even with the same sgRNA handle[202, 203]. To remove this confounder, I paired several highly-active sgRNA handles with the same guide sequence. Surprisingly, I found that the degree of off-target activity roughly corresponded with the degree of on-target activity, finding that the highly-

functional nonrepetitive sgRNAs had, on average, lower on- and off-target activity than the wild-type handle sequence. While slightly differing from the previously-determined activity for each sgRNA handle, this corresponds with prior work showing that modifications to the sgRNA or Cas9 itself which lower off-target activity by lowering the free energy of the ternary complex, therefore also lowering on-target activity to some extent[204, 205, 206, 207].

Through my validation of nonrepetitive sgRNA sequence design rules across different variants of the Cas9 enzyme, I expand the utility of mutated sgRNA sequences to cleavage-based applications, including genome editing and biosensing. The success of developing sequence- and structural-based design rules for the *S. pyogenes* Cas9 sgRNA sequence used here also suggests that similar sequence-structure-function relationships can be determined for the many other Cas systems currently in use, which may be more optimal for many other applications than the original SpCas9.

4.4 Conclusions

The utility of CRISPR-Cas9 systems is limited by the inability to effectively multiplex sgRNA expression, due to synthesis issues and genetic instability. To overcome this challenge, we designed highly-mutated, nonrepetitive sgRNA handle sequences to allow for the synthesis and cellular expression of extra-large sgRNA arrays (ELSAs). However, the initial validation of these mutated handles used a catalytically inactive version of the Cas9 nuclease, dCas9. To determine how both mutations in specific key functional regions and the overall extent of mutation affected the final DNA cleavage step in wild-type Cas9, I subselected mutated sgRNAs to determine their function in *in vitro* Cas9 cleavage assays. I found that the activity of these mutated sgRNA handles in directing dCas9 repression and Cas9 cleavage was similar, indicating that the mutated handles can function similarly in applications requiring either enzyme. I further

examined the mechanism of failure of the nonfunctional sgRNA mutants, and determined that the nonfunctional sgRNA handles were able to complex efficiently with Cas9. Finally, I considered the off-target activity of the highest-functioning sgRNA mutants, compared to the wild-type sgRNA handle sequence. I found that, while the mutants had lower off-target activity than the wild-type sequence, they had correspondingly lower on-target activity.

4.5 Methods

4.5.1 Cloning the sgRNA handle test system.

Unless stated otherwise, *Escherichia coli* K-12 DH10B (Thermo Fisher Scientific) was used for cloning. The non-repetitive sgRNA handles were synthesized as 3-sgRNA arrays on either pUC19 cloning vectors (Genscript) or as gBlock gene fragments (Integrated DNA Technologies or IDT). An existing 3-plasmid test system including pAN-PBAD-sgRNA-A2T (ColE1), pAN-PTet-dCas9 (p15A), and pAN-PA2-RFP (pSC101) was provided by the Voigt lab (Addgene). The sgRNA-expressing plasmids (ColE1) were assembled using ligase cycling reaction (LCR)[208]. Briefly, the sgRNAs and plasmid backbone were PCR amplified with Phusion DNA polymerase (NEB), 5’ phosphates were added via T4 polynucleotide kinase (NEB), and 60 nucleotide oligos were used to mediate blunt-ended ligation using Taq ligase (NEB), resulting in scarless insertion of the sgRNAs downstream of the Ara-pBAD promoter. The mRFP1-expressing target plasmid (pAN-PA2-RFP) was modified to introduce an EcoRI cut site downstream of the constitutive PA2 promoter. The resulting target plasmid was restriction digested with NheI and EcoRI, and oligonucleotides were annealed and inserted into the backbone using T4 DNA ligase (NEB) for each unique target sequence.

4.5.2 Characterization of the non-repetitive sgRNA handles.

Escherichia coli BW27783 (CGSC 12119)[209] was used for characterizing the sgRNA handles to ensure strong induction of the pBAD promoter. The BW27783 cells were chemically co-transformed with pAN-PTet-dCas9, and the modified pAN-PBAD-sgRNA and pAN-PA2-RFP plasmids, and plated on ampicillin, kanamycin, and spectinomycin plates. Picked colonies (n = 3) were used to inoculate 700 μL LB cultures, and were grown at 37°C for 9 hours in a shaker incubator. Subsequently, 5 μL of cells were diluted into 195 μL M9 minimal media with 0.4% glycerol, appropriate antibiotics, 20 mM arabinose (Sigma Aldrich), and 1.25 ng/mL anhydrous tetracycline (aTc) (or no inducers for the uninduced condition) in 96-well microplates. The cells were incubated at 37°C for 5 hours and the OD600 and mRFP1 fluorescence (Ex. 584 nm, Em. 607 nm) was recorded using a TECAN M1000 Infinite plate reader. At the end of the 5-hour growth, cells were mid-exponential phase, and a second identical dilution was done. The second plate was incubated for 12 hours. At the end of the 12-hour culture period, all cells were mid-exponential phase, 20-40 μL of the cell culture was diluted into 200 μL 1X PBS with 2 mg/mL kanamycin for flow cytometry. Flow cytometry measurements and analysis were performed same as before for promoter characterization.

4.5.4 *In vitro* Cas9 cleavage assay.

Linear amplicons were constructed, consisting of the T7 promoter, 2 guanosine residues to promote efficient transcription initiation by T7 RNA polymerase (5'-AAGCTAATACGACTCACTATAGG-3', transcription start site underlined), and the sgRNA guide and handle. The sgRNAs were transcribed using a HiScribe™ T7 High Yield RNA Synthesis Kit (NEB), and purified via phenol:chloroform extraction followed by ethanol precipitation. Gel

electrophoresis was used to confirm the integrity of each sgRNA transcript. Each sgRNA was resuspended to 300 nM in 1X TE buffer, and annealed to renature the RNA by heating to 95°C for 5 minutes and cooling at 0.2°C increments per minute to 25°C. The modified pAN-PA2-RFP vectors used for *in vivo* characterization were used as the target DNA sequence for the *in vitro* Cas9 cleavage assay. The plasmid vector was linearized by digesting with NdeI (NEB) for 6 hours. *In vitro* Cas9 cleavage reactions were performed on this linearized target DNA using purified *S. pyogenes* Cas9 nuclease (NEB). Equimolar sgRNA and Cas9 (30 nM) were incubated in 1X NEBuffer 3.1 (NEB, 100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 100 μg/mL BSA, pH 7.9) in a total volume of 30 μL for 10 minutes at 25°C to facilitate sgRNA loading. 3 nM of the corresponding linearized target DNA was subsequently added, and each reaction was incubated for 15 minutes at 37°C. After digesting with Cas9, I added 1 μL of Proteinase K (NEB) to each reaction, and incubated for 10 minutes at room temperature. I visualized the digestion products of each reaction by running on a 1X TBE, 1% agarose (SeaKem LE, Lonza), 1X GelStar (Lonza) gel. Digital photographs were taken of the gels using a blue light trans-illuminator with an orange filter, and the intensities of the digested product bands were quantified using GelAnalyzer to determine the degree of digestion. For each of the two cleaved bands, the following formula was used to determine the percent cleavage.

$$\% \text{ cleavage} = \frac{\frac{I_n}{len_n}}{\frac{I_n}{len_n} + \frac{I_{ND}}{len_{ND}}}$$

I_n is the intensity of a given product band, I_{ND} is the intensity of the uncleaved plasmid band, and len_n and len_{ND} are the lengths of the given product band (2979 or 1379 bp) and the uncleaved plasmid band (4358 bp), respectively. I reported the cleavage efficiency of each Cas9 cleavage reaction as the average of the cleavage efficiencies, determined using both product bands, across two independent replicates.

4.5.5 Electrophoretic mobility shift assay.

Electrophoretic mobility shift assays (EMSAs) were performed to measure the equilibrium formation of sgRNA:Cas9 binary complex (RNP). sgRNAs were produced using *in vitro* transcription. Briefly, linear DNA templates were constructed combining a T7 promoter, guide RNA sequence, and a selected non-repetitive sgRNA handle. sgRNAs were transcribed using the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB), and purified using phenol:chloroform extraction and ethanol precipitation. Following synthesis and confirmation of transcript integrity via agarose gel electrophoresis, sgRNAs were re-folded at a concentration of 300 nM in 1X TE buffer. Binding assays were performed with 30 nM sgRNA, with or without 30 nM Cas9, in 1X NEBuffer 3.1 buffer (NEB, 100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 100 μg/mL BSA, pH 7.9) in a total volume of 30 μL. Reactions were incubated at 25°C for 10 minutes, followed by 37°C for 15 minutes. sgRNA bands were visualized, with and without added Cas9, by running each reaction on a gel containing 1X TAE, 1% agarose (SeaKem LE, Lonza), and 1X GelStar fluorescent dye (Lonza). Digital photographs were taken of gels using a blue light trans-illuminator with an orange filter. Fluorescent band intensities were quantified using GelAnalyzer to determine the amount of unbound sgRNA. The percent complex formation was calculated as the following:

$$\% \text{ Cas9:sgRNA complex} = 1 - \frac{I_A}{I_B}$$

I_A is the intensity of the free sgRNA band when incubated with 30 nM of Cas9. I_B is the intensity of the free sgRNA band when no Cas9 is present during incubation.

Chapter 5

Direct, Cas9-based detection of SARS-CoV2 RNA using engineered tracrRNAs

5.1 Introduction

CRISPR-Cas systems enable the sequence-specific recognition and digestion of DNA or RNA target sequences. While their utility has been proven in both genome editing and gene regulation applications, their ability for specific-sequence cleavage has increasingly been harnessed for the detection of specific nucleic acid targets, including pathogen-associated RNA and DNA[163, 166, 167, 174, 188, 210, 211, 212]. Current CRISPR molecular diagnostics, using Cas12a and Cas13 nucleases, rely on the unique property of these nucleases to enable collateral off-target cleavage after binding and cleaving on-target, pathogen-derived nucleic acids[211, 213, 214]. These platforms combine the rapid signal generation inherent in isothermal nucleic acid amplification techniques with an additional sequence recognition step, increasing test specificity and accuracy. However, these platforms do not allow for either direct RNA detection or efficient, one-pot multiplex detection of many different nucleic acids simultaneously.

Multiplex nucleic acid detection has great utility in medical diagnostics, maximizing the information per test, while potentially decreasing costs time-to-diagnosis[215, 216, 217]. Many natural CRISPR systems, including the foundational and well-studied *S. pyogenes* CRISPR-Cas9 system, have been proven to operate in a highly multiplexed manner in both natural and engineered systems[171, 172, 173, 218, 219, 220, 221]. However, Cas9 is only able to directly bind to and cleave target DNA, and not RNA[187]. Additionally, the target DNA sequence being both the input analyte and output signal limits the ability to couple DNA cleavage to alternate output methods and further

multiplexing, by limiting the ways the target sequence can be modified to allow for different mechanisms of signal readout.

Here, I re-purpose the dual-RNA-guided Cas9 system, utilizing a crRNA and tracrRNA to direct Cas9 activity, for direct RNA detection. I used previously determined sequence-structure-function rules for single-RNA-guided Cas9 systems to re-engineer tracrRNAs to bind to any target RNA and direct Cas9 activity to a non-analyte DNA target, and validated 17 functional SARS-CoV2 N-gene detecting tracrRNA sequences[221]. I further investigated ways to streamline the detection protocol, and identified requirements for *in vitro* processing of target RNA-derived crRNAs. Finally, I coupled N-gene directed Cas9 cleavage to a previously-developed Cas9-based isothermal nucleic acid amplification technique (CRISDA) to enable the development of efficient RNA detection using an engineered Cas9 platform[222].

5.1.1 Hypotheses

1. The dual-RNA-guided CRISPR-Cas9 system, with the exception of specific nucleotides in the bulge region, is insensitive to structure-conserving mutations in the repeat:antirepeat.

2. By designing tracrRNAs that hybridize with the correct structure, portions of the SARS-CoV2 N-gene can be used as crRNAs.

3. Both 5’ and 3’-end N-gene derived crRNA processing is required for Cas9 activity using N-gene derived crRNAs.

4. N-gene directed Cas9 DNA cleavage can be used to trigger strand displacement amplification.

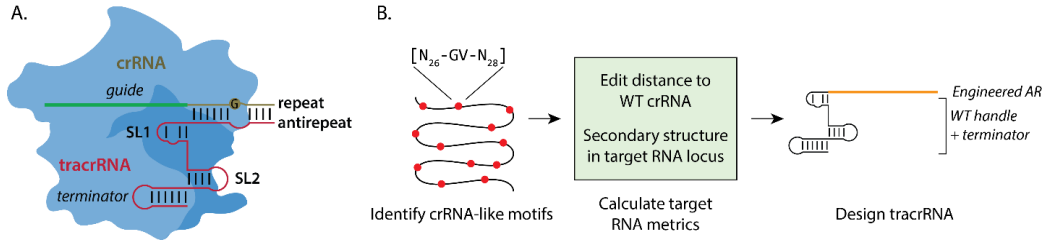

Figure 5-1: Engineering tracrRNAs to design a dual-guided Cas9 RNA detection system. (A) Dual-guided Cas9 requires three separate molecules: a crRNA (brown) consisting of guide and repeat regions; tracrRNA (red) consisting of an antirepeat, stem-loop 1 (SL1), stem-loop 2 (SL2), and a transcriptional terminator; and the Cas9 protein. The sequence of the repeat:antirepeat duplex can be highly mutated and retain its function, and in the repeat region, only G27 was found to contribute significantly to function. (B) Designing target RNA-detecting tracrRNAs. All instances of the degenerate crRNA motif were identified in the SARS-CoV2 N-gene. The edit distance to the wild-type crRNA and the degree of secondary structure in the target RNA region were then determined. Finally, a cognate tracrRNA was designed for each potential crRNA sequence in the target RNA.

5.2 Results

5.2.1 Computational design of RNA-detecting tracrRNAs in dual-guided Cas9 system

The wild-type *S. pyogenes* CRISPR system consists of three separate molecules: crRNA, tracrRNA, and Cas9[187]. The crRNA sequence consists of a 20 nt guide sequence and a 36-nt repeat, which base-pairs with the antirepeat sequence at the 5' end of the tracrRNA (**Figure 1A**). The crRNA-tracrRNA complex, or dual-guide RNA (dgRNA), is then bound by Cas9, and directs its nuclease activity towards a target DNA sequence corresponding to the guide RNA sequence. Importantly, Cas9's recognition of the dgRNA is largely directed by its structural, and not sequence, features. This phenomenon has been exploited to design mutated sgRNAs for multiplex gene regulation or allosteric activation[156, 221, 223]. Although some nucleotides are essential for efficient direction of Cas9 activity, in the repeat:antirepeat duplex, previous work has found that only G27 and U44 in the bulge region were found to contribute significantly to Cas9 activity[221].

The fact that Cas9 can recognize and bind to the dgRNA via a largely sequence-agnostic mechanism can be exploited to re-engineer Cas9 as an RNA-detecting platform. Instead of designing the guide region of a dg/sgRNA to target a specific DNA sequence, the antirepeat region of a tracrRNA can be designed to bind to a target RNA and cofold into a Cas9 binding-competent structure. The Cas9:dgRNA complex is then directed to cut a specific DNA sequence, also specified by the target RNA sequence. I harnessed this re-engineered RNA detection platform to design and characterize tracrRNAs capable of detecting different loci in the SARS-CoV2 N-gene.

To design N-gene detecting tracrRNAs, I first scanned the N-gene for motifs matching a previously-determined sequence constraint for the sgRNA repeat region (**Figure 1B**). I initially identified 275 potential crRNA sequences within the N-gene. I then ranked the candidate target-derived crRNAs by two metrics: the edit distance of the repeat to the wild-type crRNA repeat, and the degree of intramolecular structure within the putative target crRNA locus. Firstly, my goal was to develop an isothermal workflow to detect the presence of viral RNA with designed tracrRNAs, reducing equipment requirements. I therefore selected target crRNAs that had at least 9 single-stranded nucleotides in the repeat region, to enable efficient isothermal strand displacement and annealing of the N-gene repeat with the corresponding tracrRNA antirepeat. Secondly, anticipating that Cas9 cleavage would be most efficient with a repeat:antirepeat most similar to the wild-type sequence, I selected crRNAs that had an edit distance of no greater than 11 to the wild-type crRNA repeat. I finally designed tracrRNA sequences by re-designing the tracrRNA antirepeat to co-fold into the correct structure with the N-gene crRNA repeat, with the rest of the tracrRNA sequence being unchanged from the wild-type. After filtering for low edit distance and relatively high repeat single-strandedness, I subselected 20 candidate tracrRNA sequences.

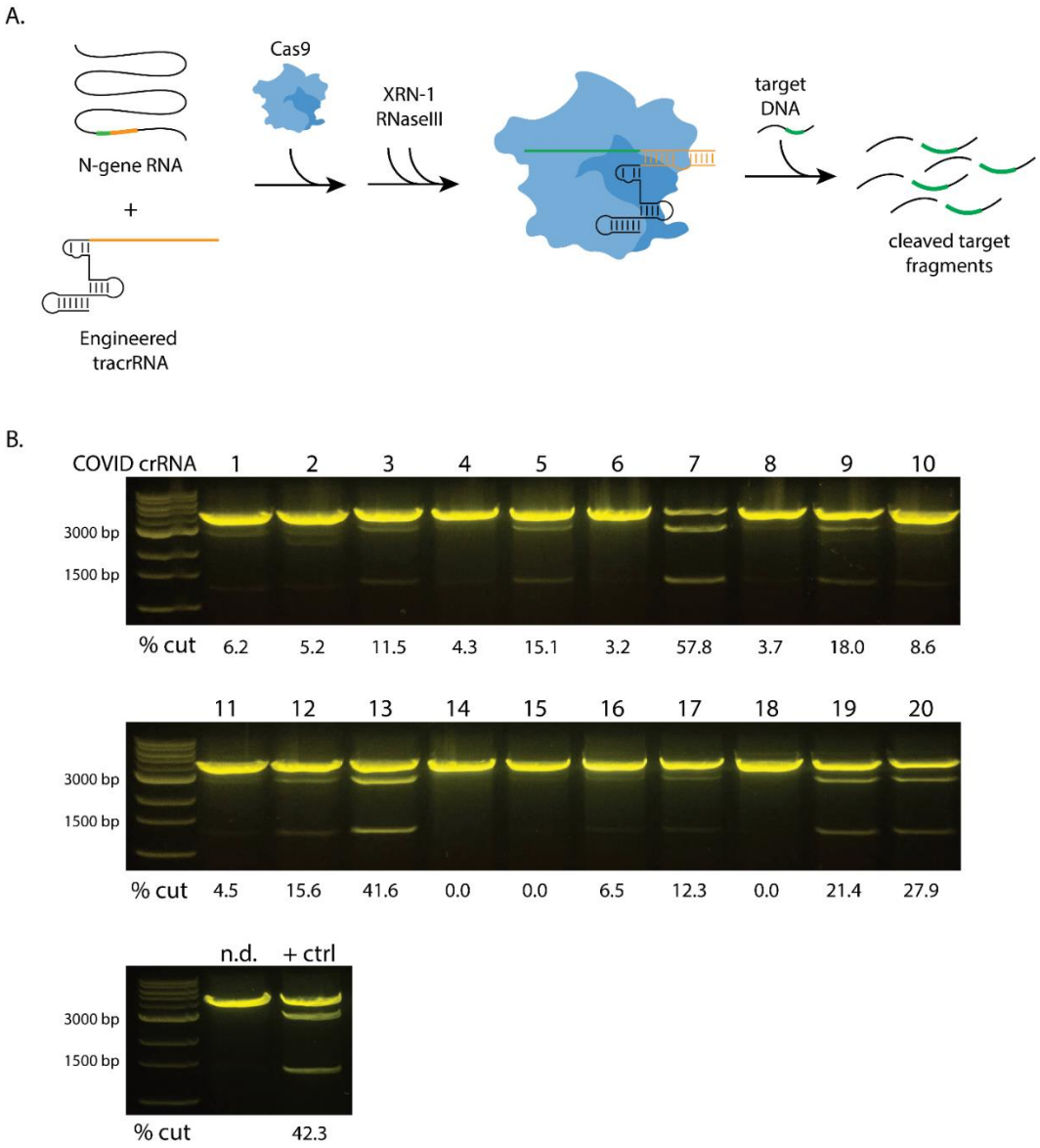

Figure 5-2: Testing N-gene detecting tracrRNAs. (A) Schematic of N-gene guided Cas9 cleavage. An engineered tracrRNA is combined with N-gene RNA. Subsequently, Cas9 is added, and the Cas9-RNA complex is then processed with XRN-1 and RNaseIII. The activity of the processed Cas9-RNA complex is assayed by measuring cleavage activity of a target RNA. (B) Cas9 cleavage activity of designed tracrRNAs combined with N-gene RNA against target DNA. The molecular-weight corrected relative intensities of the uncleaved DNA band (4358 bp), compared to the two cleaved product bands (2979 and 1379 bp), were used to calculate the % cleaved for each reaction shown.

5.2.2 Initial validation of engineered tracrRNAs that detect SARS-CoV2 N-gene

Briefly, to screen these initially-designed N-gene detecting tracrRNAs, I combined each tracrRNA with N-gene, and incubated to allow isothermal strand displacement to occur. To the tracrRNA:N-gene complex, I added Cas9, to load the assembled complex into the nuclease. I then added two RNases, XRN-1 and RNaseIII, and finally added linearized target DNA to observe Cas9 cleavage efficiency (**Figure 2A**). RNaseIII, an endonuclease that specifically binds to and cleaves RNA duplexes longer than 20 bp, is known to be essential for 3' end crRNA processing *in vivo*, cleaving the crRNA:tracrRNA duplex from the larger pre-crRNA transcript[224]. The exact enzyme or enzymes responsible for processing the 5' end of the crRNA are not known, but 5'->3' exonucleases are assumed to participate[224, 225]. I first observed that the addition of these enzymes did not disrupt dual-guided Cas9 activity using a short, wild-type crRNA and wild-type tracrRNA (**Supplementary Figure 1**).

I then tested the full set of 20 tracrRNAs to determine which were functional in directing activity, and whether either edit distance from the wild-type repeat:antirepeat sequence or structure in the target crRNA affected subsequent Cas9 activity. I found that 17 of the 20 designed tracrRNAs were able to generate cleaved product DNA bands (**Figure 2B**). The average cleavage efficiency was 13%, ranging from 0% to 58%. However, I found that there was no significant correlation between Cas9 cleavage activity and either the number of single-stranded nucleotides in the target crRNA repeat or the edit distance between the wild-type repeat and target repeat (multiple linear regression: $R^2 = 0.073$, $p = 0.88$ and $p = 0.30$, respectively). While the overall efficiency of both the wild-type dgRNA and designed N-gene:tracrRNA complexes was lower than I had observed in previous sgRNA cleavage assays, I attributed this to the extended incubation period prior to the DNA cleavage step that was introduced in this protocol.

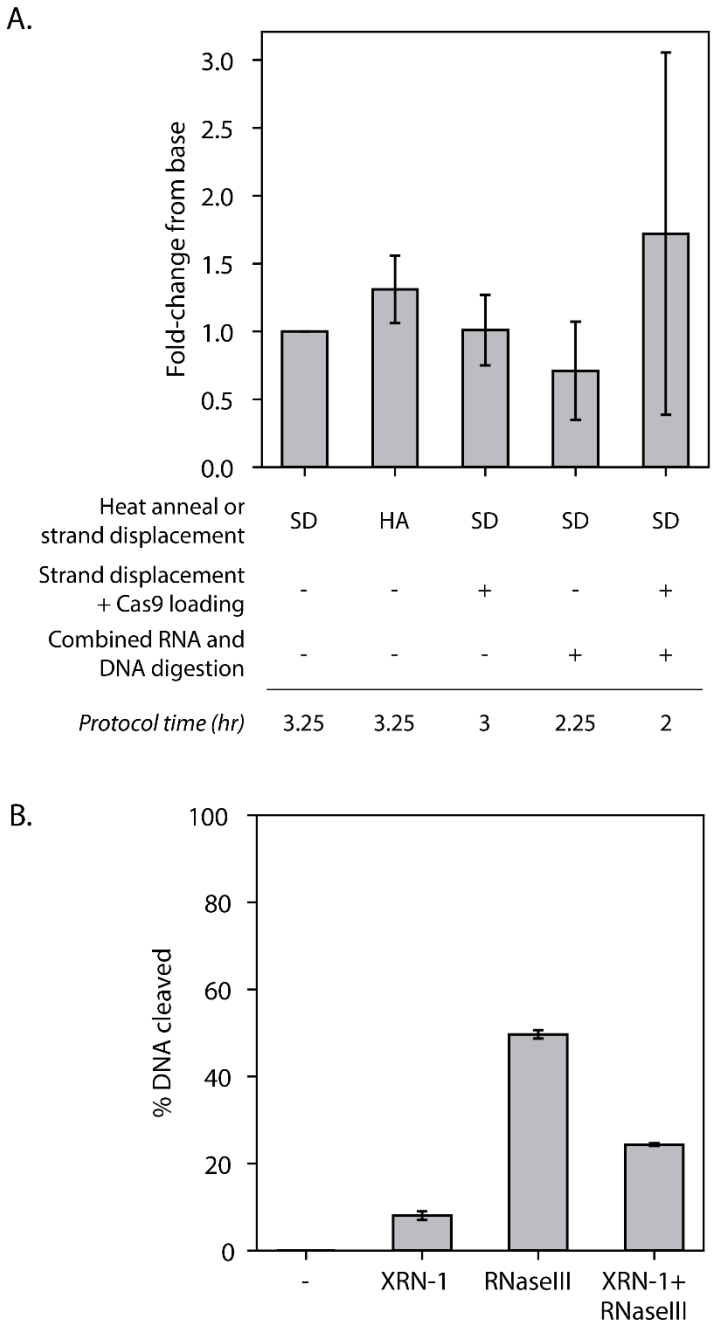

Figure 5-3: Protocol optimization. (A) Relative effects of different protocol optimizations using three different tracrRNAs. I compared the effect of varying the N-gene-tracrRNA annealing process, combining isothermal strand displacement and Cas9 loading, and combining the RNase and DNA target digestion steps. Ialso compared the total reaction times of each protocol variation. The bars and error bars shown are mean and 95% confidence interval of the relative changes in efficiency for each protocol variant, averaged across the three tracrRNAs tested (7, 9, and 15). (B)

Effect of RNase treatment on Cas9 activity with tracr7. Bars and error bars shown are the mean and 95% confidence interval of 2 independent replicates.

5.2.3 *In vitro* Cas9 cleavage protocol optimization

To optimize the protocol, it would be desirable to both increase Cas9 cleavage efficiency and decrease the amount of time and processing steps involved. I first investigated whether certain steps were able to be combined or substituted to streamline the protocol or increase cleavage efficiency. Firstly, I tested the effect of three different tracrRNA-N-gene annealing methods. The original protocol involved a 1-hour strand displacement, followed by a 15-minute Cas9:RNA loading step. Instead of isothermal strand displacement, I attempted to heat-anneal the two complementary RNA species together. I also attempted to combine the isothermal strand-displacement and Cas9 loading steps, as Cas9 contributes to strand displacement and final complex assembly efficiency[224, 226]. I next measured the effect of combining the two nuclease steps – RNase processing of the Cas9:RNA complex, and Cas9 digestion of target DNA. I measured the relative effect of these modifications on an extremely high-functioning tracrRNA, a moderate-functioning tracrRNA, and an extremely low-functioning tracrRNA (tracrRNAs 7, 9, and 15, respectively). I found no significant effect of any of these modifications to the protocol on the output cleavage efficiency, suggesting that the time-saving optimizations I attempted can be implemented without negatively impacting N-gene detecting efficiency (**Figure 4A**).

To reduce the enzymatic requirements for N-gene detection, I next decided to investigate which, if any, of the RNases used in the original workflow were required for efficient N-gene detection. I tested the effect of varying RNase treatments on the highest-functioning tracrRNA, tracr7. While it is known that RNaseIII digestion is required to cleave the 3’ end of the crRNA repeat and enable Cas9 activity, it has also previously been demonstrated that 5’ extensions to the

crRNA guide sequence also negatively impact Cas9 activity[224, 227]. I found that omitting the two nucleases entirely resulted in an absence of any observable DNA cleavage. Surprisingly, I found that either nuclease by itself was sufficient to promote some degree of Cas9 cleavage, although RNaseIII was much more efficient. Additionally, the combined XRN-1 and RNaseIII nuclease treatment resulted in 2-fold lower subsequent Cas9 activity than treating with RNaseIII alone. Interestingly, I had previously found that XRN-1 was not able to digest N-gene RNA, as the enzyme is not able to initiate degradation from the 5'-triphosphate at the 5' end of T7 transcripts (**Supplementary Figure 2**). It is possible the extended incubation at 37 °C during the strand displacement, Cas9 loading, and nuclease treatment degraded enough of the N-gene RNA to allow XRN-1 to initiate at the revealed 5' monophosphates it requires to initiate digestion.

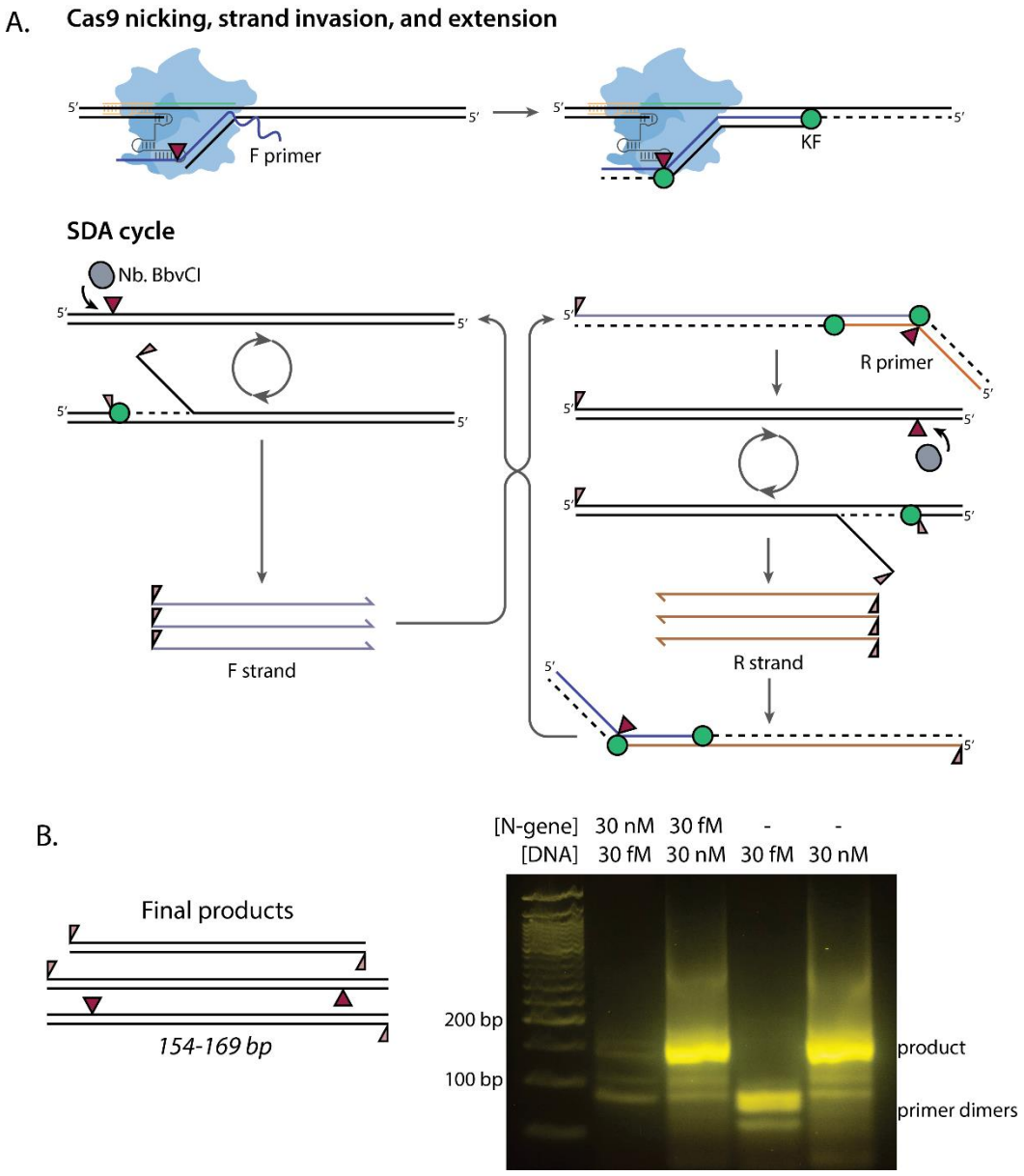

Figure 5-4: N-gene directed, Cas9-mediated strand-displacement amplification (CRISDA). (A) CRISDA schematic. N-gene directed Cas9 first cleaves a double-stranded DNA target. The F primer then binds the nontarget strand, and Klenow fragment (KF, green circle) elongates to form an intermediate DNA product. Nb.BbvCI nickase (gray oval) binds and cuts at its recognition site (purple triangle), KF extends from the nick, and Nb.BbvCI nicks again at the reconstituted site, producing F strand. R primer is then able to bind to F strand, KF extends, and R strand is then similarly produced. F primer then binds to R strand, continuing the strand displacement amplification (SDA) cycle. (B) Initial results from N-gene directed CRISDA. Successful CRISDA results in the accumulation of amplicons ranging from 154-169 BP.

5.2.4 N-gene directed Cas9-triggered strand displacement amplification

While I was able to achieve detection of 30 nM SARS CoV-2 N-gene RNA using a simple fragment analysis-based assay, real-world conditions necessitate the detection of target RNA concentrations down to 1 copy/uL or lower[228]. To facilitate the development of a protocol capable of ultrasensitive SARS-CoV2 detection, I modified a Cas9-based isothermal amplification system, CRISDA, previously described in the literature (**Figure 4A**). CRISDA is a two-stage strand displacement amplification (SDA) method, which relies on target DNA nicking and subsequent nick extension by a strand-displacing DNA polymerase. N-gene directed Cas9 first cuts the nontarget strand of the target DNA sequence. A primer (F) then binds to the free nontarget strand. Together with T4 gp32, a single-stranded DNA binding protein that lowers the melting temperature of duplex DNA, Klenow fragment (KF) then extends from the 3' of the primer and 3' of the nicked target, displacing the target strand and forming an intermediate double-stranded DNA product. The second stage of the process consists of canonical SDA starting from this intermediate product. The F primer contains a restriction site for Nb.BbvCI nickase, which nicks the intermediate product. KF extends from the nick, reconstituting the nickase site, resulting in linear amplification of a single-stranded product (F strand). Another primer (R), also containing a Nb.BbvCI tail, then binds to this single-stranded product and is extended by KF. A similar cyclical nicking and extension results in the production of a second single-stranded product (R strand), which can then bind to F primer and begin the cycle again. This exponential amplification process leads to the production of 3 final DNA products (**Figure 4B**), which can be visualized via gel electrophoresis.

To execute CRISDA reactions, I first produced Cas9 RNP-DNA complexes as I had done previously. I then added primers and enzymes to initiate the CRISDA reaction. I varied the amounts of both N-gene (0, 30 fM, or 30 nM) and target DNA (30 fM or 30 nM) to observe the

production of on-target and nonspecific amplification under different conditions. Notably, at 30 fM of target DNA, no on-target product was amplified when no N-gene RNA was present, while product bands were visible when 30 nM of N-gene was added. To maintain efficient Cas9-DNA binding while detecting smaller amounts of target RNA, I then performed the same experiment with less N-gene (30 fM) but more target DNA (30 nM). I observed on-target product bands with and without added N-gene, indicating that the second stage of the CRISDA reaction was able to take place at high DNA template concentrations. It is possible that the high DNA concentration, combined with the presence of T4 gp32, allowed the F or R primers to invade the DNA target and initiate SDA[229]. A second experiment (**Supplementary Figure 3**), systematically varying the concentration of target DNA, demonstrated that high DNA concentrations allow for nonspecific target amplification in the current reaction setup. However, at a low enough DNA template concentration, SDA can only be efficiently initiated by Cas9 cleavage of target DNA.

5.3 Discussion

While existing CRISPR nucleic acid diagnostics are capable of sensitive and rapid detection of target nucleic acids, the Cas enzymes involved are limited by their nonspecific signal output. *S. pyogenes* Cas9, on the other hand, can be guided to detect many different target DNA sequences, resulting in distinct cleaved products corresponding to each distinct target sequence. I successfully leveraged previously discovered sequence-structure-function rules, developed for single-guide RNA directed Cas9, to re-engineer dual-RNA guided Cas9 as a direct RNA detection platform[221]. Using previously-discovered design rules, I designed 17 tracrRNAs that bound to and detected SARS-CoV2 N-gene RNA, directing Cas9 cleavage of a corresponding DNA target sequence. 85% of the designs were functional in the *in vitro* screening assay,

indicating that these design rules function to efficiently identify highly-targetable regions of a target RNA.

While the original workflow consisted of separate tracrRNA-N-gene annealing, Cas9 loading, RNase processing, and DNA digestion steps, I was able to demonstrate that several of these steps could be combined to run concurrently, decreasing detection time without negatively impacting DNA cleavage efficiency. RNA processing and DNA digestion are also able to be combined without reducing reaction efficiency, further streamlining the protocol. I also demonstrated an *in vitro* RNA processing workflow, and determined what kinds of RNA processing are required for efficient DNA cleavage in this system. Surprisingly, I found that reaction efficiency was highest when RNaseIII processed the N-gene:tracrRNA duplex, but without XRN-1 trimming the 5' end of the N-gene derived crRNA.

I finally sought to develop a post-cleavage amplification system, based on the previously-developed CRISDA method, to allow for direct Cas9-based detection of viral nucleic acids. While current CRISPR-based nucleic acid detection relies on a pre-cleavage amplification of target nucleic acids, post-cleavage amplification requires less knowledge of the target sequence, potentially allowing for the detection of pathogens that have mutated from the original sequence. Using the CRISDA system, I obtained specific amplification induced by the addition of 30 nM of N-gene RNA. While this concentration is significantly higher than the limits of detection currently demonstrated by qPCR and isothermal amplification-CRISPR systems, further protocol optimization could reduce this limit of detection to diagnostically-relevant concentrations. Additionally, sequence-specific barcodes could be added to different target DNA species to allow for multiplexed RNA detection.

S. pyogenes Cas9 is known to be capable of operating in a highly multiplexed manner. While Cas9's lack of collateral cleavage activity means that it cannot be used in current CRISPR nucleic acid detection platforms, its sequence-specific cleavage could be transduced into specific

nucleic acid detection by specifically tagging target DNA, via fluorescent probes, barcode sequences, or similar. Each target cleavage event could then be read out through multiplexed fluorescence detection or next-generation sequencing detection of either cleaved or amplified products[217, 230]. Although I did not demonstrate simultaneous detection of many target RNAs in one reaction, further work could demonstrate multiplex RNA detection using these methods.

5.4 Conclusions

CRISPR-based nucleic acid detection is hindered by the difficulties in multiplex detection using Cas12a and Cas13-based systems. I used a sequence constraint derived from prior work to identify potential crRNA-like sequences in the SARS-CoV2 N-gene, and designed corresponding N-gene sensing tracrRNA sequences. After selecting and testing 20 of the best tracrRNA sensors, I found that 17 directed Cas9 to cleave a designed DNA target sequence. After this initial screen, I then optimized the reaction conditions to minimize time-to-result while maintaining Cas9 activity. Finally, I developed a Cas9-triggered strand displacement amplification-based signal amplification strategy, with the future goal of greatly increasing assay sensitivity.

5.5 Methods

5.5.1 Computational design of tracrRNAs

I identified target RNA sites and designed tracrRNAs using a custom Python script. To design tracrRNAs capable of detecting the presence of SARS-CoV2 N-gene RNA and activating SpCas9 cleavage, I first identified all instances of the motif [N_{26} – GV – N_{28}] in the N-gene. The

additional V constraint was applied to prevent base-paring of the wild-type tracrRNA sequence with that nucleotide, altering the structure of the bulge in the repeat:antirepeat duplex. For each potential crRNA sequence, I then calculated the Hamming distance of the repeat region to the wild-type crRNA sequence. Using RNAfold v2.4, I then calculated the folded structure of a 55-nt window around the beginning of the repeat region of the identified crRNA-like motif, and counted the number of base-paired nucleotides in the guide and repeat regions of the target crRNA. For each identified potential N-gene derived crRNA, I designed a cognate tracrRNA consisting of an engineered antirepeat joined to the wild-type handle sequence and the wild-type tracrRNA terminator. The engineered antirepeat sequence consisted of 28 nt designed to base-pair with the upper stem of the crRNA repeat, the canonical tracrRNA bulge sequence (AAGU), and 6 nt designed to base-pair with the lower stem of the crRNA repeat.

5.5.2 Construction of target DNA and N-gene and tracrRNA RNA

I designed T7 DNA templates consisting of the T7 RNA polymerase promoter, followed by 2 guanosine residues to promote efficient transcription initiation, and then each tracrRNA sequence. I used Primerize to design oligos for PCR assembly. To produce T7 template DNA for the SARS-CoV2 N-gene, I ordered N-gene positive control plasmid (IDT), and designed primers to bind to and amplify it. The forward primer consisted of the T7 promoter, followed by 2 guanosine residues, and finally the primer binding site. For each T7 amplicon, I used Q5 DNA polymerase (NEB) to amplify the sequence, and gel extracted each correct product band. After constructing the T7 DNA templates, I used the HiScribe T7 Kit (NEB) to produce RNA. After a 16-hour in vitro transcription reaction, I used phenol:chloroform extraction, followed by ethanol precipitation, to clean up the transcribed RNA. I then verified RNA integrity via gel electrophoresis.

To produce target plasmids, I obtained the pAN-PA2-RFP target plasmid. I designed new target sequence inserts as oligos (IDT) with flanking EcoRI and NheI sites, and annealed them to create the new target inserts. I digested the original target plasmid with EcoRI and NheI and gel-extracted the plasmid backbone. I then ligated the insert into the backbone using T4 DNA ligase (NEB), and heat-shock transformed the ligated product into chemically competent E. coli DH10B. After sequence confirming each target construct, I purified plasmid DNA, linearized via NdeI (NEB) digestion, and purified the DNA for use in subsequent experiments.

5.5.3 Initial screening of N-gene targeting tracrRNAs

I screened each of the designed tracRNAs for function in binding to N-gene RNA and directing SpCas9 nuclease activity. I first re-annealed the N-gene and tracrRNA to remove any intermolecular structure by diluting samples to 300 nM in 1X TE and then heating to 95 °C for 5 minutes, followed by cooling to 4 °C for 5 minutes. Unless otherwise noted, I then combined 30 nM N-gene RNA and 150 nM each respective tracrRNA in 30 uL of 1X CutSmart buffer (NEB). This mixture was incubated at 37 °C for 1 hr to allow for strand displacement and N-gene:tracrRNA hybridization. I then added 30 nM of Cas9 and incubated for 15 minutes at 37 °C to load the dual-guide RNA complex into Cas9. Subsequently, I added 1 uL each of XRN-1 (NEB) and RNaseIII (Ambion), and incubated for another hour at 37 °C. Finally, I added 3 nM of the corresponding linearized target plasmid, and digested for 1 hour at 37 °C. After digesting with Cas9, I added 1 μL of Proteinase K (NEB) to each reaction, and incubated for 10 minutes at room temperature. I visualized the digestion products of each reaction by running on a 1X TBE, 1% agarose (SeaKem LE, Lonza), 1X GelStar (Lonza) gel. Digital photographs were taken of the gels using a blue light trans-illuminator with an orange filter, and the intensities of the digested product bands were

quantified using GelAnalyzer to determine the degree of digestion. For each of the two cleaved bands, the following formula was used to determine the percent cleavage.

$$\% \text{ cleavage} = \frac{\frac{I_n}{len_n}}{\frac{I_n}{len_n} + \frac{I_{ND}}{len_{ND}}}$$

I_n is the intensity of a given product band, I_{ND} is the intensity of the uncleaved plasmid band, and len_n and len_{ND} are the lengths of the given product band (2979 or 1379 bp) and the uncleaved plasmid band (4358 bp), respectively. I reported the cleavage efficiency of each Cas9 cleavage reaction as the average of the cleavage efficiencies, determined using both product bands, across two independent replicates.

5.5.4 CRISDA

Cas9:tracr7:N-gene:target DNA complexes were prepared and processed as described above, with either 30 fM or 30 nM of N-gene RNA, and either 30 fM or 30 nM of target DNA. CRISDA reactions were then initiated by combining 10 uL of the Cas9 digestion reaction with 2 uL of primer mix and 8 uL of enzyme mix. The primer mix consisted of 500 nM of each primer and 2.5 mM dNTPs, while the enzyme mix contained 0.5 U uL^{-1} Nb.BbvCI (NEB), 1 U uL^{-1} Klenow fragment exo- (NEB), 5 uM T4 gp32 (NEB), and 2.5X CutSmart (NEB). The reactions were then run at 37 °C for 90 minutes. Reaction products were run on a 4% agarose (SeaKem LE, Lonza), 1X TAE gel pre-stained with 1X GelStar (Lonza).

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Over the course of my Ph.D. work, my objective was to implement useful sensor systems using nucleic acid-based sensing elements. I identified two potential sensor platforms – cell-free expression systems and CRISPR-Cas9 systems. Cell-free systems are portable and inexpensive expression platforms, which are currently used to detect diverse target compounds, while CRISPR-Cas9 is a highly-programmable nucleic acid detection platform. For each system, I pursued two objectives. The first was the fundamental characterization of the sensor platform, to examine how each system responded to perturbations, whether chemical or mutational. The second was the development and implementation of novel classes of nucleic acid-based sensors in these platforms, to extend these platforms' utility towards new classes of analyte.

I first sought to examine the effects of the common cosolutes PEG-8000, Ficoll-400, and magnesium glutamate on cell-free expression processes. While the amounts of these cosolutes are commonly tuned to optimize protein expression within these systems, the steps of gene expression these cosolutes effect, as well as their mechanisms of action, are poorly understood. I characterized the effects of these cosolutes on mRNA and protein level dynamics, and derived their effect on the maximum translation rate from each member of a library of ribosome binding site constructs. I was able to leverage my understanding of translation in *E. coli* and extend a previously-developed thermodynamic equilibrium model of translation initiation to incorporate the effects of post-initiation events on the overall translation rate. I determined the relative effects of each cosolute on both initiation and elongation rates, and finally connected these effects to previously-developed theory on the effects of macromolecular crowders on macromolecular

association equilibria and diffusion rates. Finally, I considered the tradeoffs between maximizing protein expression, minimizing the time delay of expression, and maximizing translation rate tuneability that emerge when adding increasing amounts of crowding agent.

After considering the effects of these crowding agents on translational processes in cell-free expression systems, I then moved on to implementing protein-detecting riboswitches in cell-free expression systems. Using a previously-developed model of translational riboswitch regulation, the Riboswitch Calculator, I designed 35 riboswitch sensors for MS2 coat protein, human monomeric C-reactive protein, and human interleukin-32γ. Out of the original 35 sensors, I obtained 34 functional protein sensors. Importantly, at least one of the C-reactive protein sensors obtained a clinically-relevant limit of detection. Overall, I demonstrated the utility of *in silico* sensor sequence design and *in vitro* testing for the rapid and efficient design and test of potential sensor sequences. Further, I examined how protein targets, versus the small-molecule targets the Riboswitch Calculator model was initially validated on, affected the model's predictive accuracy. By incorporating the effects of reduced cell-free translational tunability, direct repression of some proteins, and the effect of supplied aptamer constraint on the model's predictions, I improved the model predictive accuracy to an R^2 of 0.64.

The next sensor platform I examined was the *S. pyogenes* CRISPR-Cas9 system. We had previously designed highly-mutated, nonrepetitive sgRNA handle sequences to allow for the synthesis and cellular expression of extra-large sgRNA arrays (ELSAs). However, the initial work done used a mutated and catalytically inactive version of the Cas9 nuclease, dCas9. To determine how both mutations in specific key functional regions and the overall extent of mutation affected the final DNA cleavage step in wild-type Cas9, I subselected mutated sgRNAs to determine their function in *in vitro* Cas9 cleavage assays. I found that the activity of these mutated sgRNA handles in directing dCas9 repression and Cas9 cleavage was similar, indicating that sgRNA mutations did not affect DNA scission rates subsequent to DNA binding. I further

examined the mechanism of failure of the nonfunctional sgRNA mutants, and determined it was likely that, while Cas9 recognized these nonfunctional mutants, the mutations prevented the conformational shifts required for DNA binding and cleavage. Finally, I considered the off-target activity of the highest-functioning sgRNA mutants, compared to the wild-type sgRNA handle sequence, in both Cas9 and dCas9. I found that, while the mutants had lower off-target activity than the wild-type sequence, they had correspondingly lower on-target activity as well, an observation consistent with the behavior of other off-target activity reducing CRISPR-Cas9 modifications.

After determining how sgRNA handle mutations affect Cas9 DNA binding and DNA cleavage activity, I finally moved on to developing CRISPR-Cas9 into an RNA detection platform. Using a sequence constraint derived from previous work, I identified potential crRNA-like sequences in the SARS-CoV2 N-gene. I subselected 20 sequences based on the amount of predicted secondary structure in the N-gene, and sequence similarity to the wild-type crRNA sequence, and then designed corresponding tracrRNA sequences that would base-pair with the genomic material and form a Cas9-recognizable dual-guide RNA. Of the 20 originally-designed COVID-detecting tracrRNAs, 17 directed Cas9 to cleave a designed DNA target sequence. After this initial screen, I then optimized the reaction conditions to minimize time-to-result while maintaining Cas9 activity. Finally, I developed a signal amplification strategy to work towards increasing the sensitivity of the assay, obtaining amplification of a DNA target.

Overall, I demonstrate here a robust and generalizable sensor platform development process. In order to develop novel sensor systems, it is first useful to identify a sensor platform with useful or disruptive characteristics, such as multiplexability or rapid time-to-result, and then characterize how perturbations to that system, such as different chemical additives or sequence mutations, might affect sensor performance. Subsequently, computational modeling of sensor

sequence function can be used to efficiently search the sequence space to identify functional sensors.

6.2 Future Work

While the work presented in this dissertation demonstrates the concept of an efficient and powerful design-build-test workflow for nucleic acid-based sensors, future work could expand on the potential of this idea in a number of ways. Firstly, there are many more aptameric sensing elements for targets of interest described in the literature, especially for targets of environmental, medical, or industrial biotechnological interest. Additionally, there are many more nucleic acids of interest to be detected, like different pathogens or other nucleic acid biomarkers of disease. We could build on the sensing platforms developed here to develop low-cost sensors for these important analytes. Further, there are other nucleic acid sensor modalities that I did not explore here, such as aptamer-based structure-switching sgRNAs and light-up RNA sensors. These sensors operate under a similar principle to those I developed here, allowing for the engineering of more useful sensors using these platforms.

The concept of computational design of nucleic acid secondary structure switches is powerful, but the thermodynamic equilibrium models of sensor function used in this work do not always capture the reality of the sensor system. Some systems, such as COVID-sensing tracrRNAs, can operate isothermally and in the absence of active transcription via a strand-displacement mechanism. Aptameric sensors, in contrast, operate in a highly nonequilibrium regime. The sensitivity of the protein-sensing riboswitches appears to be governed by the binding affinity of the aptamer, and not the equilibrium folding of the entire ribosome binding site, indicating that the aptamer encountering the ligand governs the folding kinetics of the rest of the RNA molecule. Prior work on riboswitches in the lab, as well as unpublished work, demonstrates

that aptamer-based sensors cannot generally refold to switch states, and instead must rely on co-transcriptional folding, revealing the importance of both active transcription and RNA degradation in ensuring a population of induced-state RNA switches when the target analyte is added. Future work could probe the importance of kinetics in aptamer-based RNA switch function, via both computational modeling of cotranscriptional folding and experimental determination of RNA folding pathways. Additionally, it may be interesting to examine how the placement of the aptamer, the 'decision-making' module of the molecule, within the transcribed sensor sequence affects sensor function.

The protein- and nucleic acid-detecting systems I develop here are novel and have many favorable characteristics, but more work can be done to maximize the utility of these platforms. Firstly, increasing the number of analytes that are capable of being detected simultaneously could decrease cost and potentially time-to-result for many different kinds of assay. The sensor concepts I develop in this dissertation can potentially be coupled with outputs that are highly multiplexable, such as protein riboswitches that control the transcription of a barcoded RNA, or modifying a Cas9-triggered isothermal amplification cascade to allow for the simultaneous amplification of different barcoded amplicons, allowing for the one-pot detection of many distinct targets. To enable the efficient and multiplex detection of targets, it is also important to increase the sensitivity and specificity of the assay. This can be accomplished via improving the recognition elements, such as using SELEX to improve the binding affinity and target selectivity of an aptameric sensing element. It can also be improved by developing or further optimizing signal amplification methods such as isothermal amplification. Finally, the systems I develop here require the use of specialized technology, such as thermal cyclers and plate readers, to effectively sense their target analytes. It would be useful to couple these sensor systems to more portable, human-useful output systems.

Appendix A

Chapter 2 Supplementary Data

Table A-1: RBS sequences used in this study.

RBS Name	Library	5'UTR
A1	SD	TCTAGACCCGCAGCCATATACGGCAGCGGGAAAACACCCAAACCAT
A4	SD	TCTAGACCCGCAGCCATATACGGCAGCGGGACACAAACGACGACAT
A7	SD	TCTAGACCCGCAGCCATATACGGCAGCGGGAAAATAACGAGCACAT
A10	SD	TCTAGACCCGCAGCCATATACGGCAGCGGGAAAATAACGAGGCCAT
A13	SD	TCTAGACCCGCAGCCATATACGGCAGCGGGAAAACAAGGAGCACAT
A16	SD	TCTAGACCCGCAGCCATATACGGCAGCGGGAAAATAAGGAGGACAT
A19	SD	TCTAGAACCCGCCATATACGGCGGGACACACACAAGGAGACCAT
U2	standby site	TCTAGAGGCCGACGCAAGCCCATATCGGGGCTTCCGTCGGCCAATAAGGAGGTAAAAA
U10	standby site	TCTAGAGGCCGACGCAAGATATCGCTTCCGTCGGCCAAATAAGGAGGTAAAAA
U11	standby site	TCTAGAGGCCGACGCAAGATATCGCTTCCGTCGGCCAAAATAAGGAGGTAAAAA
U47	standby site	TCAGGCCGACGCAAGCCCATATCGGGGCTTCCGTCGGCCATAAGGAGGTAAAAA
U48	standby site	TCAAGGCCGACGCAAGCCCATATCGGGGCTTCCGTCGGCCATAAGGAGGTAAAAA
U49	standby site	TCAAAAGGCCGACGCAAGCCCATATCGGGGCTTCCGTCGGCCATAAGGAGGTAAAAA
U51	standby site	TCAAAAAAAAAAGGCCGACGCAAGCCCATATCGGGGCTTCCGTCGGCCATAAGGAGGTAAAAA

Table A-2: RBS Calculator v2.1 predicted free energies and translation initiation rates (TIRs) for RBSs in this study.

RBS Name	Library	ΔG_{total}	$\Delta G_{SD:antiSD}$	ΔG_{mRNA}	$\Delta G_{standby}$	Predicted TIR
A1	SD	2.49	1.85	-9.22	3.31	814
A4	SD	0.98	-0.37	-11.46	2.66	1607
A7	SD	-2.01	-3.34	-14.41	2.64	6173
A10	SD	-4.72	-6.05	-17.12	2.64	20899
A13	SD	-6.42	-7.75	-18.82	2.64	44913
A16	SD	-10.04	-11.37	-22.44	2.64	229001
A19	SD	-9.00	-7.75	-17.48	0.06	143245
U2	Standby site	-1.98	-13.15	-36.30	6.77	6090
U10	Standby site	-3.60	-13.15	-29.00	5.79	12622
U11	Standby site	-4.68	-13.15	-29.00	4.89	20548
U47	Standby site	1.88	-13.15	-36.30	11.45	1071
U48	Standby site	0.69	-13.15	-36.30	10.17	1834
U49	Standby site	-0.92	-13.15	-36.30	7.83	3787
U51	Standby site	-6.93	-13.15	-36.30	2.64	56575

Table A-3: SYBR qPCR primers and primer efficiencies.

	SYBR Efficiency	
Gene Target	**mRFP1**	**iclR**
Forward Primer	ACGTTATCAAAGAGTTCATGCGTTTC	GACTGTCatgGTCGCACC
Reverse Primer	CCTTCACCTTCACCTTCGATTTCG	GTGCCACACTGCCATTGG
Amplicon Length (nt)	88	143
Primer Efficiency	1.01%	95.35%
No Dilution Ct Value	17.15	22.25
Dilution 1:10 Ct Value	20.45	25.57
Dilution 1:100 Ct Value	23.70	29.06
Dilution 1:1000 Ct Value	27.13	32.68
Dilution 1:10000 Ct Value	30.17	35.89

Dilution 1:100000 Ct Value	33.70	
y-intercept	-3.30	-3.44
slope	36.93	42.84
R^2	0.9999	0.9997

Table A-4: 2-state model fit initiation and elongation parameters for the base, 4% PEG-8000, 4% Ficoll-400, and 16.67 mM Mg-glut conditions.

	c_1 (initiation coefficient, a.u. hr^{-1})		**k_{elong} (a.u. hr^{-1})**	
Cosolute condition	**parameter value**	**std err**	**parameter value**	**std err**
base	4.06	0.18	228.54	36.80
4% PEG-8000	32.19	3.97	125.39	15.82
4% Ficoll400	11.55	1.30	102.45	16.51
16.67 mM Mg-glut	4.77	0.54	40.75	6.50

Appendix B

Chapter 3 Supplementary Data

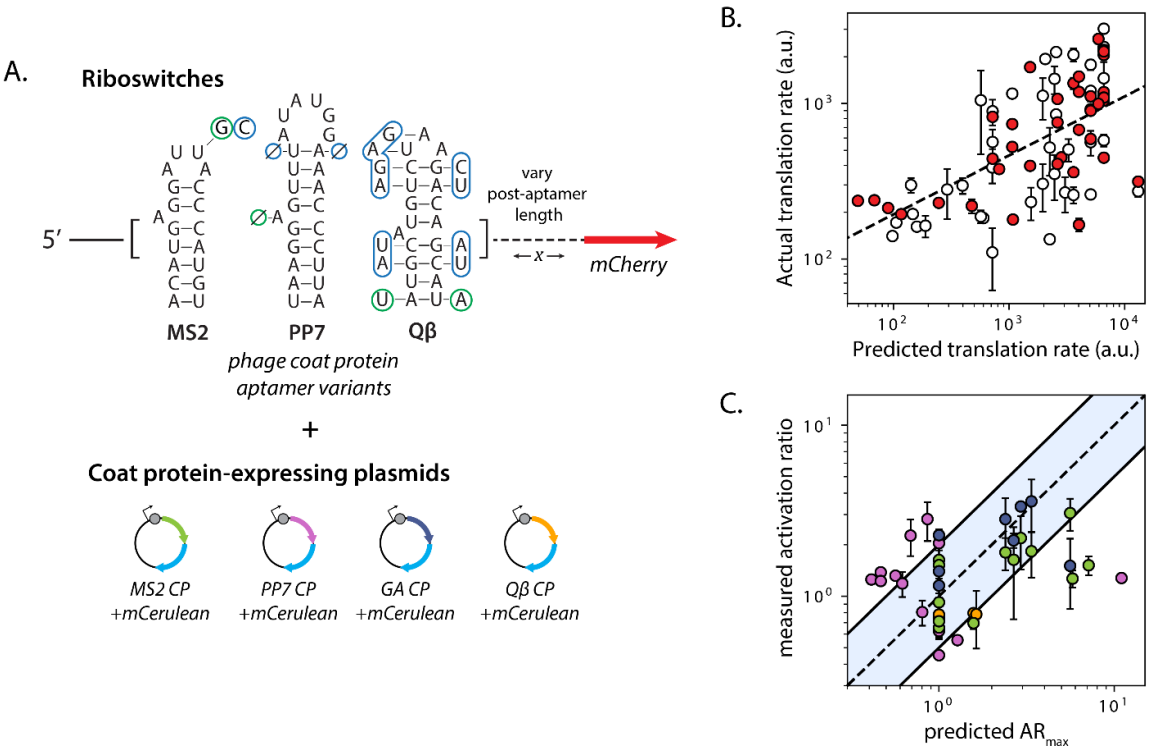

Figure B-1: Phage coat protein aptamer-based riboswitches. The Riboswitch Calculator model is predictive of uninduced and induced expression from riboswitches which detect the phage coat proteins from E. coli bacteriophages MS2, PP7, GA, and Qβ, from Katz et al. A. Riboswitches and inducer plasmids used. Variants of the MS2, PP7, and Qβ hairpins were placed at varying distances from the mCherry CDS, and coexpressed with coat protein-mCerulean fusion proteins in E. coli. B. Predicted and actual expression from coat protein riboswitches. White circles represent uninduced expression, while red circles represent maximal induced expression from riboswitches. Black dotted line is best-fit line comparing predicted and actual expression for both maximally induced and uninduced riboswitches (R^2 = 0.35). C. Measured vs. predicted riboswitch activation ratios. Circle colors correspond to the inducer protein in (A) (green circles, MS2-mCerulean; pink circles, PP7-mCerulean; blue circles, GA-mCerulean; orange circles, Qβ-mCerulean). Dotted black line represents measured activation ratio equivalent to predicted ARmax, while solid black lines represent 2-fold over- or under-predicted. The correlation between predicted and actual activation ratio is lower than that between predicted and actual expression, with R^2 = 0.12. However, the 59% of riboswitch activation ratios are predicted within 2-fold. Error bars for (B) and (C) represent the standard deviation (n=2).

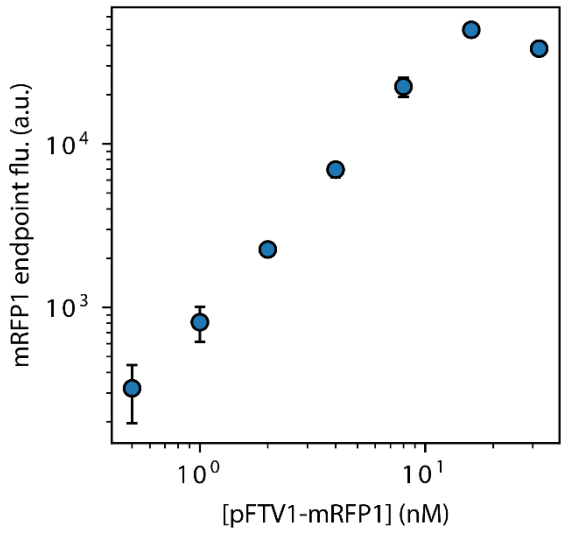

Figure B-2: mRFP1 fluorescence vs. [pFTV1-mRFP1]. Fluorescence (blue circles) is proportional to mRFP-expressing plasmid concentration up to 32 nM. Error bars represent 95% confidence interval (n=2).

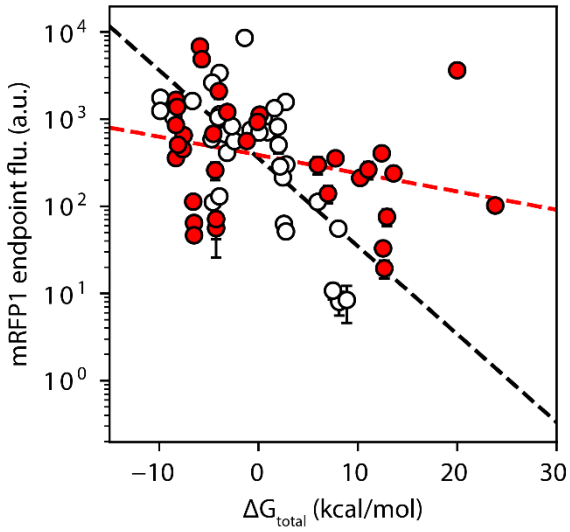

Figure B-3: mRFP1 fluorescence vs. Riboswitch Calculator-predicted free energies of translation initiation for all 35 riboswitches. The Riboswitch Calculator predicts the free energy of translation initiation (ΔGtotal) for riboswitches in the uninduced and induced states. White circles are uninduced mRFP1 endpoint fluorescence, while red circles represent induced mRFP1 fluorescence. Dotted lines are the best-fit line of the data to log(mRFP1 endpoint flu.)=log(K)-βΔG¬_total (black dotted line – uninduced riboswitch fluorescence, red dotted line – induced

riboswitch fluorescence), where the parameter K is a fitting coefficient and β represents the thermodynamic temperature of the system. In vivo, β = 0.45 across diverse bacterial species (Farasat et al.) However, here I find that for the riboswitches in the uninduced state, β = 0.23, representing a reduced range of cell-free translation. R2 = 0.48 for uninduced riboswitches. The correlation between measured mRFP1 fluorescence and predicted ΔGtotal is much lower for induced riboswitches, with R2 = 0.10. Error bars represent 95% confidence interval (n=6 for MS2 on-riboswitches, n=8 for all other riboswitches).

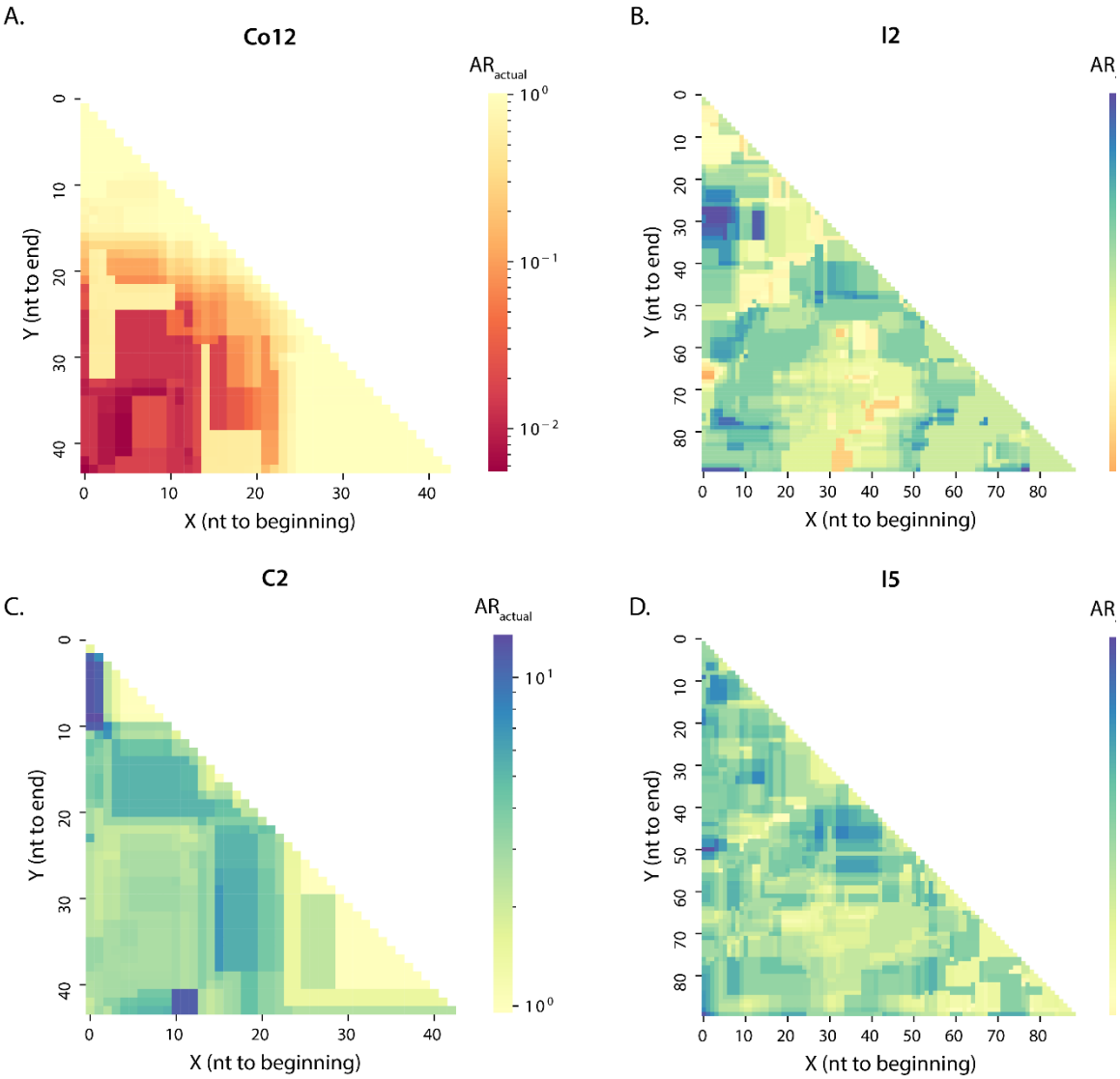

Figure B-4: Effect of varying aptamer constraint region on other CRP and IL32 riboswitches. The constrained aptamer region of riboswitches Co12, C2, I2, and I5 were varied, and the effect on predicted ARactual was calculated. For all, X represents the distance from the start of the full aptamer sequence to the beginning of the subconstrained region, and Y represents the distance from the start of the full aptamer sequence to the end of the subconstrained region. A. Varying size of CRP aptamer constraint in riboswitch Co12. B. Varying size of IL32 aptamer constraint in riboswitch I2. C. Varying size of CRP aptamer constraint in riboswitch C2. D. Varying size of IL32 aptamer constraint in riboswitch I5.

Table B-1: Riboswitch sequences tested in this study.

Riboswitch name	Riboswitch 5'UTR sequence
M1	TCTAGAAATGAGAAGAGACACATGAGGATCACCCATGTCTGAGAAAAGATAAGGAG GGTGAGG
M2	TCTAGACAAGTAAAGAGACACATGAGGATCACCCATGTCTGAAGTAAGATAAGGAGG GTGAGG
M3	TCTAGACAAAGAATGAAACACATGAGGATCACCCATGTCTGGAGACAGACAAGGAGG GTGAGT
M4	TCTAGACAAAAAATGAAACACATGAGGATCACCCATGTCTGGAGACAGACAAGGAGG GTGAGT
M5	TCTAGACTTAAAATGAGACACATGAGGATCACCCATGTCTGGAGACAGGCAAGGAGG GTGATA
Mo1	TCTAGATTGGTGCTCAGTCCCGCCGCGAGGGACATGAGGATCACCCATGTGGGGTAA ATGAGGAGGAAGGG
Mo2	TCTAGACAGTCTCATGTTCCATCCGATGGTAGGGAACATGAGGATCACCCATGTAGGG GCATGAGATGGTATC
Mo3	TCTAGATGCCTCATGTCCCTCCAGAGTGGGACATGAGGATCACCCATGTAGACTACAA GGAGGGTATGGG
Mo4	TCTAGAGTAGCGGTTCGGGTGGTCTTCAGATTCACATGAGGATCACCCATGTGTTCTG GGGAGGAGCC
Mo5	TCTAGAGGTTCCTCCATGTCCCAGGATGGGACATGAGGATCACCCATGTAGCTGTAAG GAGGACAGG
Mo6	TCTAGATGCCTCATGTCCCTCCAGAGTGGGACATGAGGATCACCCATGTAGGGCACAA GGAGGTATGGG
Mo7	TCTAGACAGCCCTCATGTACGCTCCGATGAGTAACATGAGGATCACCCATGTGGGGCA TGAGGATGGTATC
Mo8	TCTAGATCCTCATGCTCGCCTCCAGAGTGGGACATGAGGATCACCCATGTAGACTGGG GATGAGGTTATGTGG
Mo9	TCTAGACAGTCTCATGTTCCCCAATCCGATGGTTAACATGAGGATCACCCATGTAGGG GCATGAGATGGTATC
M10	TCTAGATGCCTCATGTCCCTCCAGAGTGGGACATGAGGATCACCCATGTAGGCACAAG GAGGGTATGGG
C1	TCTAGATATGCTGAAGTACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCGTCAAGGCCATTAAGGAGGAGTACG
C2	TCTAGAGAAACTGAAATTCGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCGACAAGGCCATTAAGGAGGTGAAAG
C3	TCTAGAAAAGCTGAAGTACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCCGCAAGGCCATGAAGGAGGCGTAGG
C4	TCTAGAAATAAGCTAATACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCCCATTGGCTAATAAGGAGGGTGAAA
C5	TCTAGACAAAAGCAGTTACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCTCACGGGTCAATAAGGAGGGTGGGT

C6	TCTAGACATATACAGCGACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCTCACAGGCCAATAAGGAGGGTGAGG
C7	TCTAGAAGTATTGCAGCACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCGACAAGGCTAATAAGGAGGGTGAGA
C8	TCTAGAAAAGCTGAACAACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCCGCAAGGCCATGAAGGAGGCGTAGG
C9	TCTAGAAAAGCTGAAGAACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAG AGATTGCCGCAAGGCCATTAAGGAGGCGTGGC
C10	TCTAGAAACAAACAGTGACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGA GATTGCCCACGGGCCAATAAGGAGGGTGAGT
Co1	TCTAGAGCTCGCCACACCGCCGCACTGCCTGTAAGGTGGTCGGTGTGGCGAGTGTGT TAGGAGAGATTGCGAAGAAAGGTGTGGGGGA
Co5	TCTAGAATTGAATCTCCTAGACACTCCACGCCTGTAAGGTGGTCGGTGTGGCGAGTGT GTTAGGAGAGATTGCACGGTATAAGGAGGTAGGAG
Co3	TCTAGAGTCTCGCCACCGCCGCCCACCTGCCTGTAAGGTGGTCGGTGTGGCGAGTGT GTTAGGAGAGATTGCAAGTGTACAAGGGGGTGGCGAT
Co2	TCTAGAGCTCGCCGCACCCAGCCCCTTAGCCTGTAAGGTGGTCGGTGTGGCGAGTGT GTTAGGAGAGATTGCGTAGTCGTGAGACAAAGGAGGTTTTGG
Co4	TCTAGAGTTATCTCCAAGTACACTCCACGCCTGTAAGGTGGTCGGTGTGGCGAGTGTG TTAGGAGAGATTGCACGTATAAGGAGGTAGGGG
I1	TCTAGAATTTAAAAAGAAAGATCTGTGGGTTCACTGCAGACTTGACGAAGCTTCCGGA GAGAAGGGTCAAAGTTGTGCGGGAGTGTGTTGTGGAATGGATCCACATCTACGAATT CCTTAGATGATATAAGGAGGGTGAAA
I2	TCTAGATAGAAAGATAAACGGTAAGAGGGTTCACTGCAGACTTGACGAAGCTTCCGG AGAGAAGGGTCAAAGTTGTGCGGGAGTGTGTTGTGGAATGGATCCACATCTACGAAT TCCGTGCAAGAGATAAGGAGGGTGAAA
I3	TCTAGAATGAATAAATAAGAGTTACAGGGTTCACTGCAGACTTGACGAAGCTTCCGG AGAGAAGGGTCAAAGTTGTGCGGGAGTGTGTTGTGGAATGGATCCACATCTACGAAT TCGCTACATTGGATAAGGAGGGTGAGT
I4	TCTAGAATTAATTAATATGAGTGACGGGGTTCACTGCAGACTTGACGAAGCTTCCGGA GAGAAGGGTCAAAGTTGTGCGGGAGTGTGTTGTGGAATGGATCCACATCTACGAATT CGCTACTGTGGATAAGGAGGGTGAGG
I5	TCTAGAATGCAATTGATAAAGTGTCAGGGTTCACTGCAGACTTGACGAAGCTTCCGGA GAGAAGGGTCAAAGTTGTGCGGGAGTGTGTTGTGGAATGGATCCACATCTACGAATT CGCTACACTAAATAAGGAGGGTGAAA

Table B-2: Alternate aptamer constraint sequences and structures. MS2 is the MS2 aptamer, AC3-3 is the IL32 aptamer, and VT2d0 is the CRP aptamer.

Aptamer name	Aptamer sequence	Aptamer structure
MS2	ACATGAGGATCACCCATGT	(((((x((xxxx)))))))
AC3-3 (full)	GGGTTCACTGCAGACTTGACGAAGCTTCCGGAGAGAAGGGTCAAAGTTGTGCGGGAGTGTGTTGTGGAATGGATCCACATCTACGAATTC	((((((x((((((((((((xxxx(((xxxxx)))xxxx)))x)))))x)))))xxxx(((x((((((xxxxx))))))xx))))))))
AC3-3a	CTGCAGACTTGACGAAGCTTCCGGAGAGAAGGGTCAAAGTTGTGCGG	((((((((((((xxxx(((xxxxx)))xxxx)))x)))))x))))))
AC3-3b	GACTTGACGAAGCTTCCGGAGAGAAGGGTCAAAGTT	((((((((xxxx(((xxxxx)))xxxx)))x)))))
AC3-3c	GACGAAGCTTCCGGAGAGAAGGGTC	(((xxxx(((xxxxx)))xxxx)))
AC3-3d	GACGAAGCTTCCGGAGAGAAGGGTC	(((xxxxx)))
AC3-3e	GTGTTGTGGAATGGATCCACATCTAC	(((x((((((xxxxx))))))xx)))
AC3-3f	TGTGGAATGGATCCACA	((((((xxxxx))))))
VT2d0 (full)	GCCTGTAAGGTGGTCGGTGTGGCGAGTGTGTTAGGAGAGATTGC	xxxxxxxxx((((((xxxx(((((xxxx))))) xxxxx))))))
VT2d0-a	GTGGTCGGTGTGGCGAGTGTGTTAGGAGAGATTGC	((((((xxxx(((((xxxx)))))xxxxx))))))
VT2d0-b	TGGCGAGTGTGTTA	(((((xxxx)))))

Table B-3: Promoters, RBSs and coding sequences of proteins used in this study.

CDS name	Promoter	RBS sequence	CDS sequence
mRFP1	J23100	TCTAGAAAAAAAAATAAGGAGGTAAAAA	ATGGCGAGCTCTgaagacgttatcaaagagttcatgcgtttcaaagttcgtatggaaggttccgttaacggtcacgagttcgaaatcgaaggtgaaggtgaaggtcgtccgtacgaaggtacccagaccgctaaactgaaagttaccaaaggtggtccgctgccgttcgcttgggacatcctgtccccgcagttccagtacggttccaaagcttacgttaaacacccggctgacatcccggactacctgaaactgtccttcccggaaggtttcaaatgggaacgtgttatgaacttcgaagacggtggtgttgttaccgttacccaggactcctccctgcaagacggtgagttcatctacaaagttaaactgcgtggtaccaacttcccgtccgacggtccggttatgcagaaaaaaaccatgggttgggaagcttccaccgaacgtatgtacccggaagacggtgctctgaaaggtgaaatcaaaatgcgtctgaaactgaaagacggtggtcactacgacgctgaagttaaaaccacctacatggctaaaaaaccggttcagctgccgggtgcttacaaaaccgacatcaaactggacatcacctcccacaacgaagactacaccatcgttgaacagtacgaacgtgctgaaggtcgtcactccaccggtgcttaataa
MS2 coat protein	J23100	TCTAGAacgaacgaaattataaggaggtaaa	atggcatcaaactttacccaatttgtattggtcgacaacggcggtactggggatgtgactgtggcaccttcgaactttgcgaacggtgtagcggagtggatctcgtctaattcgcgatcacaggcatataaagtgacttgctcagtacgacaaagttctgctcaaaatcgcaagtatacgataaaagttgaggttccgaaagtggcgactcaaacggttggcggtgtggagttgccggtagctgcttggcgctcgtatttgaacatggagctgactataccgatttttgcgactaatagcgactgcgaactgattgtgaaagcgatgcaaggcctgttgaaggatgggaatcctattccgtcggcaatcgcggctaactcgggtatatactga

Appendix C

Chapter 4 Supplementary Data

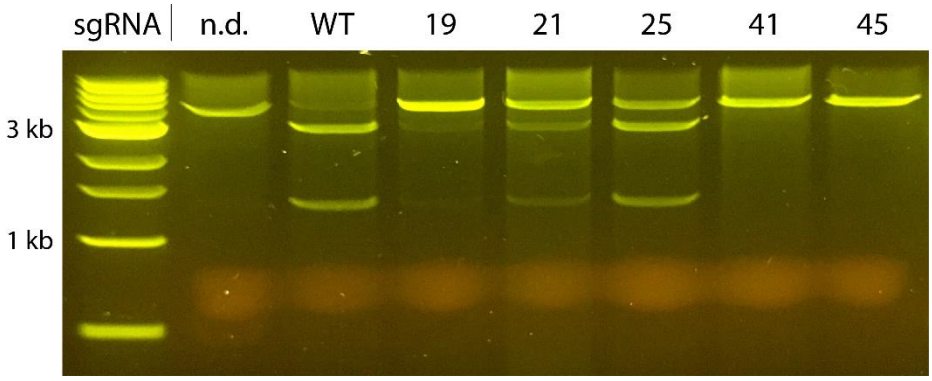

Figure C-1: 16-hr DNA digestion with low-functioning sgRNAs. Lanes from left to right: undigested target DNA, wild-type sgRNA, sgRNA 19, sgRNA 21, sgRNA 25, sgRNA 41, sgRNA 45. There was no detectable digestion product for sgRNA 41 or sgRNA 45.

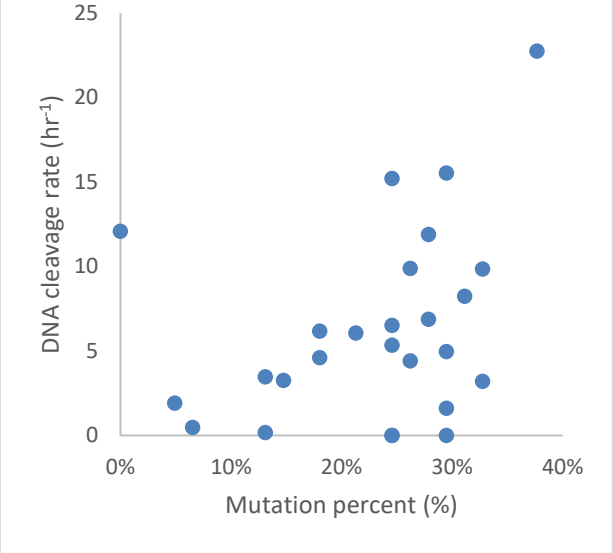

Figure C-2: Mutation percentage vs. Cas9 DNA cleavage rate is not significant. ($R2 = 0.11$, $p = 0.097$).

Table C-1: Nonrepetitive sgRNA guide and handle sequences, and set ID.

sgRNA#	Set	guide	handle
WT	-	AGCTTCTG AGTAGTCTA AGT	GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTA GTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGT GCTTTTTTT
5	1	AGCACACA CGCGCTAG AGAC	GTTTTATAGCTAGAAATAGCAAGATAAAATAAGGCTAG TCCATTATCAACTTGAAAAAGTGTTCCATCGGGTCCG AATTTTCGGACCTTTTCTCCGCA
7	1	CTAGTCAC GTCTCTAA GCAC	GTTTTAGATCACGAAAGTGAAAGTTAAAATAAGCCTAG CCCGTTACCAACTGGAAACAGTGACTTAAGACCGCCG GTCTTGTCCACTACCTTGCAGTAATGCGGTGGACAGG ATCGGCGGTTTTCTTTTCT
19	1	ATACTCGA CTTCGAAG CTTG	GTTTTAGAGCTACACATAGCAAGTTAAAATAAAGGTAG TCCGTTATCAGTTTGAAAAAACGTCTGGTAGTCCTGG TAAGACGCGAACAGCGTCGCATCAGGCATATTGCCAA CTAG
21	1	GAGTCGTT CGCGAGCT AGCT	GTTGTAGAGCTAGAAATAGCGAGTTACAATAAGGCTA GTCCGTTATGAACTTGAAAAAGTGGTGAAAAATGGCG CCCATCGGCGCCATTTTTTTATGA
25	1	TCTAAGCTC GCTGTAGA CTA	GTTTTACACCTAGAAATAGGAAGGTAAAATAAGGCTG GTCCGTTATCACCTCGAAAGAGGGTTATTGAAGACGC TTAACAGCGTCTTTTTTTGTTT
26	1	TCACGACT AGAGCTAG CACG	GTTGTAGAGCTAGCAATAGCAGGTTACAATAAGGCTC GTCCGTTATAAACATGAAAATGTGACTAAAAAGGCCG CTCTGCGGCCTTTTTTCTTTT
27	1	ACTAGTTGT CGTAACAT CTA	GATTTCGAGCTAGGCATAGCAAGTGAAATTAAGGCTG GTCCATTAACACCTTGAAAAAGGGAACAATAAGGCCT CCCTTTAGGGGGGGCCTTTTTTATTGA
41	2	GAGTCGTT CGCGAGCT AGCT	GTGTTAGAGTCAGATATGACATGTTAACATTAGGCTA GTCCGGGGTGAAGTTGAAAAACTGAAGCAGAAGGCC ATCCTGACGGATGGCCTTTTTGCGTTT
42	2	CTCTAATCG CGCTTCTA GAG	TTACTAGAGTGACAAATCACAAGTTAGTAAAAGGCTA GACCGTTATAATCCCGAACGGGAGGCACAAAAGCCA GTCTGGAAACAGGCTGGCTTTTTTTTGCG
43	2	CATATACTA CATACTCTA TT	TTTTCAGATTTGGAAACAAAACGTTGAAAAAAGGCAA GTCCGTTATGAACGCGAAAGCGTGCGAAAAAACCCG CTTCGGCGGGTTTTTTTATAG
44	2	GTATAATAG AAAACTAGA CA	GTATGCGAGGTAGAAATACCCAGTGCATATCAGGCTA GTCCGATATCATGTTGAAGAACAGATAACAAAGCCGG GTAATTCCCGGCTTTGTTGTATC
45	2	TCTAAGCTC GCTGTAGA CTA	CGGTTAGGATAAGAAATTATAAGTTAACCGTAGGCTA GCCCGTTATAAACTGGAAACAGTGTCCTTTCACGGCC CATTCCTCATGGATGGGCCGTTTATTTCCCC
46	2	TCACGACT AGAGCTAG CACG	GATGTAGATGTAGAAATACAAGGTTACATTAAGGCCC GTCCGTAATCAACTTGAAGAAGTGTTCCATCGGGTCC GAATTTTCGGACCTTTTCTCCGCA
47	2	GTCGCTCG AGACACGA AAAG	GTTTTGGACCTAGAAATAGGAAGTCAAAATAAGGCTG GACCGACATGTAATCGAAAGATTTAGTCAAAAGCCTC CGGTCGGAGGCTTTTGACTTTC
48	2	TCACTAGTT GTCTGAGC TCT	CCTTAAGAGCTAGCAATAGCAAGTTTAAGGAAGGCAA GCCCGTTATCATCCTGAATAGGACAAAAAAAATCCTTA GCTTTCGCTAAGGATGATTTCTACT

49	2	ACTTACAGTATAAATTATAG	GTAATAGAGATGGATACATCAAGTTATTATAAGGCTCGACCGTTAACAGTCTGAAAAGACGATTCAAGACCCCCGCACCGAAAGGTCCGGGGGTTTTTTTTACT
51	2	CTAGATCGTTTAACTAAGTA	GTTGGAGAGCAAGACATTGCAAGTTCCAATAAGGCGTGTCCGATAAAAGCTTGAGAAAGCAAAAACGAAGGCCGGAGCATGCTCCGGCCTTTTTTATCTC
52	2	AATTAGTACTTCTTATAGTT	ATCTGAGAGCCAAAAATGGCAAGTTCAGATAAGGCCAGACCGTTACCAGCTTAAATAAGCGATCCTAAAGCCCCGAATTTTTTATAAATTCGGGGCTTTTTTACTAG
53	2	ATCTAGTTCGCTATACACGT	GCTTCAGATCCAGAAATGGAAAGTTGAAGTGAGGCAGGTCCGGTAGCAACTCGAAAGAGTGAGAAAAGAGGGGAGCGGGAAACCGCTCCCCTTTTTTCGTTT
54	2	GATACAGTCAGTCGCTTGAC	AGTTTAGAGAATGCAAATTCAAGTTAAACTAAGGCGAGTCCGGTATAATCGTGTAAACGAGGTGAAAAATGGCGCCCATCGGCGCCATTTTTTTATGA
55	2	AAGTCTAGTACGCTTACATA	GGAATAGAAAACAAAAGTTTAAGTTATTCTAAGGCCAGTCCGGAATCATCCTAAAAAGGAGTTATTGAACACCCGAAAGGGTGTTTTTTTGTTT
56	2	TCGAAGTATATACTAACGAG	GTGCTAGAGTCGTAAACGACAAGTTAGCATTAGGCTTGTCCGCAATGAACCTGAAAAGGTGATCCAAAACGCCGCGTTCAGCGGCGTTTTTTCTGCT
57	2	CTCATAGTAGTACATACGAA	CATTTTGGCGTCGAAAGACGAAGTAAAATGAAGGCGAGACCGATATCAACTGGAAGCAGTGTCTGGTAGTCCTGGTAAGACGCGAACAGCGTCGCATCAGGCATATTGCCAACTAG
58	2	AGACAGAGTCGCTATCTAAA	TTTTTAGAGGAAGGAATTCCAAGTTAAAAAAAGGCAGGACCGGGAACATGTTGAAAAACAGGCAAAAAAGCGCCTTTAGGGCGCTTTTTTACATT
59	2	TGTCTAACTACACTTGACAG	CTTACCGAACTAGGAATAGTAAGTGGTAAGAAGGCCTGACCGTAATAAGCCTGAAAAGGCGACCAAAAAGGGGGGATTTTATCTCCCCTTTAATTTTTC

Table C-2: Off-target sequences and CFD scores. Mutated nucleotides are in red, and PAM sequence is in blue.

Sequence name	Target sequence	CFD score
On-target	GAGTCGTTCGCGAGCTAGCTAGG	1
MM1	GAGTCGTTCGCGAGCTAGC**g**AGG	0.176
MM2	GAGTCGTTCGCGAGC**c**AGC**g**AGG	0.118
MM3	GAGTCGTTCG**t**GAGC**c**AGC**g**AGG	0.036
MM4	GAGTCGTTCG**t**G**g**GC**c**AGC**g**AGG	0.024
Noncanonical PAM	GAGTCGTTCGCGAGCTAGCTA**a**G	0.259

Appendix D

Chapter 5 Supplementary Data

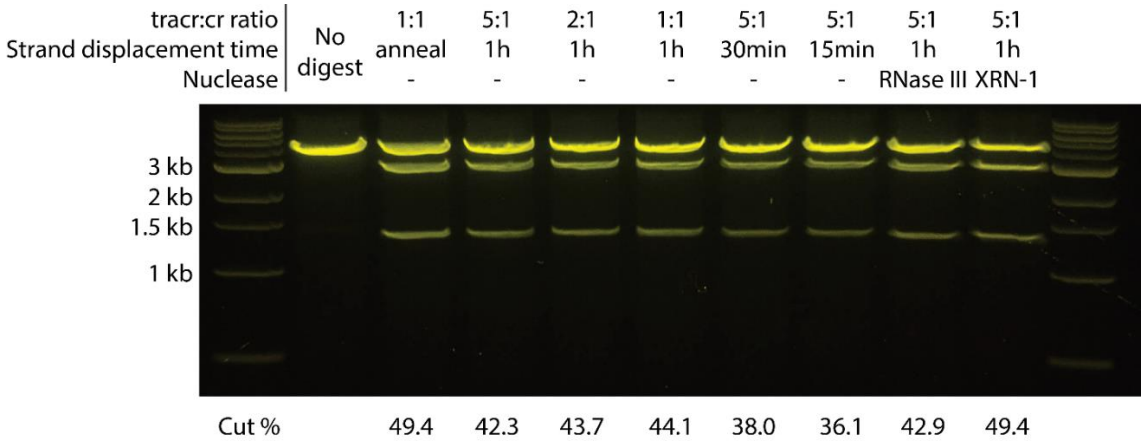

Figure D-1: Initial proof-of-concept workflow using wild-type dual guide RNA. The tracrRNA:crRNA ratio, heat annealing vs. length of strand displacement time, and nuclease digestion steps were systematically varied, and the impact on the produced Cas9 RNP's ability to cleave target DNA was measured.

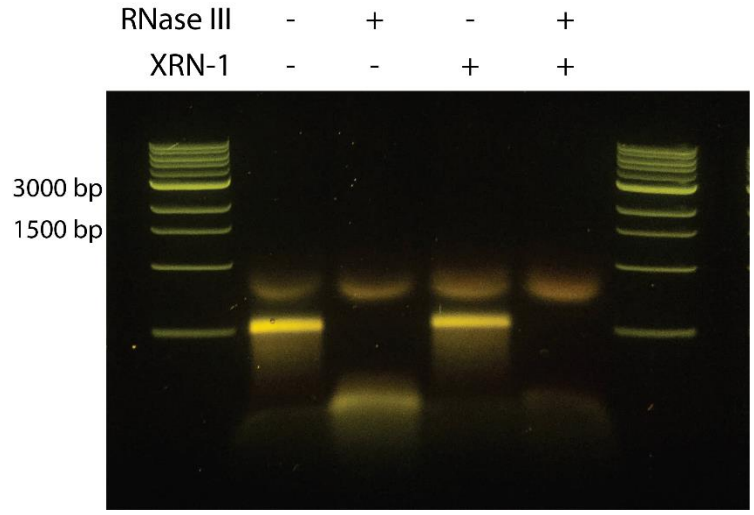

Figure D-2: RNase digestion of N-gene T7 transcripts. The 1260 nt N-gene was digested with either RNaseIII, XRN-1, or both nucleases simultaneously, and run on a non-denaturing agarose gel.

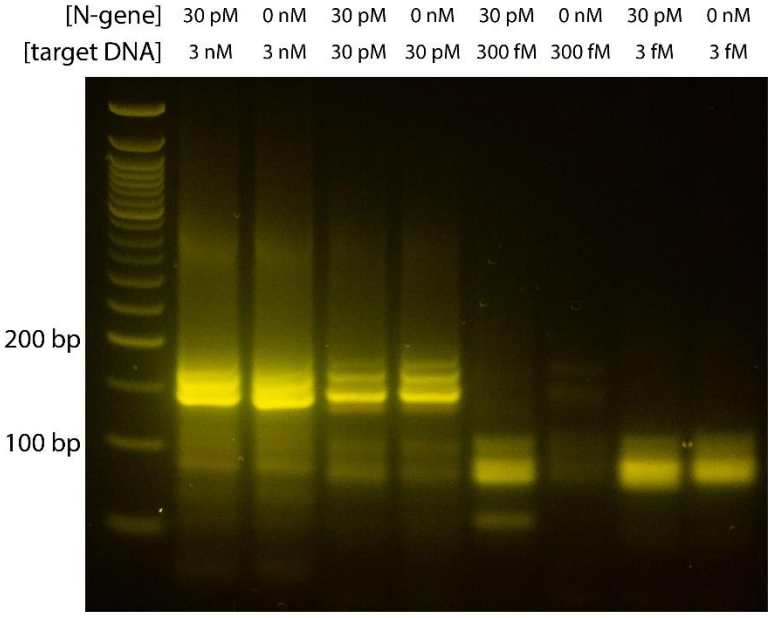

Figure D-3: Nonspecific and specific SDA at different DNA template concentrations. Cas9 RNPs were prepared with 30 pM of N-gene, and subsequent CRISDA reactions were set up with DNA template concentrations varying between 3 nM and 3 fM, with or without the Cas9 RNP added. SDA product bands are between 154-169 bp long.

Table D-1: N-gene targeting tracrRNA sequences and corresponding crRNAs. crRNA loc corresponds to the position of the constrained crRNA 'G' in the SARS-CoV2 genome. Guide seq and crRNA seq are the processed guide and crRNA derived from the N-gene.

ID	crRNA loc	guide seq	crRNA seq	tracrRNA
1	28444	CTTGGTTCACCGCTCTCACT	CAACATGGCAAGGAAGACCTTAAATTCCCTCGAGGA	TCCTCGAGGGAATTTAAGGTCTTCCTTGAAGTATGTTGAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
2	28445	TTGGTTCACCGCTCTCACTC	AACATGGCAAGGAAGACCTTAAATTCCCTCGAGGAC	GTCCTCGAGGGAATTTAAGGTCTTCCTTAAGTCATGTTAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
3	28450	TCACCGCTCTCACTCAACAT	GGCAAGGAAGACCTTAAATTCCCTCGAGGACAAGGC	GCCTTGTCCTCGAGGGAATTTAAGGTCTAAGTCTTGCCAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
4	28453	CCGCTCTCACTCAACATGGC	AAGGAAGACCTTAAATTCCCTCGAGGACAAGGCGTT	AACGCCTTGTCCTCGAGGGAATTTAAGGAAGTTTCCTTAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
5	28507	TTCCAATTAACACCAATAGC	AGTCCAGATGACCAAATTGGCTACTACCGAAGAGCT	AGCTCTTCGGTAGTAGCCAATTTGGTCAAAGTTGGACTAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
6	28510	CAATTAACACCAATAGCAGT	CCAGATGACCAAATTGGCTACTACCGAAGAGCTACC	GGTAGCTCTTCGGTAGTAGCCAATTTGGAAGTATCTGGAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
7	28573	TTCGTGGTGGTGACGGTAAA	ATGAAAGATCTCAGTCCAAGATGGTATTTCTACTAC	GTAGTAGAAATACCATCTTGGACTGAGAAAGTTTTCATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
8	28696	CTGAGGGAGCCTTGAATACA	CCAAAAGATCACATTGGCACCCGCAATCCTGCTAAC	GTTAGCAGGATTGCGGGTGCCAATGTGAAAGTTTTTGGAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
9	28705	CCTTGAATACACCAAAAGAT	CACATTGGCACCCGCAATCCTGCTAACAATGCTGCA	TGCAGCATTGTTAGCAGGATTGCGGGTGAAGTAATGTGAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
10	28706	CTTGAATACACCAAAAGATC	ACATTGGCACCCGCAATCCTGCTAACAATGCTGCAA	TTGCAGCATTGTTAGCAGGATTGCGGGTAAGTCAATGTAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
11	28740	AATCCTGCTAACAATGCTGC	AATCGTGCTACAACTTCCTCAAGGAACAACATTGCC	GGCAATGTTGTTCCTTGAGGAAGTTGTAAAGTACGATTAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
12	28850	CTCATCACGTAGTCGCAACA	GTTCAAGAAATTCAACTCCAGGCAGCAGTAGGGGAA	TTCCCCTACTGCTGCCTGGAGTTGAATTAAGTTTGAACAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT

13	29089	AAGCATACAATGTAACACAA	GCTTTCGGCAGACGTGGTCCAGAACAAACCCAAGGA	TCCTTGGGTTTGTTCTGGACCACGTCTGAAGTGAAAGCAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
14	29093	ATACAATGTAACACAAGCTT	TCGGCAGACGTGGTCCAGAACAAACCCAAGGAAATT	AATTTCCTTGGGTTTGTTCTGGACCACGAAGTTGCCGAAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
15	29259	GAAGTCACACCTTCGGGAAC	GTGGTTGACCTACACAGGTGCCATCAAATTGGATGA	TCATCCAATTTGATGGCACCTGTGTAGGAAGTAACCACAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
16	29269	CTTCGGGAACGTGGTTGACC	TACACAGGTGCCATCAAATTGGATGACAAAGATCCA	TGGATCTTTGTCATCCAATTTGATGGCAAAGTTGTGTAAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
17	29308	TGGATGACAAAGATCCAAAT	TTCAAAGATCAAGTCATTTTGCTGAATAAGCATATT	AATATGCTTATTCAGCAAAATGACTTGAAAGTTTTGAAAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
18	29325	AATTTCAAAGATCAAGTCAT	TTTGCTGAATAAGCATATTGACGCATACAAAACATT	AATGTTTTGTATGCGTCAATATGCTTATAAGTAGCAAAAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
19	29338	AAGTCATTTTGCTGAATAAG	CATATTGACGCATACAAAACATTCCCACCAACAGAG	CTCTGTTGGTGGGAATGTTTTGTATGCGAAGTAATATGAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT
20	29398	AAAAGGACAAAAAGAAGAAG	GCTGATGAAACTCAAGCCTTACCGCAGAGACAGAAG	CTTCTGTCTCTGCGGTAAGGCTTGAGTTAAGTATCAGCAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTT

Table D-2: CRISDA primers and target gBlock.

Name	**Sequence**
F7 primer	CCAGCAAACCTATCACCGAGGCTGAGGACCGTCACCACCACGAAATTGGGGTATGCACGTAACA
R primer	GCTGAAAACTCGAGGGAACTGCTGAGGTACTTCCAATACCCAACGCC
Target DNA	TCTTAAACTCACGGCAGCTGggaTTTACCGTCACCACCACGAAATTGGGGTATGCACGTAACATGGAGGGGATAGGCATACTTGTCCAGCAATTCCTCTCAGAAATCGCCATACCCAAAACCGGTGAAGAGAAAGGTGTCCGAGGCGTTGGGTATTGGAAGTA

Appendix E

Glossary

16S rRNA. The RNA component of the 30S ribosomal subunit in prokaryotes, the 3’ end of which recognizes and hybridizes with the Shine-Dalgarno sequence of a bacterial mRNA.

30S pre-initiation complex (30S PIC). A transient molecular assembly formed during translation initiation, formation of which is thought to be rate-limiting to overall translation, consisting of the 30S ribosomal subunit, mRNA, initiator tRNA (fMet-tRNA), and initiation factors 1, 2, and 3.

30S ribosomal subunit. The smaller of the 2 ribosomal subunits (30S and 50S). The 30S recognizes and binds to the ribosome binding site, the first step of translation.

5’ untranslated region (5’ UTR). The sequence between the 5’ end of an mRNA and the beginning of a protein coding sequence, sequence and structural elements of which regulate the translation initiation rate.

Aptamer. A relatively short (generally less than 100 nucleotide) single-stranded DNA or RNA sequence which base-pairs with itself to form a structure that specifically interacts with and binds a ligand.

Activation ratio (AR). The fold-enhancement of signal from a riboswitch when its ligand is added.

Biological recognition element. A specific sequence or structure of protein or nucleic acid that strongly and specifically interacts with a target ligand.

Base pair (bp). Unit of length of double-stranded nucleic acid.

Cas12a. A Cas nuclease that binds and cleaves double-stranded DNA, activating nonspecific endonuclease activity which leads to it cutting other DNA strands in solution with it.

Cas13. A Cas nuclease that binds and cleaves RNA, activating nonspecific endonuclease activity which leads to it cutting other RNA strands in solution with it.

Cas9. A Cas nuclease that complexes with a sgRNA or dgRNA and is guided to bind to and cleave a specific DNA sequence.

Coding sequence (CDS). The portion of DNA or mRNA that encodes a specific protein.

Cutting frequency determination (CFD). Model prediction for Cas9 binding or cutting activity against off-target DNA sequences.

Cosolute. An additional solute present in solution with a solute of interest, addition of which may affect the solute of interest.

C-reactive protein (CRP). A human protein whose levels are associated with inflammation. Can be found in its native state (pentameric, or pCRP) or monomeric/modified (mCRP).

CRISPR–Cas9-triggered nicking endonuclease-mediated Strand Displacement Amplification (CRISDA). An isothermal DNA amplification technique where strand displacement amplification is triggered by 2 Cas9 nicking events.

CRISPR. The family of RNA-guided endonucleases, originally identified in prokaryotic genomes as clustered regularly interspaced short palindromic repeats (CRISPR arrays) and CRISPR-associated proteins (Cas proteins).

CRISPRi. A system to regulate transcription of specific genes with dCas9 DNA binding.

crRNA. One of two RNAs in the native CRISPR-Cas9 system, which consists of the guide sequence and the first part of the handle sequence.

dCas9. A catalytically inactive version of Cas9, which can bind to its DNA target but not cut it.

dgRNA. Dual-guide RNA, consisting of crRNA hybridized to tracrRNA.

***E. coli* DH10B.** A specific strain of the bacterium *E. coli*, used for cloning genetic constructs.

ELISA. Enzyme-linked immunosorbent assay. Assay to detect specific target compounds, often proteins, using antibodies specific for those targets.

EMSA. Electrophoretic mobility shift assay. A technique to assay the degree of macromolecular binding to nucleic acids by observing the difference in electrophoretic mobility between the bound and unbound forms.

Ficoll-400. An approximately 400 kDa branched sucrose polymer.

Genetically-encoded sensor. A protein or nucleic acid that binds to a specific target and transduces the signal to a change in gene expression.

Guide sequence. The first 20 nucleotides of a crRNA or sgRNA, the sequence of which determines the DNA sequence Cas9 will recognize and bind or cut.

Handle sequence. The non-guide sequence of the dgRNA or sgRNA, which is specifically recognized and bound by Cas9.

Huggins equation. Relates the specific viscosity (n_s) of a polymer solution to the polymer's concentration (c). The Huggins coefficient (k_H) is determined empirically, but typically ranges from 0.3 to 0.5.

$$\left[\frac{n_s}{c}\right] = [n] + k_H[n]^2 c$$

Hybridization. Base-pairing of two complementary nucleic acid strands.

Intereukin-32γ (IL32γ). The γ isoform of interleukin-32, a proinflammatory cytokine.

K_d. The dissociation constant of an association equilibria, a commonly-used metric of binding affinity.

Klenow factor (KF). A strand-displacing DNA polymerase formed from enzymatic cleavage of *E. coli* DNA polymerase I.

LC-MS/MS. An analytical technique combining liquid chromatography (LC) to separate compounds in a sample with tandem mass spectrometry (MS/MS) to analyze the composition of each compound.

Linear discriminant analysis (LDA). A classification technique which can be used to predict which features result in a specific outcome.

Ligand. A compound which is specifically bound by a biological recognition element.

mRFP1. A monomeric red fluorescent protein (RFP), which is commonly used as a reporter protein to determine gene expression levels from different genetic regulatory elements.

mRNA. Messenger RNA, which is transcribed from DNA and encodes a protein.

MS2. A bacteriophage which infects *E. coli* and utilizes a strong interaction between an RNA hairpin and capsid protein for regulating its expression in the cell and assembling the viral particle.

Nb.BbvCI. A nicking endonuclease, which specifically cuts one strand of DNA at its target sequence.

Nucleotides (nt). A unit of length for single-stranded nucleic acids.

N-terminal. Referring to the part of the translated protein closest to the initially-translated region.

Nucleic acid secondary structure prediction. Algorithms that predict the base-paired structure and free energy of a nucleic acid sequence.

Protospacer-adjacent motif (PAM). An additional specific DNA sequence adjacent to the target sequence that is required for Cas9 recognition of the target DNA.

PEG-8000. A linear polymer of polyethylene glycol which is approximately 8000 Da.

Prokaryote. Bacterial or archaeal organisms.

Repeat:antirepeat (R:AR). A region of the sgRNA or dgRNA handle sequence which is formed by crRNA:tracrRNA hybridization or sgRNA intramolecular base pairing.

Ribosome binding site (RBS). The portion of the 5'UTR and N-terminal CDS that interacts with the 30S ribosomal subunit.

RBS Calculator. A sequence-to-function biophysical model of translation initiation that predicts the translation rate of a gene based on its 5'UTR and CDS sequences.

Riboswitch. A 5'UTR that contains an aptamer and changes its secondary structure, and therefore its translation initiation rate, when the aptamer binds its cognate ligand.

Riboswitch Calculator. A sequence-to-function model of riboswitch function that uses the RBS Calculator model to predict the translation initiation rate of an aptamer-containing 5'UTR, with and without the ligand present.

RNaseIII. A riboendonuclease from *E. coli* that cleaves RNA duplexes greater than 20 base pairs long.

Repression ratio (RR). The fold-reduction of signal from a riboswitch when its ligand is added.

RT-qPCR. An ultrasensitive nucleic acid quantitation technique that uses reverse transcription to produce complementary DNA (cDNA) from an RNA template, followed by PCR to amplify the cDNA and fluorescent probes that bind to the amplicon product.

***Streptococcus pyogenes*.** A bacterium whose CRISPR-Cas9 system was the first to be extensively characterized.

SARS-CoV2 / COVID-19. A novel coronavirus and its corresponding disease that was the causative agent of the 2020-2021 pandemic.

Scaled particle theory (SPT). An equation of state describing the increase in chemical potential of a molecule when particles of a specific size and geometry are added that occupy volume in the system.

Strand displacement amplification (SDA). An isothermal DNA amplification technique that uses a nicking endonuclease to nick a strand of DNA, and a strand-displacing DNA polymerase to extend from the nick.

SELEX. An evolutionary aptamer discovery technique that begins with a large, randomized pool of single-stranded nucleic acids and selects those with the highest binding affinity over multiple rounds of selection against a specific target.

sgRNA. Single-guide RNA, consisting of the original crRNA and tracrRNA incorporated into one molecule.

Shine-Dalgarno sequence (SD sequence). A 9-nucleotide sequence in a ribosome binding site that directly base-pairs with the 3' end of the 16S rRNA.

Signal transduction. The process by which an input sensing event is turned into an output.

Stem loop 1 (SL1). A structural motif in an sgRNA handle sequence.

Stem loop 2 (SL2). A structural motif in an sgRNA handle sequence.

Stokes-Einstein equation. An equation describing the relationship of the diffusion coefficient (D) of a molecule in solution with the temperature (T), the dynamic viscosity (η), and the particle's radius (r).

$$D = \frac{k_B T}{6\pi\eta r}$$

T7 *in vitro* transcription. A commonly-used method to produce large amounts of RNA *in vitro*, using the highly-processive T7 RNA polymerase, which transcribes RNA from a DNA template containing the T7 promoter sequence.

tracrRNA. The second RNA in the native CRISPR-Cas9 system, which base-pairs with the crRNA and consists of the second part of the handle sequence.

TX-TL. A cell-free coupled transcription-translation system. It consists of a crude *E. coli* cell extract containing expression machinery and central metabolic enzymes, an energy solution, an amino acid solution, and added template DNA to express protein.

XRN-1. A 5'->3' riboexonuclease that initiates at 5' monophosphate.

References

1. Rajapaksha P, Elbourne A, Gangadoo S, Brown R, Cozzolino D, Chapman J. A review of methods for the detection of pathogenic microorganisms. *Analyst* **144**, 396-411 (2019).

2. Nadim F, Hoag GE, Liu S, Carley RJ, Zack P. Detection and remediation of soil and aquifer systems contaminated with petroleum products: an overview. *Journal of Petroleum Science and Engineering* **26**, 169-178 (2000).

3. Simoni RD, Hill RL, Vaughan M, Tabor H. A Classic Instrument: The Beckman DU Spectrophotometer and Its Inventor, Arnold O. Beckman. *Journal of Biological Chemistry* **278**, 79-81 (2003).

4. Morris PJT, Royal Society of C, Science M, Chemical Heritage F, Conference on from the Test-tube to the Autoanalyzer: the Development of Chemical Instrumentation in the Twentieth C. From classical to modern chemistry : the instrumental revolution. (ed^(eds). Royal Society of Chemistry in association with the Science Museum, and the Chemical Heritage Foundation.

5. Fung AWS, Sugumar V, Ren AH, Kulasingam V. Emerging role of clinical mass spectrometry in pathology. *J Clin Pathol* **73**, 61-69 (2020).

6. Yalow RS, Berson SA. Assay of Plasma Insulin in Human Subjects by Immunological Methods. *Nature* **184**, 1648-1649 (1959).

7. Hamilton. Hamilton ELISA NIMBUS | View the automated ELISA workstation. (ed^(eds) (2021).

8. Holmes JC, Morrell FA. Oscillographic Mass Spectrometric Monitoring of Gas Chromatography. *Applied Spectroscopy* **11**, 86-87 (1957).

9. Hashimoto Y. Development of a Miniature Mass Spectrometer and an Automated Detector for Sampling Explosive Materials. *Mass Spectrom (Tokyo)* **6**, A0054 (2017).

10. Herzog S, Reth M, Jumaa H. Regulation of B-cell proliferation and differentiation by pre-B-cell receptor signalling. *Nat Rev Immunol* **9**, 195-205 (2009).

11. Turner SJ, Doherty PC, McCluskey J, Rossjohn J. Structural determinants of T-cell receptor bias in immunity. *Nat Rev Immunol* **6**, 883-894 (2006).

12. Peng HP, Lee KH, Jian JW, Yang AS. Origins of specificity and affinity in antibody-protein interactions. *Proc Natl Acad Sci U S A* **111**, E2656-2665 (2014).

13. Ulrich LE, Koonin EV, Zhulin IB. One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol* **13**, 52-56 (2005).

14. Ulrich LE, Zhulin IB. The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Res* **38**, D401-407 (2010).

15. Koonin EV, Makarova KS. Origins and evolution of CRISPR-Cas systems. *Philos Trans R Soc Lond B Biol Sci* **374**, 20180087 (2019).

16. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res* **43**, D298-299 (2015).

17. Sun ZZ, Hayes CA, Shin J, Caschera F, Murray RM, Noireaux V. Protocols for implementing an Escherichia coli based TX-TL cell-free expression system for synthetic biology. *J Vis Exp*, e50762 (2013).

18. Campbell RE, *et al.* A monomeric red fluorescent protein. *Proc Natl Acad Sci U S A* **99**, 7877-7882 (2002).

19. England CG, Ehlerding EB, Cai W. NanoLuc: A Small Luciferase Is Brightening Up the Field of Bioluminescence. *Bioconjugate Chemistry* **27**, 1175-1187 (2016).

20. Wang X, Gao N, Zhou Q. Concentration responses of toxicity sensor with Shewanella oneidensis MR-1 growing in bioelectrochemical systems. *Biosens Bioelectron* **43**, 264-267 (2013).

21. Fernandez-Lopez R, Ruiz R, de la Cruz F, Moncalian G. Transcription factor-based biosensors enlightened by the analyte. *Front Microbiol* **6**, 648 (2015).

22. Tang DJ, *et al.* A SAM-I riboswitch with the ability to sense and respond to uncharged initiator tRNA. *Nat Commun* **11**, 2794 (2020).

23. McCormick JW, Pincus D, Resnekov O, Reynolds KA. Strategies for Engineering and Rewiring Kinase Regulation. *Trends Biochem Sci* **45**, 259-271 (2020).

24. McGinn J, Marraffini LA. Molecular mechanisms of CRISPR-Cas spacer acquisition. *Nat Rev Microbiol* **17**, 7-12 (2019).

25. team TA. AlphaFold: a solution to a 50-year old grand challenge in biology | DeepMind. (ed^(eds).

26. L FMM, Currin A, Dixon N. Directed evolution of the PcaV allosteric transcription factor to generate a biosensor for aromatic aldehydes. *J Biol Eng* **13**, 91 (2019).

27. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505-510 (1990).

28. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818-822 (1990).

29. Winkler W, Nahvi A, Breaker RR. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* **419**, 952-956 (2002).

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

31. Shabalina SA, Koonin EV. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol* **23**, 578-587 (2008).

32. Olina AV, Kulbachinskiy AV, Aravin AA, Esyunina DM. Argonaute Proteins and Mechanisms of RNA Interference in Eukaryotes and Prokaryotes. *Biochemistry (Mosc)* **83**, 483-497 (2018).

33. Bayat P, *et al.* SELEX methods on the road to protein targeting with nucleic acid aptamers. *Biochimie* **154**, 132-155 (2018).

34. Bashir A, *et al.* Machine learning guided aptamer refinement and discovery. *Nat Commun* **12**, 2366 (2021).

35. Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. The Vienna RNA websuite. *Nucleic Acids Res* **36**, W70-74 (2008).

36. Espah Borujeni A, Mishler DM, Wang J, Huso W, Salis HM. Automated physics-based design of synthetic riboswitches from diverse RNA aptamers. *Nucleic Acids Res* **44**, 1-13 (2016).

37. Moore SJ, *et al.* Rapid acquisition and model-based analysis of cell-free transcription-translation reactions from nonmodel bacteria. *Proc Natl Acad Sci U S A* **115**, E4340-E4349 (2018).

38. Chappell J, Jensen K, Freemont PS. Validation of an entirely in vitro approach for rapid prototyping of DNA regulatory elements for synthetic biology. *Nucleic Acids Res* **41**, 3471-3481 (2013).

39. Senoussi A, *et al.* Quantitative Characterization of Translational Riboregulators Using an in Vitro Transcription-Translation System. *ACS Synth Biol* **7**, 1269-1278 (2018).

40. Lehr FX, *et al.* Cell-Free Prototyping of AND-Logic Gates Based on Heterogeneous RNA Activators. *ACS Synth Biol* **8**, 2163-2173 (2019).

41. Wang H, Li J, Jewett MC. Development of a Pseudomonas putida cell-free protein synthesis platform for rapid screening of gene regulatory elements. *Synthetic Biology* **3**, (2018).

42. Karim AS, Jewett MC. A cell-free framework for rapid biosynthetic pathway prototyping and enzyme discovery. *Metab Eng* **36**, 116-126 (2016).

43. Casini A, *et al.* A Pressure Test to Make 10 Molecules in 90 Days: External Evaluation of Methods to Engineer Biology. *J Am Chem Soc* **140**, 4302-4316 (2018).

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

45. Thavarajah W, Silverman AD, Verosloff MS, Kelley-Loughnane N, Jewett MC, Lucks JB. Point-of-Use Detection of Environmental Fluoride via a Cell-Free Riboswitch-Based Biosensor. *ACS Synth Biol* **9**, 10-18 (2020).

46. Alam KK, *et al.*, (2019).

47. Wen KY, *et al.* A Cell-Free Biosensor for Detecting Quorum Sensing Molecules in P. aeruginosa-Infected Respiratory Samples. *ACS Synth Biol* **6**, 2293-2301 (2017).

48. Zimmerman SB, Trach SO. Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. *J Mol Biol* **222**, 599-620 (1991).

49. Jewett MC, Swartz JR. Mimicking the Escherichia coli cytoplasmic environment activates long-lived and efficient cell-free protein synthesis. *Biotechnol Bioeng* **86**, 19-26 (2004).

50. Kuznetsova IM, Zaslavsky BY, Breydo L, Turoverov KK, Uversky VN. Beyond the excluded volume effects: mechanistic complexity of the crowded milieu. *Molecules* **20**, 1377-1409 (2015).

51. Kontur WS, Capp MW, Gries TJ, Saecker RM, Record MT, Jr. Probing DNA binding, DNA opening, and assembly of a downstream clamp/jaw in Escherichia coli RNA polymerase-lambdaP(R) promoter complexes using salt and the physiological anion glutamate. *Biochemistry* **49**, 4361-4373 (2010).

52. Ge X, Luo D, Xu J. Cell-free protein expression under macromolecular crowding conditions. *PLoS One* **6**, e28707 (2011).

53. Moriizumi Y, Tabata KV, Miyoshi D, Noji H. Osmolyte-Enhanced Protein Synthesis Activity of a Reconstituted Translation System. *ACS Synth Biol* **8**, 557-567 (2019).

54. Shin J, Noireaux V. An E. coli cell-free expression toolbox: application to synthetic gene circuits and artificial cells. *ACS Synth Biol* **1**, 29-41 (2012).

55. Espah Borujeni A, Channarasappa AS, Salis HM. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res* **42**, 2646-2659 (2014).

56. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol* **27**, 946-950 (2009).

57. Kai L, Dotsch V, Kaldenhoff R, Bernhard F. Artificial environments for the co-translational stabilization of cell-free expressed proteins. *PLoS One* **8**, e56637 (2013).

58. Karim AS, Heggestad JT, Crowe SA, Jewett MC. Controlling cell-free metabolism through physiochemical perturbations. *Metab Eng* **45**, 86-94 (2018).

59. Espah Borujeni A, Salis HM. Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism. *J Am Chem Soc* **138**, 7016-7023 (2016).

60. Reis AC, Salis HM. An Automated Model Test System for Systematic Development and Improvement of Gene Expression Models. *ACS Synth Biol* **9**, 3145-3156 (2020).

61. Farasat I, Kushwaha M, Collens J, Easterbrook M, Guido M, Salis HM. Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. *Mol Syst Biol* **10**, 731 (2014).

62. Niess A, Siemann-Herzberg M, Takors R. Protein production in Escherichia coli is guided by the trade-off between intracellular substrate availability and energy cost. *Microb Cell Fact* **18**, 8 (2019).

63. Klumpp S, Scott M, Pedersen S, Hwa T. Molecular crowding limits translation and cell growth. *Proc Natl Acad Sci U S A* **110**, 16754-16759 (2013).

64. Dinç CÖ, Kibarer G, Güner A. Solubility profiles of poly(ethylene glycol)/solvent systems. II. comparison of thermodynamic parameters from viscosity measurements. *Journal of Applied Polymer Science* **117**, 1100-1119 (2010).

65. Borg A, Ehrenberg M. Determinants of the rate of mRNA translocation in bacterial protein synthesis. *J Mol Biol* **427**, 1835-1847 (2015).

66. Yamagami R, Bingaman JL, Frankel EA, Bevilacqua PC. Cellular conditions of weakly chelated magnesium ions strongly promote RNA stability and catalysis. *Nat Commun* **9**, 2149 (2018).

67. Wohlgemuth I, Pohl C, Rodnina MV. Optimization of speed and accuracy of decoding in translation. *EMBO J* **29**, 3701-3709 (2010).

68. Lebowitz JL, Helfand E, Praestgaard E. Scaled Particle Theory of Fluid Mixtures. *The Journal of Chemical Physics* **43**, 774-779 (1965).

69. Ling K, Jiang H, Zhang Q. A colorimetric method for the molecular weight determination of polyethylene glycol using gold nanoparticles. *Nanoscale Res Lett* **8**, 538 (2013).

70. Allen SH, Wong KP. The hydrodynamic and spectroscopic properties of 16 S RNA from Escherichia coli ribosome in reconstitution buffer. *J Biol Chem* **253**, 8759-8766 (1978).

71. Gabler R, Westhead EW, Ford NC. Studies of ribosomal diffusion coefficients using laser light-scattering spectroscopy. *Biophys J* **14**, 528-545 (1974).

72. Werner A. Predicting translational diffusion of evolutionary conserved RNA structures by the nucleotide number. *Nucleic Acids Res* **39**, e17 (2011).

73. Sullivan CJ, *et al.* A cell-free expression and purification process for rapid production of protein biologics. *Biotechnol J* **11**, 238-248 (2016).

74. Stark JC, *et al.* On-demand, cell-free biomanufacturing of conjugate vaccines at the point-of-care. *bioRxiv*, 681841 (2019).

75. Fritz BR, Jamil OK, Jewett MC. Implications of macromolecular crowding and reducing conditions for in vitro ribosome construction. *Nucleic Acids Res* **43**, 4774-4784 (2015).

76. Rustad M, Eastlund A, Jardine P, Noireaux V. Cell-free TXTL synthesis of infectious bacteriophage T4 in a single test tube reaction. *Synthetic Biology* **3**, (2018).

77. Jung JK, *et al.* Cell-free biosensors for rapid detection of water contaminants. *Nat Biotechnol* **38**, 1451-1459 (2020).

78. Takahashi MK, *et al.* Characterizing and prototyping genetic networks with cell-free transcription-translation reactions. *Methods* **86**, 60-72 (2015).

79. de los Santos EL, Meyerowitz JT, Mayo SL, Murray RM. Engineering Transcriptional Regulator Effector Specificity Using Computational Design and In Vitro Rapid Prototyping: Developing a Vanillin Sensor. *ACS Synth Biol* **5**, 287-295 (2016).

80. Shin J, Noireaux V. Efficient cell-free expression with the endogenous E. Coli RNA polymerase and sigma factor 70. *J Biol Eng* **4**, 8 (2010).

81. Pardee K, *et al.* Portable, On-Demand Biomolecular Manufacturing. *Cell* **167**, 248-259 e212 (2016).

82. Tjorve KMC, Tjorve E. A proposed family of Unified models for sigmoidal growth. *Ecol Model* **359**, 117-127 (2017).

83. Tsao CY, Hooshangi S, Wu HC, Valdes JJ, Bentley WE. Autonomous induction of recombinant proteins by minimally rewiring native quorum sensing regulon of E. coli. *Metab Eng* **12**, 291-297 (2010).

84. Zhang W, Ames BD, Tsai SC, Tang Y. Engineered biosynthesis of a novel amidated polyketide, using the malonamyl-specific initiation module from the oxytetracycline polyketide synthase. *Appl Environ Microbiol* **72**, 2573-2580 (2006).

85. Noguchi N, Takada K, Katayama J, Emura A, Sasatsu M. Regulation of transcription of the mph(A) gene for macrolide 2'-phosphotransferase I in Escherichia coli: characterization of the regulatory gene mphR(A). *J Bacteriol* **182**, 5052-5058 (2000).

86. Branco R, Cristovao A, Morais PV. Highly sensitive, highly specific whole-cell bioreporters for the detection of chromate in environmental samples. *PLoS One* **8**, e54005 (2013).

87. King JM, *et al.* Rapid, sensitive bioluminescent reporter technology for naphthalene exposure and biodegradation. *Science* **249**, 778-781 (1990).

88. Rogers JK, Guzman CD, Taylor ND, Raman S, Anderson K, Church GM. Synthetic biosensors for precise gene control and real-time monitoring of metabolites. *Nucleic Acids Res* **43**, 7648-7660 (2015).

89. Green AA, Silver PA, Collins JJ, Yin P. Toehold switches: de-novo-designed regulators of gene expression. *Cell* **159**, 925-939 (2014).

90. Chappell J, Westbrook A, Verosloff M, Lucks JB. Computational design of small transcription activating RNAs for versatile and dynamic gene regulation. *Nat Commun* **8**, 1051 (2017).

91. Davidson ME, Harbaugh SV, Chushak YG, Stone MO, Kelley-Loughnane N. Development of a 2,4-dinitrotoluene-responsive synthetic riboswitch in E. coli cells. *ACS Chem Biol* **8**, 234-241 (2013).

92. Joe MH, *et al.* Pigment-based whole-cell biosensor system for cadmium detection using genetically engineered Deinococcus radiodurans. *Bioprocess Biosyst Eng* **35**, 265-272 (2012).

93. Kylilis N, *et al.* Whole-Cell Biosensor with Tunable Limit of Detection Enables Low-Cost Agglutination Assays for Medical Diagnostic Applications. *ACS Sens* **4**, 370-378 (2019).

94. Riangrungroj P, Bever CS, Hammock BD, Polizzi KM. A label-free optical whole-cell Escherichia coli biosensor for the detection of pyrethroid insecticide exposure. *Sci Rep* **9**, 12466 (2019).

95. Shaw WM, *et al.* Engineering a Model Cell for Rational Tuning of GPCR Signaling. *Cell* **177**, 782-796 e727 (2019).

96. Pardee K, *et al.* Paper-based synthetic gene networks. *Cell* **159**, 940-954 (2014).

97. Voyvodic PL, *et al.* Plug-and-play metabolic transducers expand the chemical detection space of cell-free biosensors. *Nat Commun* **10**, 1697 (2019).

98. Iyer S, Doktycz MJ. Thrombin-mediated transcriptional regulation using DNA aptamers in DNA-based cell-free protein synthesis. *ACS Synth Biol* **3**, 340-346 (2014).

99. Byun JY, Lee KH, Shin YB, Kim DM. Cascading Amplification of Immunoassay Signal by Cell-Free Expression of Firefly Luciferase from Detection Antibody-Conjugated DNA in an Escherichia coli Extract. *ACS Sens* **4**, 93-99 (2019).

100. Leca-Bouvier B, Blum LJ. Biosensors for Protein Detection: A Review. *Analytical Letters* **38**, 1491-1517 (2005).

101. Lequin RM. Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). *Clin Chem* **51**, 2415-2418 (2005).

102. Jannetto PJ, Fitzgerald RL. Effective Use of Mass Spectrometry in the Clinical Laboratory. *Clin Chem* **62**, 92-98 (2016).

103. Levine MZ, Gregorio NE, Jewett MC, Watts KR, Oza JP. Escherichia coli-Based Cell-Free Protein Synthesis: Protocols for a robust, flexible, and accessible platform technology. *J Vis Exp*, (2019).

104. Pardee K. Perspective: Solidifying the impact of cell-free synthetic biology through lyophilization. *Biochem Eng J* **138**, 91-97 (2018).

105. Dunn MR, Jimenez RM, Chaput JC. Analysis of aptamer discovery and technology. *Nature Reviews Chemistry* **1**, 0076 (2017).

106. Etzel M, Morl M. Synthetic Riboswitches: From Plug and Pray toward Plug and Play. *Biochemistry* **56**, 1181-1198 (2017).

107. Silverman AD, Kelley-Loughnane N, Lucks JB, Jewett MC. Deconstructing Cell-Free Extract Preparation for in Vitro Activation of Transcriptional Genetic Circuitry. *ACS Synth Biol* **8**, 403-414 (2019).

108. Chushak Y, Harbaugh S, Zimlich K, Alfred B, Chávez J, Kelley-Loughnane N. Characterization of synthetic riboswitch in cell-free protein expression systems. *RNA Biology*, 1-12 (2021).

109. Espah Borujeni A, Channarasappa AS, Salis HM. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Research* **42**, 2646-2659 (2014).

110. Kensch O, Connolly BA, Steinhoff H-J, McGregor A, Goody RS, Restle T. HIV-1 Reverse Transcriptase-Pseudoknot RNA Aptamer Interaction Has a Binding Affinity in the Low Picomolar Range Coupled with High Specificity *. *Journal of Biological Chemistry* **275**, 18271-18278 (2000).

111. Wang MS, Black JC, Knowles MK, Reed SM. C-reactive protein (CRP) aptamer binds to monomeric but not pentameric form of CRP. *Anal Bioanal Chem* **401**, 1309-1318 (2011).

112. Cai Q, *et al.* A simplified and robust protocol for immunoglobulin expression in Escherichia coli cell-free protein synthesis systems. *Biotechnol Prog* **31**, 823-831 (2015).

113. Siegal-Gaskins D, Tuza ZA, Kim J, Noireaux V, Murray RM. Gene circuit performance characterization and resource usage in a cell-free "breadboard". *ACS Synth Biol* **3**, 416-425 (2014).

114. Takahashi MK, *et al.* Rapidly characterizing the fast dynamics of RNA genetic circuitry with cell-free transcription-translation (TX-TL) systems. *ACS Synth Biol* **4**, 503-515 (2015).

115. Valegard K, Murray JB, Stockley PG, Stonehouse NJ, Liljas L. Crystal structure of an RNA bacteriophage coat protein-operator complex. *Nature* **371**, 623-626 (1994).

116. Lowary PT, Uhlenbeck OC. An RNA mutation that increases the affinity of an RNA-protein interaction. *Nucleic Acids Res* **15**, 10483-10493 (1987).

117. Katz N, *et al.* Synthetic 5' UTRs Can Either Up- or Downregulate Expression upon RNA-Binding Protein Binding. *Cell Syst* **9**, 93-106 e108 (2019).

118. Black S, Kushner I, Samols D. C-reactive Protein. *J Biol Chem* **279**, 48487-48490 (2004).

119. Slevin M, *et al.* Monomeric C-reactive protein--a key molecule driving development of Alzheimer's disease associated with brain ischaemia? *Sci Rep* **5**, 13281 (2015).

120. Eisenhardt SU, *et al.* Dissociation of pentameric to monomeric C-reactive protein on activated platelets localizes inflammation to atherosclerotic plaques. *Circ Res* **105**, 128-137 (2009).

121. Khreiss T, Jozsef L, Potempa LA, Filep JG. Conformational rearrangement in C-reactive protein is required for proinflammatory actions on human endothelial cells. *Circulation* **109**, 2016-2022 (2004).

122. Zouki C, Haas B, Chan JS, Potempa LA, Filep JG. Loss of pentameric symmetry of C-reactive protein is associated with promotion of neutrophil-endothelial cell adhesion. *J Immunol* **167**, 5355-5361 (2001).

123. Heinhuis B, Netea MG, van den Berg WB, Dinarello CA, Joosten LA. Interleukin-32: a predominantly intracellular proinflammatory mediator that controls cell activation and cell death. *Cytokine* **60**, 321-327 (2012).

124. El-Far M, *et al.* Proinflammatory isoforms of IL-32 as novel and robust biomarkers for control failure in HIV-infected slow progressors. *Sci Rep* **6**, 22902 (2016).

125. Gui M, Zhang H, Zhong K, Li Y, Sun J, Wang L. Clinical significance of interleukin-32 expression in patients with rheumatoid arthritis. *Asian Pac J Allergy Immunol* **31**, 73-78 (2013).

126. Kim S-H, *et al.* Generation of Antagonistic RNA Aptamers Specific to Proinflammatory Cytokine Interleukin-32. *Bulletin of the Korean Chemical Society* **31**, 3561-3566 (2010).

127. Huttenhofer A, Noller HF. Footprinting mRNA-ribosome complexes with chemical probes. *EMBO J* **13**, 3892-3901 (1994).

128. Moore SJ, *et al.* Rapid acquisition and model-based analysis of cell-free transcription–translation reactions from nonmodel bacteria. *Proceedings of the National Academy of Sciences* **115**, E4340 (2018).

129. Underwood KA, Swartz JR, Puglisi JD. Quantitative polysome analysis identifies limitations in bacterial cell-free protein synthesis. *Biotechnol Bioeng* **91**, 425-435 (2005).

130. Auslander S, Stucheli P, Rehm C, Auslander D, Hartig JS, Fussenegger M. A general design strategy for protein-responsive riboswitches in mammalian cells. *Nat Methods* **11**, 1154-1160 (2014).

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

132. Saito H, Fujita Y, Kashida S, Hayashi K, Inoue T. Synthetic human cell fate regulation by protein-driven RNA switches. *Nat Commun* **2**, 160 (2011).

133. Kennedy AB, Vowles JV, d'Espaux L, Smolke CD. Protein-responsive ribozyme switches in eukaryotic cells. *Nucleic Acids Res* **42**, 12306-12321 (2014).

134. Aptagen L. Apta-Index™ (Aptamer Database) - Library of 500+ Aptamers. (ed^(eds).

135. Ngubane NA, Gresh L, Pym A, Rubin EJ, Khati M. Selection of RNA aptamers against the M. tuberculosis EsxG protein using surface plasmon resonance-based SELEX. *Biochem Biophys Res Commun* **449**, 114-119 (2014).

136. Pan Q, *et al.* Aptamers That Preferentially Bind Type IVB Pili and Inhibit Human Monocytic-Cell Invasion by Salmonella enterica Serovar Typhi. *Antimicrobial Agents and Chemotherapy* **49**, 4052 (2005).

137. Berglund JA, Charpentier B, Rosbash M. A high affinity binding site for the HIV-1 nucleocapsid protein. *Nucleic Acids Res* **25**, 1042-1049 (1997).

138. Gopinath SCB, *et al.* An RNA aptamer that distinguishes between closely related human influenza viruses and inhibits haemagglutinin-mediated membrane fusion. *Journal of General Virology* **87**, 479-487 (2006).

139. Hirao I, Madin K, Endo Y, Yokoyama S, Ellington AD. RNA Aptamers That Bind to and Inhibit the Ribosome-inactivating Protein, Pepocin *. *Journal of Biological Chemistry* **275**, 4943-4948 (2000).

140. Chang TW, Janardhanan P, Mello CM, Singh BR, Cai S. Selection of RNA Aptamers Against Botulinum Neurotoxin Type A Light Chain Through a Non-Radioactive Approach. *Appl Biochem Biotechnol* **180**, 10-25 (2016).

141. Jellinek D, Lynott CK, Rifkin DB, Janjic N. High-affinity RNA ligands to basic fibroblast growth factor inhibit receptor binding. *Proc Natl Acad Sci U S A* **90**, 11227-11231 (1993).

142. Zhai L, Wang T, Kang K, Zhao Y, Shrotriya P, Nilsen-Hamilton M. An RNA aptamer-based microcantilever sensor to detect the inflammatory marker, mouse lipocalin-2. *Anal Chem* **84**, 8763-8770 (2012).

143. Moosavian SA, Jaafari MR, Taghdisi SM, Mosaffa F, Badiee A, Abnous K. Development of RNA aptamers as molecular probes for HER2(+) breast cancer study using cell-SELEX. *Iran J Basic Med Sci* **18**, 576-586 (2015).

144. Ling K, *et al.* A self-assembling RNA aptamer-based nanoparticle sensor for fluorometric detection of Neomycin B in milk. *Anal Bioanal Chem* **408**, 3593-3600 (2016).

145. Santos-Cancel M, White RJ. Collagen Membranes with Ribonuclease Inhibitors for Long-Term Stability of Electrochemical Aptamer-Based Sensors Employing RNA. *Anal Chem* **89**, 5598-5604 (2017).

146. Schoukroun-Barnes LR, Wagan S, White RJ. Enhancing the analytical performance of electrochemical RNA aptamer-based sensors for sensitive detection of aminoglycoside antibiotics. *Anal Chem* **86**, 1131-1137 (2014).

147. Ferapontova EE, Olsen EM, Gothelf KV. An RNA aptamer-based electrochemical biosensor for detection of theophylline in serum. *J Am Chem Soc* **130**, 4256-4258 (2008).

148. Mishler DM, Gallivan JP. A family of synthetic riboswitches adopts a kinetic trapping mechanism. *Nucleic Acids Res* **42**, 6753-6761 (2014).

149. Desai SK, Gallivan JP. Genetic screens and selections for small molecules based on a synthetic riboswitch that activates protein translation. *J Am Chem Soc* **126**, 13247-13254 (2004).

150. Leistra AN, Gelderman G, Sowa SW, Moon-Walker A, Salis HM, Contreras LM. A Canonical Biophysical Model of the CsrA Global Regulator Suggests Flexible Regulator-Target Interactions. *Sci Rep* **8**, 9892 (2018).

151. Babitzke P, Baker CS, Romeo T. Regulation of translation initiation by RNA binding proteins. *Annu Rev Microbiol* **63**, 27-44 (2009).

152. Kuzmanovic DA, Elashvili I, Wick C, O'Connell C, Krueger S. Bacteriophage MS2: molecular weight and spatial distribution of the protein and RNA components by small-angle neutron scattering and virus counting. *Structure* **11**, 1339-1348 (2003).

153. Williams RD, *et al.* Monomeric C-Reactive Protein in Serum With Markedly Elevated CRP Levels Shares Common Calcium-Dependent Ligand Binding Properties With an in vitro Dissociated Form of C-Reactive Protein. *Front Immunol* **11**, 115 (2020).

154. Paige JS, Nguyen-Duc T, Song W, Jaffrey SR. Fluorescence imaging of cellular metabolites with RNA. *Science* **335**, 1194 (2012).

155. Kim H, Jaffrey SR. A Fluorogenic RNA-Based Sensor Activated by Metabolite-Induced RNA Dimerization. *Cell Chemical Biology* **26**, 1725-1731.e1726 (2019).

156. Kundert K, *et al.* Controlling CRISPR-Cas9 with ligand-activated and ligand-deactivated sgRNAs. *Nature Communications* **10**, 2127 (2019).

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

158. Xiang JS, Kaplan M, Dykstra P, Hinks M, McKeague M, Smolke CD. Massively parallel RNA device engineering in mammalian cells with RNA-Seq. *Nat Commun* **10**, 4327 (2019).

159. Strobel B, *et al.* High-throughput identification of synthetic riboswitches by barcode-free amplicon-sequencing in human cells. *Nat Commun* **11**, 714 (2020).

160. Song W, Strack RL, Jaffrey SR. Imaging bacterial protein expression using genetically encoded RNA sensors. *Nat Methods* **10**, 873-875 (2013).

161. Halper SM, Cetnar DP, Salis HM. An Automated Pipeline for Engineering Many-Enzyme Pathways: Computational Sequence Design, Pathway Expression-Flux Mapping, and Scalable Pathway Optimization. *Methods Mol Biol* **1671**, 39-61 (2018).

162. Dominguez AA, Lim WA, Qi LS. Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. *Nature reviews Molecular cell biology* **17**, 5 (2016).

163. Barrangou R, Horvath P. A decade of discovery: CRISPR functions and applications. *Nature microbiology* **2**, 17092 (2017).

164. Halperin SO, Tou CJ, Wong EB, Modavi C, Schaffer DV, Dueber JE. CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature*, 1 (2018).

165. Peters JM, *et al.* Enabling genetic analysis of diverse bacteria with Mobile-CRISPRi. *Nature Microbiology* **4**, 244-250 (2019).

166. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420 (2016).

167. Hess GT, Tycko J, Yao D, Bassik MC. Methods and applications of CRISPR-mediated base editing in eukaryotic genomes. *Molecular cell* **68**, 26-43 (2017).

168. Klann TS, Black JB, Gersbach CA. CRISPR-based methods for high-throughput annotation of regulatory DNA. *Current opinion in biotechnology* **52**, 32-41 (2018).

169. Adamson B, *et al.* A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867-1882. e1821 (2016).

170. Swiech L, *et al.* In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nature biotechnology* **33**, 102-106 (2015).

171. Yao L, Cengic I, Anfelt J, Hudson EP. Multiple gene repression in cyanobacteria using CRISPRi. *ACS synthetic biology* **5**, 207-212 (2015).

172. Zhao Y, *et al.* CRISPR/dCas9-mediated multiplex gene repression in Streptomyces. *Biotechnology journal* **13**, 1800121 (2018).

173. Kim SK, Seong W, Han GH, Lee D-H, Lee S-G. CRISPR interference-guided multiplex repression of endogenous competing pathway genes for redirecting metabolic flux in Escherichia coli. *Microbial cell factories* **16**, 188 (2017).

174. Ordon J, *et al.* Generation of chromosomal deletions in dicotyledonous plants employing a user-friendly genome editing toolkit. *The Plant Journal* **89**, 155-168 (2017).

175. Hughes RA, Ellington AD. Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harbor perspectives in biology* **9**, a023812 (2017).

176. Stapley J, Feulner PG, Johnston SE, Santure AW, Smadja CM. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Phil Trans R Soc B* **372**, 20160455 (2017).

177. Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. *The ISME journal* **3**, 199 (2009).

178. Casini A, *et al.* A pressure test to make 10 molecules in 90 days: External evaluation of methods to engineer biology. *Journal of the American Chemical Society* **140**, 4302-4316 (2018).

179. Najm FJ, *et al.* Orthologous CRISPR–Cas9 enzymes for combinatorial genetic screens. *Nature biotechnology* **36**, 179 (2018).

180. Jack BR, *et al.* Predicting the genetic stability of engineered DNA sequences with the EFM calculator. *ACS synthetic biology* **4**, 939-943 (2015).

181. Brophy JA, Voigt CA. Principles of genetic circuit design. *Nature methods* **11**, 508 (2014).

182. Lovett ST. Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. *Molecular microbiology* **52**, 1243-1253 (2004).

183. Jiang F, *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**, 867-871 (2016).

184. Nishimasu H, *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949 (2014).

185. Jinek M, *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997 (2014).

186. Briner AE, *et al.* Guide RNA functional modules direct Cas9 activity and orthogonality. *Molecular cell* **56**, 333-339 (2014).

187. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).

188. Qi LS, *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-1183 (2013).

189. Zuo Z, Liu J. Allosteric regulation of CRISPR-Cas9 for DNA-targeting and cleavage. *Curr Opin Struct Biol* **62**, 166-174 (2020).

190. Dagdas YS, Chen JS, Sternberg SH, Doudna JA, Yildiz A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Science advances* **3**, eaao0027 (2017).

191. Sternberg SH, LaFrance B, Kaplan M, Doudna JA. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110-113 (2015).

192. Jones SK, Jr., *et al.* Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nat Biotechnol* **39**, 84-93 (2021).

193. Farasat I, Salis HM. A biophysical model of CRISPR/Cas9 activity for rational design of genome editing and gene regulation. *PLoS Computational Biology* **12**, e1004724 (2016).

194. Zhang L, Rube HT, Vakulskas CA, Behlke MA, Bussemaker HJ, Pufall MA. Systematic in vitro profiling of off-target affinity, cleavage and efficiency for CRISPR enzymes. *Nucleic Acids Res* **48**, 5037-5053 (2020).

195. Feng H, Guo J, Wang T, Zhang C, Xing XH. Guide-target mismatch effects on dCas9-sgRNA binding activity in living bacterial cells. *Nucleic Acids Res* **49**, 1263-1277 (2021).

196. Thyme SB, Akhmetova L, Montague TG, Valen E, Schier AF. Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat Commun* **7**, 11750 (2016).

197. Fu Y, *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* **31**, 822-826 (2013).

198. Hsu PD, *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827-832 (2013).

199. Tycko J, *et al.* Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nat Commun* **10**, 4063 (2019).

200. Doench JG, *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature biotechnology*, (2016).

201. Wright AV, *et al.* Rational design of a split-Cas9 enzyme complex. *Proc Natl Acad Sci U S A* **112**, 2984-2989 (2015).

202. Doench JG, *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1262-1267 (2014).

203. Calvo-Villamanan A, *et al.* On-target activity predictions enable improved CRISPR-dCas9 screens in bacteria. *Nucleic Acids Res* **48**, e64 (2020).

204. Kleinstiver BP, *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490-495 (2016).

205. Slaymaker IM, Gao L, Zetsche B, Scott DA, Yan WX, Zhang F. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-88 (2016).

206. Tsai SQ, *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-197 (2015).

207. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32**, 279-284 (2014).

208. Kok Sd, *et al.* Rapid and reliable DNA assembly via ligase cycling reaction. *ACS synthetic biology* **3**, 97-106 (2014).

209. Khlebnikov A, Datsenko KA, Skaug T, Wanner BL, Keasling JD. Homogeneous expression of the PBAD promoter in Escherichia coli by constitutive expression of the low-affinity high-capacity AraE transporter. *Microbiology* **147**, 3241-3247 (2001).

210. Kellner MJ, Koob JG, Gootenberg JS, Abudayyeh OO, Zhang F. SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nat Protoc* **14**, 2986-3012 (2019).

211. Chen JS, *et al.* CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **360**, 436-439 (2018).

212. Broughton JP, *et al.* CRISPR-Cas12-based detection of SARS-CoV-2. *Nat Biotechnol* **38**, 870-874 (2020).

213. Gootenberg JS, *et al.* Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438-442 (2017).

214. Abudayyeh OO, *et al.* C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).

215. Wang D, *et al.* Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci U S A* **99**, 15687-15692 (2002).

216. Popowitch EB, O'Neill SS, Miller MB. Comparison of the Biofire FilmArray RP, Genmark eSensor RVP, Luminex xTAG RVPv1, and Luminex xTAG RVP fast multiplex assays for detection of respiratory viruses. *J Clin Microbiol* **51**, 1528-1533 (2013).

217. Aynaud MM, *et al.* A multiplexed, next generation sequencing platform for high-throughput detection of SARS-CoV-2. *Nat Commun* **12**, 1405 (2021).

218. Mangericao TC, Peng Z, Zhang X. Computational prediction of CRISPR cassettes in gut metagenome samples from Chinese type-2 diabetic patients and healthy controls. *BMC Syst Biol* **10 Suppl 1**, 5 (2016).

219. Marraffini LA. The CRISPR-Cas system of Streptococcus pyogenes: function and applications. In: *Streptococcus pyogenes : Basic Biology to Clinical Manifestations* (ed^(eds Ferretti JJ, Stevens DL, Fischetti VA) (2016).

220. Bradde S, Nourmohammad A, Goyal S, Balasubramanian V. The size of the immune repertoire of bacteria. *Proc Natl Acad Sci U S A* **117**, 5144-5151 (2020).

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

222. Zhou W, Hu L, Ying L, Zhao Z, Chu PK, Yu XF. A CRISPR-Cas9-triggered strand displacement amplification method for ultrasensitive DNA detection. *Nat Commun* **9**, 5012 (2018).

223. Iwasaki RS, Ozdilek BA, Garst AD, Choudhury A, Batey RT. Small molecule regulated sgRNAs enable control of genome editing in E. coli by Cas9. *Nat Commun* **11**, 1394 (2020).

224. Deltcheva E, *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607 (2011).

225. Cody WB, Scholthof HB. Native Processing of Single Guide RNA Transcripts to Create Catalytic Cas9/Single Guide RNA Complexes in Planta. *Plant Physiol* **184**, 1194-1206 (2020).

226. Karvelis T, Gasiunas G, Miksys A, Barrangou R, Horvath P, Siksnys V. crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. *RNA Biol* **10**, 841-851 (2013).

227. Kocak DD, Josephs EA, Bhandarkar V, Adkar SS, Kwon JB, Gersbach CA. Increasing the specificity of CRISPR systems with engineered RNA secondary structures. *Nat Biotechnol* **37**, 657-666 (2019).

228. Kilic T, Weissleder R, Lee H. Molecular and Immunological Diagnostic Tests of COVID-19: Current Status and Challenges. *iScience* **23**, 101406 (2020).

229. Toley BJ, *et al.* Isothermal strand displacement amplification (iSDA): a rapid and sensitive method of nucleic acid amplification for point-of-care diagnosis. *Analyst* **140**, 7540-7549 (2015).

230. Ackerman CM, *et al.* Massively multiplexed nucleic acid detection with Cas13. *Nature* **582**, 277-282 (2020).

VITA

Grace Vezeau

Publications

Vezeau, G. E., Gadila, L. R., & Salis, H. M. Automated design of genetically-encoded protein-sensing riboswitches in a cell-free expression system. *In preparation*.

Cetnar, D., Hossain, A., **Vezeau, G.**, Salis, H.M. (2021). Comprehensive modeling and design of 5′ UTRs for RNA stability using a 62,000 unique 5′ UTR member library. *In preparation*.

Vezeau, G. E. & Salis, H.M. Tuning Cell-free Composition Controls the Time-delay, Dynamics, and Productivity of TX-TL Expression. *Submitted to ACS Synthetic Biology*.

Manzano, I., Taylor, N., Csordas, M., **Vezeau, G.,** Salis, H., Zydney, A. L. (2020). Purification of Cas9-RNA complexes by ultrafiltration. Biotechnology Progress.

Manzano, I., **Vezeau, G.**, Salis, H., & Zydney, A. L. (2020). RNA size and 3-dimensional structure determine ultrafiltration behavior of small RNA molecules. Separation and Purification Technology.

Reis, A. C., Halper, S. M., **Vezeau, G. E.**, Cetnar, D. P., Hossain, A., Clauer, P. R., & Salis, H. M. (2019). Simultaneous repression of multiple bacterial genes using nonrepetitive extra-long sgRNA arrays. Nature Biotechnology.

Tamiev, D., Lantz, A., **Vezeau, G.**, Salis, H., & Reuel, N. F. (2019). Controlling Heterogeneity and Increasing Titer from Riboswitch-Regulated Bacillus subtilis Spores for Time-Delayed Protein Expression Applications. ACS Synthetic Biology.

Walsh, M. J., Goodnow, S. D., **Vezeau, G. E.**, Richter, L. V., & Ahner, B. A. (2015). Cysteine enhances bioavailability of copper to marine phytoplankton. Environmental science & technology.

Presentations

Vezeau, G. E., Gadila, L. R., Salis, H. M. (2021). SEED 2021.

Vezeau, G. E., Salis, H. M. (2019). SEED 2019.

Vezeau, G. E. & Salis, H. M. EBRC Spring Retreat 2018.

Vezeau, G. E. & Salis, H. M. International Workshop on Biodesign Automation 2017.

Vezeau, G. E. & Salis, H. M. EBRC Fall Retreat 2016.

Vezeau, G. E., Montgomery, M., Urli, J., & Salis, H. M. SynBERC Fall Retreat 2015.

Patents

Salis, H.M., Reis, A., Halper, S., **Vezeau, G.**, Cetnar, D., Hossain, A., Clauer, P. R. Engineered CRISPR/Cas9 Systems for Simultaneous Long-term Regulation of Multiple Targets. 206032-0112-00WO. *Application submitted*