Engineering of Synthetic DNA/RNA Modules for Manipulating Gene Expression and Circuit Dynamics

by

Qi Zhang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved October 2020 by the
Graduate Supervisory Committee:

Xiao Wang, Chair
Alexander Green
David Brafman
Xiaojun Tian
Christopher Plaisier

ARIZONA STATE UNIVERSITY

December 2020

ABSTRACT

Gene circuit engineering facilitates the discovery and understanding of fundamental biology and has been widely used in various biological applications. In synthetic biology, gene circuits are often constructed by two main strategies: either monocistronic or polycistronic constructions. The Latter architecture can be commonly found in prokaryotes, eukaryotes, and viruses and has been largely applied in gene circuit engineering. In this work, the effect of adjacent genes and noncoding regions are systematically investigated through the construction of batteries of gene circuits in diverse scenarios. Data-driven analysis yields a protein expression metric that strongly correlates with the features of adjacent transcriptional regions (ATRs). This novel mathematical tool helps the guide for circuit construction and has the implication for the design of synthetic ATRs to tune gene expression, illustrating its potential to facilitate engineering complex gene networks.

The ability to tune RNA dynamics is greatly needed for biotech applications, including therapeutics and diagnostics. Diverse methods have been developed to tune gene expression through transcriptional or translational manipulation. Control of RNA stability/degradation is often overlooked and can be the lightweight alternative to regulate protein yields. To further extend the utility of engineered ATRs to regulate gene expression, a library of RNA modules named degradation-tuning RNAs (dtRNAs) are designed with the ability to form specific 5' secondary structures prior to RBS. These modules can modulate transcript stability while having a minimal interference on translation initiation. Optimization of their functional structural features enables gene expression level to be tuned over a wide dynamic range. These engineered dtRNAs are capable of regulating gene circuit dynamics as well as noncoding RNA levels and can be further expanded into cell-free system for gene expression control *in vitro*. Finally, integrating dtRNA with synthetic toehold sensor enables improved paper-based viral diagnostics, illustrating the potential of using synthetic dtRNAs for biomedical applications.

DEDICATION

To my parents, Ruisha and Xiaoping, the other family members and my dear friends for unconditional encouragement and belief in me.

To my wife Huan, for her infinite support and love.

ACKNOWLEDGMENTS

First and foremost, I am sincerely thankful for my advisor, Dr. Xiao Wang. He gave me the opportunity to work in his laboratory since I was a master student when we first met. His scientific insight, tireless effort and guidance have taught me how to be a qualified scientist, how to deal with scientific issues and how to be a great leader in one project. Under his mentorship, I have grown up from a student who requires constant guidance from other senior students to be a mature scientist that are able to propose ideas and perform planned works independently.

I would also like to thank my committee members Dr. Alexander Green, Dr. David Brafman, Dr. Xiaojun Tian, and Dr. Christopher Plaisier and previous committee members Dr. Samira Kiani and Dr. Mo Ebrahimkhani for their time and support to my works. I sincerely appreciate your insightful comments on the direction of my projects and suggestions toward the difficulties I have ever confronted.

In addition, I would like to thank my previous and current laboratory members, particularly Fuqing Wu, Xingwen Chen, Kylie Standage-Beier, Rong Zhang, Shan Zhu, Zhilong Mi, Ziqi Zhu, Lezhi Wang, David Menn, Riqi Su and Samat BAYAKHMETOV for their help and discussions on all my projects. I would also like to thank my friends from the other labs including Duo Ma, Kaiyue Wu, Jiajie Zhu, Joshua Cutts, Nick Brookhouser, Sreedevi Ramam, Gayathri Srinivasan, Zhaoqing Yan and Stefan Tekel for their encouragements and friendship during these challenging years for which I will never forget.

Lastly, I would like to thank my academic advisor Laura Hawes, School of Biological and Health Systems Engineering, the Graduate Professional Student Association (GPSA) and National Institute of Health (NIH) for their generous support for my work as a Ph.D. student.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

PREFACE

Chapter 2 presented in this PhD dissertation document, has been previously published as describe below:

Wu, F.*, Zhang, Q.* & Wang, X. Design of Adjacent Transcriptional Regions to Tune Gene Expression and Facilitate Circuit Construction. *Cell Syst* **6**, 206-215.e6 (2018).

CHAPTER 1

INTRODUCTION

1.1 Synthetic biology

Synthetic biology is a highly interdisciplinary subject that encompasses a wide range of research areas, including system biology, molecular biology, genetic engineering, chemistry, biophysics, and mathematics. The development of biology and related life science subjects provide us the knowledge and understanding of the basic biological phenomena and mechanisms, enabling the exploration of the hidden interactions among different types of species. An increasing number of biological networks or systems have been discovered based on these interactions which offers the broader stage for scientists to design, characterize and construct artificial biological systems with mimic function using engineering approaches[1–3]. This surely benefits from the development of chemistry and genetic engineering to rapidly synthesize various biological molecules with decreased costs for carrying out synthetic biology research. With the merging of mathematics and biophysics, now scientists are able to build biophysical models to describe or explain different biological systems with predictable manners and utilize for many practical applications in broad areas.

Although the use of the term ‘synthetic biology’ was identified nearly a hundred years ago, it was first appeared in the title of a literature in 1980 by Barbara Hobom to describe the genetically engineered bacteria using DNA recombinant technology[4]. This novel subject has only been under the burgeoning development for five decades which were divided into three time periods by the scientists for thriving, enabling science, modules era and systems era[5]. The first era describes that synthetic biology is grounded on the development of basic molecular biology and genomics. It can be traced back to about 1960s where the discovery of lac operon’s regulation first demonstrated the existing of regulatory gene circuit in natural biological system[6]. This study presented that gene expression can be regulated by the other genes, enables the regulation process to be identified as a dynamic system to produce output while controlled by the behavior of the input protein. In the same decade in 1969, scientists discovered DNA restriction enzymes, illustrating the possibility to cleave and ligate the original DNA sequence to engineer

recombinant gene circuits[7]. The further advent of genetic modification technologies in 1970s and 1980s, including molecular cloning and polymerase chain reaction (PCR), further provide additional enabling technologies to effectively and efficiently construct synthetic gene circuits, and now has become more and more widely used in molecular biology research[8]. During the 1990s, the development of automated DNA sequencing strategy and other genomic technologies enables high-throughput measurement of various types of small molecules including DNA, RNA and protein. This large-scale analysis technique allows scientist to build up libraries of database for each molecular component classified by their specific properties or functions. These large-scale data banks drive the research to understand biology into a broader path through a top-down approach, which later on generates the field of system biology. A large variety of systems, thereby, were created through this approach as scientists start to combine experimental data and computational analysis to reverse-engineer gene regulatory networks (GRNs)[9–11]. Through research, it gradually became clear about the roles and functions of individual biological components that forms a well-organized synthetic system. With the better understanding of these components, a complementary bottom-up approach was established, that is to forward-engineer complex gene networks to implement particular functions using well-characterized biological components. This approach was later termed as the discipline for synthetic biology[12].

The early development for synthetic biology is the starting point of modules era in which biologists created first generation of simple synthetic gene circuits or modules to carry out functions that are analogous to electrical circuits[13,14]. The first two engineered synthetic gene circuits using bottom-up approach was reported in early 2000: the genetic toggle switch and the repressilator[15,16]. By using similar sets of small biological components including inducible promoters, repressors (proteins that inhibit gene expression) and green fluorescence protein (GFP) reporter, circuits' behaviors can be monitored with the stimulation of the inducers. The genetic toggle switch consists of two inducible promoters that drive the expression of mutually inhibitory genes, forming a bistable system that is capable of either maintaining the original system memory or switching to the other state presenting with the respective inducers. This toggle switch principle was well-studied and has been employed for varying applications. One

example is to engineer microbial kill switches to control bacterial population level depending on specific environmental inputs[17]. These tools harness the principle of genetic toggle switch and are readily to be reprogrammed with diverse environmental cues, regulatory topology and killing mechanisms. Another example is to build a bacterial sensing and recording system for environmental stimuli diagnostics in mammalian gut[18]. The repressilator, on the other hand, comprises three genes that mutually inhibit each other, forming a triple negative-feedback loop with their own promoter pairs which, when triggered, lead to the expression of repressor protein in a sequential and periodic manner. Interestingly, both cases incorporated the mathematical modeling to quantitatively define the systems that are fit with experimental outputs, enabling the prediction of systems' outcome by digital inputs without the need to perform biological experiments. Following the research of genetic toggle switch and repressilator, numerous biological modules and simple gene circuits have been engineered with complex architectures and behaviors for diverse purposes. This includes building synthetic gene circuits to study the relationship between gene expression and transcriptional noise in both prokaryotic and eukaryotic cells[19–21], engineering genetic oscillators with different topologies[22–25], developing synthetic bacterial quorum sensing system which regulate circuit behaviors through cell-cell communication[26–28], and other fancy circuits such as genetic counter and timer[29,30], band-pass filter[31], pulse generator[32], optical sensor[33], edge detector[34] and a variety of genetic logical gates[35–37]. The development of these simple circuits and modules paved way for forward-engineering more complex gene networks from prokaryotes to the other highly complicated biological systems.

The transition from modules era to systems era started about ten years ago as synthetic biologists began to focus on engineering layered computational logic and memories, synthesizing scaling-up bacterial consortia and carrying out applications in medicine, biotechnology and environment[12,38,39]. For example, synthetic biosensors were engineered which harness the design principles of logic gate to detect chemical combinations in the environment[18,40]. These sensors contain environmental inducible promoter and corresponding transcriptional factors, when interacting with the chemicals such as metal ions in the environment, releasing orthogonal

outputs which can be individually measured by instruments. Gene regulatory networks has also been integrated into paper-based techniques to develop rapid and low-cost RNA-based technology for viral diagnostic applications[41,42]. Additionally, complex synthetic gene networks were used for various clinic conditions, including the treatment of infectious diseases and cancer therapies[43]. One difficulty for cancer treatment is to engineer tools that can successfully identify cancer cells from nearby normal cells. Despite, to some extent, the general chemotherapy is effective, it indiscriminately kills both cancer cells and normal cells, leading to irreversible tissue damage. Methods that can only target cancer cells through sensing with cancer specific signals are urgently needed. To achieve this, Nissim and colleagues developed a tunable dual-promoter integrator that can specifically kill cancer cells[44]. The two promoters in the integrator can sense signals such as overexpressed oncogenes that only produced by cancer cells and then drive the expression of chimeric proteins, which, when combined, further activate downstream killer gene expression to kill the cells. Normal cells does not contain enough cancerous signals, and therefore survive even with the presence of the integrator. Similar principle was also used in another study in which Xie and colleagues developed RNAi-based logic circuit system to identify and kill specific cancer cell lines[45]. The output killing gene can only be expressed while presented with cancer specific microRNAs to trigger cell apoptosis.

Although keeps making considerable achievements in current stage, additional challenges might still exist to obstacle the further development of synthetic biology. Using bottom-up design approach, it is relatively convenient to characterize a functional module such as the genetic toggle switch or the oscillator in isolation. However, its function may change while integrating into more complex synthetic gene networks, leading to unexpected behavior or system failure. Meanwhile, re-characterize the system typically consuming time and requires iterative design process which hinder the construction of more complex synthetic gene networks with multi-dimensional logic layers. Despite gene expression process shares “Central Dogma” principle among species, circuits with well-characterized behaviors in bacteria might exhibit unwanted interactions with intracellular chassis in the other cell types from bacteria to mammalian cells[46]. Furthermore, expressing exogenous synthetic circuits also competes resource with

endogenous genome expression. This effect could be minor and negligible while introducing smaller synthetic modules. However, the cellular status might be substantially hampered as the increase of circuitry complexity, resulting in elusive behavior that can no longer be negligible[47–49]. Solutions to circumvent these challenges are necessary to provide comprehensive understanding of synthetic gene circuit with robust behaviors in diverse biological systems.

1.2 RNA based regulation of gene expression

The rise of engineering synthetic regulatory modules has made considerable progress in synthetic biology during recent period. By employing certain regulatory topologies, a large variety of dynamics can be achieved through combining well-characterized modules to construct complex synthetic gene networks. In general, most engineered synthetic gene circuits relies on protein-based transcriptional regulation in the early development stage. These systems usually contain a regulatory protein and its cognate inducible promoter which drives the expression of downstream genes. While under expression, this regulatory protein serves as an activator or a repressor that can bind with its promoter DNA sequence to either enhance or inhibit the transcription process of the downstream gene. Two good examples are genetic toggle switch that contains two mutually inhibitory repressors and quorum sensing regulatory system which activates downstream gene expression while binding with cognate ligands[15,26,50]. Now the requirement of engineering complex gene circuits with orthogonal regulations has drawn scientists to focus on developing RNA-based tools for gene expression regulation.

Riboswitch or riboregulator is a noncoding RNA structure that typically locates at the 5' untranslated region (UTR) of a mRNA molecule. This RNA structure is subject to a conformational change while binding with its cognate ligand molecules, such as antisense RNA[51–54], amino acids[55–57] and the other molecules[58,59]. Some regulators such as ribozymes that do not require the specific binding can regulate gene expression through self-cleavage[60–63]. As the intermediate molecules during gene expression process, riboswitch holds the potential for regulation either at transcriptional level by folding or disrupting transcription terminator or at translational level through exposing or sequestering mRNA ribosome binding site (RBS).

One good example of riboswitch regulation at transcriptional level is the engineering of small transcriptional activating RNAs or STARs[64]. This riboswitch usually consists of a cis-regulatory RNA that can form a transcription terminator at the 5' end of mRNA to inhibit transcription process. The addition of STARs can cause the conformational change of this upstream riboswitch, removing its termination effect and thereafter activate downstream gene expression. The effect can be inverted by engineering a cis-attenuator sequence that sequesters the terminator riboswitch. Introduction of STARs that bind with the attenuator sequence completely remove its inhibition effect to release the terminator structure, resulting in gene expression switching from ON state to OFF state. The translational regulatory riboswitch has been optimized for many generations for versatility, less crosstalk and pronounced dynamic ranges. Biologists currently de-novo-designed a riboregulator named toehold switch that, instead of focusing loop-mediated interaction, introduced a linear toehold region that can be designed to recognize arbitrary trigger sequences, favoring the system with improved dynamic range as well as low component crosstalk[65]. Due to the property that only recognizing their cognate trans-regulatory RNAs, STARs and toehold switch have been incorporated to design complex synthetic circuits with multi-input logic gates and have been applied to many biomedical applications[41,42].

The development of clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated (Cas) systems hold the prominent capability for genome engineering and disease treatment[66–70]. As its specific property that the transcribed CRISPR RNA can guide its Cas protein to recognize DNA or RNA sequence through Watson-Crick base pairing to introduce nucleotide cleavage, CRIPSR/Cas system can also serve as a promising tool for gene expression regulation. Qi and colleagues first developed a CRISPR based RNA-guided platform for sequence-specific gene expression regulation in 2013[71]. This CRISPR interference (CRISPRi) system that harnesses a catalytically dead Cas9 (dCas9) protein, when coexpressed with its guide RNA that directs dCas9 binding without cleavage, introduces transcriptional inhibition. An activation version of CRISPR system namely CRISPR activation (CRISPRa) was also developed shortly after the previous publication where biologists engineered dCas9 protein by genetically fusing to a C-terminal VP64 acidic transactivation domain, leading to enhanced

gene expression levels in human cells[72,73]. Currently, diverse systems that shares similar functionality to regulate gene expression have been engineered and expended to multiple organisms from bacteria to mammalian cells, and have been widely used for constructing complex synthetic gene circuits[74–78].

1.3 Chapters in the dissertation

Chapter 2, “Design of Adjacent Transcriptional Regions to Tune Gene Expression and Facilitate Circuit Construction,” mainly discusses the features that can affect gene expression in polycistronic architectures. Here, we constructed synthetic operons with a reporter gene flanked by different ATRs, and found that ATRs with high GC content, small size, and low folding energy lead to high gene expression. Based on these results, a metric of gene expression was built that takes into account ATRs. We used the metric to design and construct logic gates with low basal expression and high sensitivity and nonlinearity. Furthermore, we rationally designed synthetic 5’ ATRs with different GC content and sizes to tune protein expression levels over a 300-fold range and used these to build synthetic toggle switches with varying basal expression and degrees of bistability. Our comprehensive model and gene expression metric could facilitate the future engineering of more complex synthetic gene circuits.

Chapter 3, “Applications of Machine Learning Techniques in Genetic Circuit Design,” presents the research that using machine learning (ML) techniques to accurately construct mathematical models for predicting gene expressions in genetic circuit designs. Specifically, classification and regressions models were built using Random Forrest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN). The obtained accuracy of the regression model using RF and ANN yielded R2 scores of 0.97 and 0.95, respectively, compared to the best score of 0.63 obtained in Chapter 2. Furthermore, a classifier model was built using the green fluorescent protein measurements obtained from the experiments. The measured GFP values were predicted with 100% accuracy by both RF and ANN classifier models while identifying various synthetic gene circuit patterns. This work highlights importance of the relevant data preparation techniques to ensure high accuracy is obtained by the utilized ML models.

Chapter 4, “Predictable control of RNA lifetime using engineered degradation-tuning RNAs,” covers the methods to control RNA stability in various systems. Here, we report a library of RNA modules called degradation-tuning RNAs (dtRNAs) that can increase or decrease transcript stability *in vivo* and in *vitro*. The dtRNAs enable modulation of transcript stability over a 40-fold dynamic range in *Escherichia coli* while having a minimal influence on translation initiation. We harness dtRNAs in mRNAs and noncoding RNAs to tune gene circuit dynamics and enhance CRISPR interference *in vivo*. Use of stabilizing dtRNAs in cell-free transcription-translation reactions increases gene expression *in vitro*. Finally, we combine dtRNAs with toehold switch sensors to enhance the performance of paper-based norovirus diagnostics, illustrating the potential of synthetic dtRNAs for biotechnological applications.

Chapter 5, “Conclusion and Future Direction,” summaries the research specific aims and proposes directions for future research.

CHAPTER 2

DESIGN OF ADJACENT TRANSCRIPTIONAL REGIONS TO TUNE GENE EXPRESSION AND FACILITATE CIRCUIT CONSTRUCTION

2.1 Introduction

Gene circuit engineering as one of the foundation technologies has helped start the burgeoning development of bacterial synthetic biology. Based on a large collection of well-characterized biological components, including promoters, ribosome binding sites, transcriptional factors, terminators, RNA elements, and other small modules, complex gene circuits with designed functions can be wired using established biological principles. Toggle switch and repressilator are two of the earliest examples of engineered gene circuits[15,16]. Now synthetic biologists are paying increasing attention to develop innovative gene circuits for spatial pattern formation[79,80], drug development[81,82], pathogen detection[41,83], *in vivo* delivery[84], and other biotechnological applications, including nitrogen fixation[85,86] and environmental bioremediation[87].

Currently, circuit assembly has two main strategies: one is monocistronic construct, in which one promoter drives one gene expression and ensures each gene is being expressed independently; the other is polycistronic construct, in which one promoter transcribes multiple genes (operon) into a single mRNA but is translated into individual products (Figure 2.1A). Operon, a cluster of genes with functional associations under control of a single promoter, is a common type of genome organization in prokaryotic cells and is also widely found in eukaryotes and viruses[88]. This operon organization strategy, here mainly referring to the genes' order and position downstream of the promoter in an operon, ensures coordinated gene expression and regulation and enables bacteria cells to rapidly respond to environmental changes. In synthetic biology, this organization (synthetic operon) facilitates rapid construction of genetic cascades and decreases the number of biological components (such as the promoters and terminators) required for complex genetic circuits, and therefore is widely used in circuit engineering[89–96].

However, it remains unknown whether/how gene expression is affected by immediately adjacent genes in a polycistronic operon. Two previous reports have indicated that gene position and transcriptional distance can affect gene expression in a synthetic operon[97,98]. But little

research has systematically studied the effects of adjacent genes in synthetic operons on the circuit's gene expression, dynamics, and functionality. This factor is more prominent for synthetic operons containing a cluster of genes and complex multi-layered genetic circuits. Deciphering the effects of adjacent transcriptional region (ATRs) on gene expression would advance our understanding of determinants of gene expression in synthetic circuits and accelerate circuit design and assembly. Such effect has been generally neglected during engineering of synthetic gene networks, leading to unexpected circuit performance or failure[99–101]. Hence, development of a predictive method to evaluate each gene's expression level in a circuit would be of great importance to circumvent the need for trial and error in circuit design and assembly.

To quantify the effects of ATRs on gene expression, here we systematically analyzed the effect of adjacent genes and noncoding regions on GFP expression levels through construction of ~120 synthetic gene circuits (operons) in Escherichia coli. Data-driven analysis yields a new protein expression metric that strongly correlates with the features of ATRs including GC content, size, and stability of mRNA folding near ribosomal binding sites (RBS). We demonstrated this metric's utility in evaluating relative expression levels of genes by incorporating it in the design and construction of logic gates with lower basal expression and higher sensitivity and nonlinearity. Furthermore, we designed synthetic 5' ATRs to tune protein expression levels over a 300-fold range. Finally, by combining ATR regulation and mathematical modeling, we illustrated the application of synthetic ATRs in quantitatively tuning nonlinear dynamics of bistable gene networks.

2.2 Results

2.2.1 Protein expression is significantly influenced by its adjacent genes and position in the operon

To examine whether protein expression is affected by its neighbors in a polycistronic setting, we first constructed a two-gene operon (gene X and GFP), which is driven by a constitutive promoter (Figure 2.1B). Flowcytometry results showed that for different X, GFP expression varies significantly. Specifically, circuits with *AraC* and *RhlR* as X showed a

comparable level of GFP fluorescence with the control (without X gene), while the others (LuxI, *TetR*, and *dnMyD88*) showed high expression variations, ranging from 6-fold to over 120-fold decrease compared with control (Figure 2.1B). Membrane protein dnMyD88 shows the most significant influence on its neighbor GFP expression. On the other hand, RT-qPCR measurements of transcripts of GFP showed much smaller variations of mRNA concentrations between different circuits, for P1:P2 (GFP N-terminal) or P3:P4 (GFP C-terminal) primer pairs (Figures 2.1C and Figure 2.2A-C). So, the variation of mRNA concentrations for each construct is insufficient to explain the fluorescence differences, which agrees with previous studies that protein and mRNA copy numbers in E. coli cells are uncorrelated[98,102].

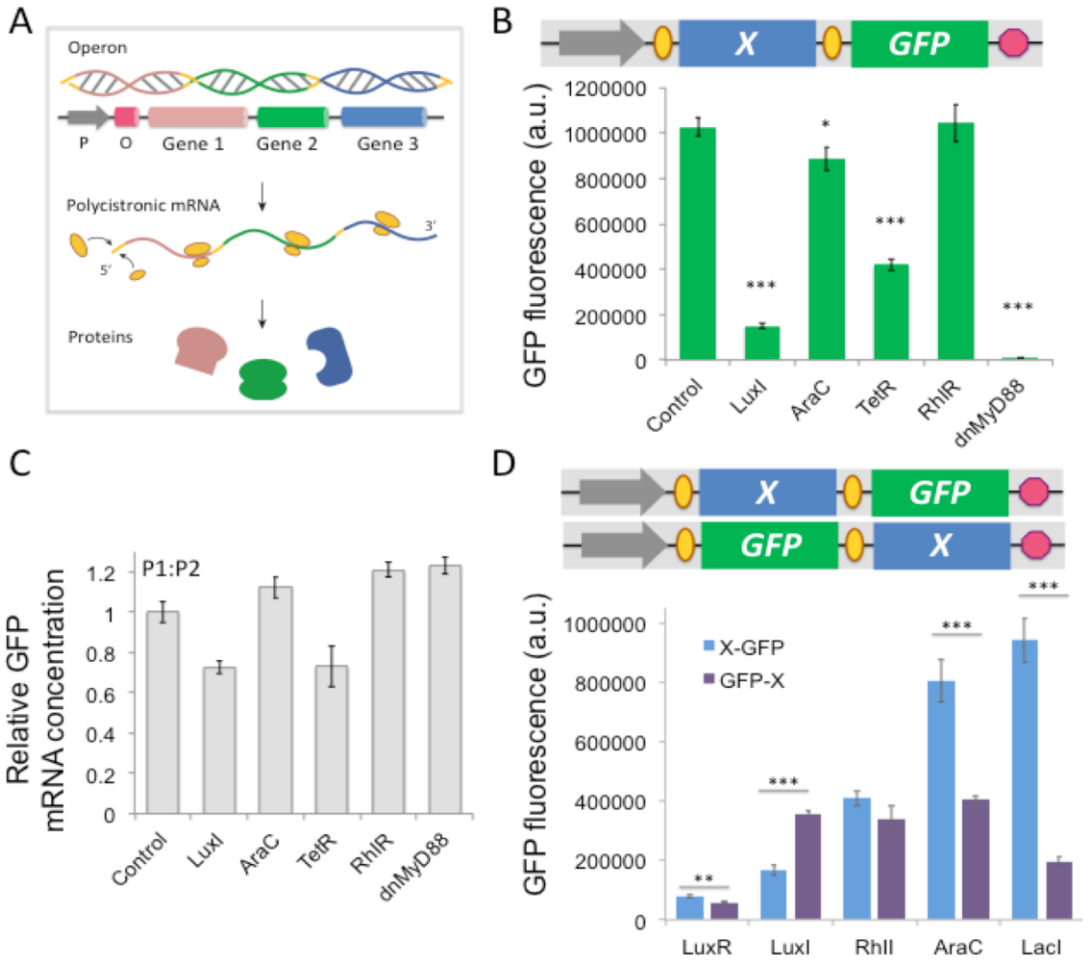


Figure 2.1 Protein Expression Is Significantly Influenced by Its Adjacent Genes and Position in Synthetic Operons (A) Illustration of the operon structure and gene expression. The three structural genes are transcribed as a polycistronic mRNA but translated into individual proteins.

P, promoter; O, operator. Yellow oval, ribosome. (B) Top: Schematic representation of synthetic bicistronic gene circuits with gene X and GFP. Gray arrow, constitutive promoter; orange oval, ribosome binding site; red hexagon, transcriptional terminator. Bottom: Flow cytometry results show GFP expression is influenced by its 5' ATRs. X represents a gene name (i.e., *LuxI*, *AraC*, *TetR*, *RhIR*, and *dnMyD88*). ''Control'' is without X gene in the circuit. Rectangles with filled colors represent different genes. Data represent the mean ± SE of eight replicates. (C) Relative GFP mRNA concentrations (normalized to 16S rRNA control) for the circuits in (B) determined by RT-qPCR. Primer pair P1:P2 was designed to amplify GFP gene from the sample cDNA. (D) Top: Schematic representation of synthetic bicistronic gene circuits with gene X and GFP, but with switched positions in the circuit. Gene position in the operon affects GFP expression. Data represent the mean ± SE of eight replicates. *$p < 0.05$, **$p < 0.001$, and ***$p < 0.0001$ by Student's t test.

Next, we further investigated the influence of a gene's position on its expression. As shown in Figure 2.1D, higher GFP expression is observed when GFP is arranged distal to the promoter for the bicistronic constructs that X gene is *RhlI*, *AraC*, or *LacI*, while there are cases showing a similar level of GFP fluorescence (*LuxR*) or higher (*LuxI*) when GFP is arranged right downstream of the promoter. Results from tricistronic constructs also indicate that GFP expression is varied for different positions in the circuit and adjacent genes (Figures 2.2D-G). Moreover, for different Xs with the same position, GFP shows substantial variations, consistent with results shown in Figure 2.1B. Altogether, these results demonstrate that a gene's sequence and position in operons influence the expression of adjacent genes.

2.2.2 Quantitative characterizations of ATR effects on synthetic operons

To quantify the impact of ATRs on protein expression, we designed and constructed ~80 circuits with different neighbor protein-coding genes and varying sizes (X and Y) to cover a wide range of GFP gene position and neighbor features (GC content, size, and mRNA secondary structure). These genes are commonly used in synthetic biology, including transcriptional factors, quorum-sensing components, and other functional genes. To ensure experimental consistency, all circuits were constructed using the same constitutive promoter, RBS, terminator, and expression vector.

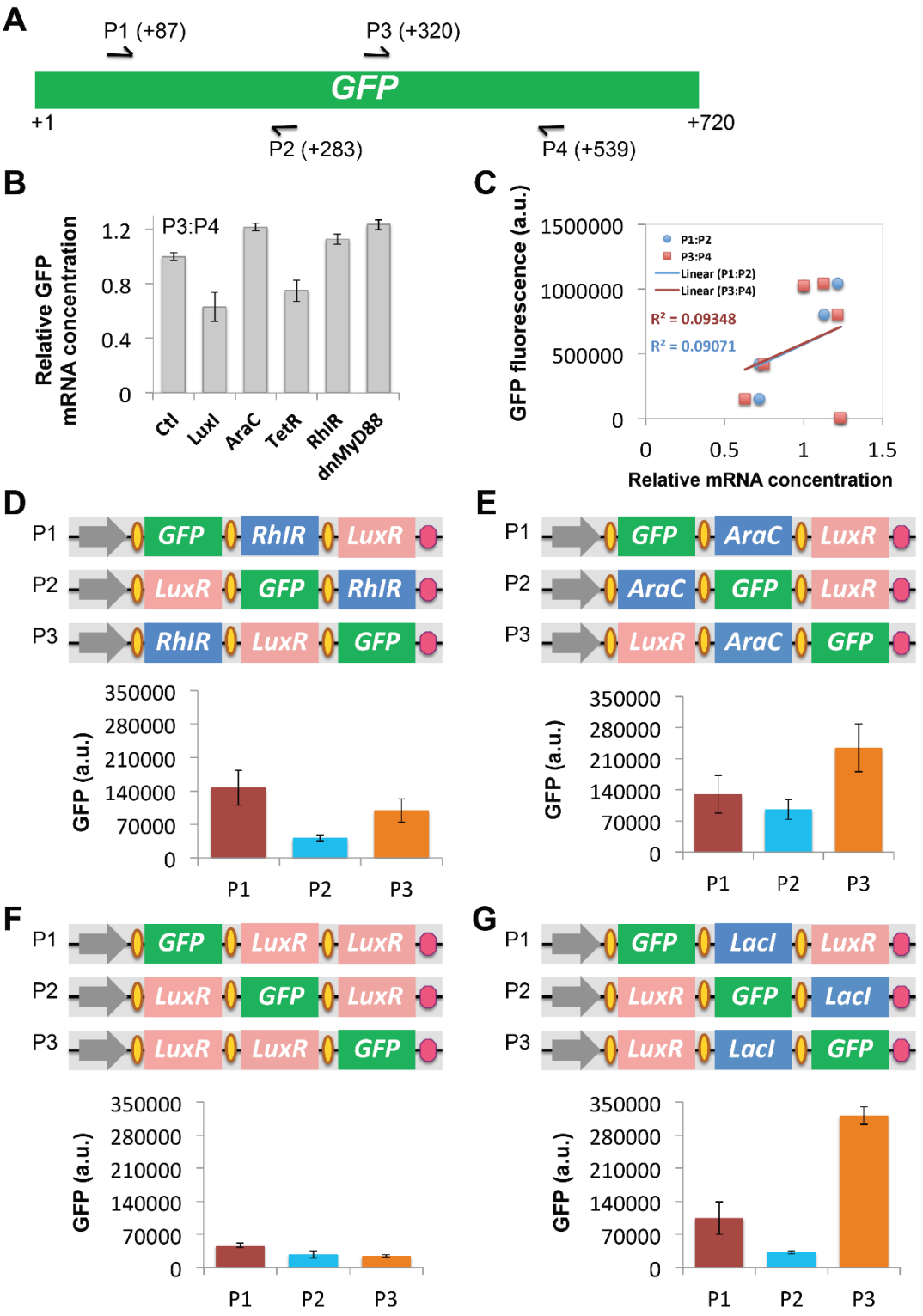


Figure 2.2 RT-qPCR Result for the Circuits in Figure 2.1B, and Gene Position in the Tricistronic Circuit Impacts GFP Expression (A) Two pairs of primers (P1:P2, and P3:P4) designed to amplify GFP gene from the sample cDNA. The binding sites of the four primers are also indicated. (B)

RT-qPCR result using primer pair P3:P4 to amplify GFP gene. The GFP mRNA concentrations were normalized to the 16S rRNA control. Error bar represents standard deviation of three biological replicates. (C) Correlation between the GFP fluorescence intensities and the relative GFP mRNA concentrations. Little correlation was found using primer pair P1:P2 or P3:P4. These results indicate there is little correlation between GFP protein fluorescence intensity and mRNA level for the circuits in Figure 2.1B. (D) GFP is arranged at proximal (P1) or middle (P2) or distal (P3) positions to the constitute promoter in the tricistronic circuit with two more genes *LuxR* and *RhlR*. Circuit with GFP at P1 position shows the highest GFP expression. (E) GFP is arranged at P1, P2, and P3 positions in the tri-cistronic circuit with genes LuxR and AraC. Circuit with GFP at P3 position shows the highest GFP expression. (F) GFP is arranged at P1, P2, and P3 positions in the tri-cistronic circuit with two copies of *LuxR* genes. Circuit with GFP at P1 position shows the highest GFP expression. (G) GFP is arranged at P1, P2, P3 positions to the constitute promoter in the tri-cistronic circuit with genes *LuxR* and *LacI*. Circuit with GFP at P3 position shows the highest GFP expression. Data represent the standard deviation of eight replicates. Gray arrow: constitutive promoter; Orange oval: ribosome binding site; Red hexagon: transcriptional terminator. Rectangles with filled colors represent different genes.

First, GFP was arranged to the distal end of synthetic bicistronic and tricistronic operons, and the DNA sequence starting from the transcription start site after the promoter to the beginning of the RBS of GFP is denoted as 5' ATRs (Figure 2.3A). Log transformation was applied to the original data because of its large variability ranging from 21,000 to 1,900,000 (GFP fluorescence, arbitrary unit) and inconstant variance. GFP expression increased with the total 5' ATRs GC content, while 5' ATR length had a negative effect on GFP expression. Sliding window analysis of 5' ATR GC content suggested that the GC content of the whole 5' ATR has the highest fitting efficiency (Figure 2.4A). We hypothesize that high GC content could increase total mRNA stability, while a long transcription process could decrease the probability of complete GFP transcription/translation and increase the probability of degradation. In addition, previous studies reported that RNA secondary structure near the RBS influences a gene's expression, so local folding energy from the -70-nt to +38-nt region around GFP's RBS (GFP's translation starting site is denoted as +1) was calculated. Consistent with previous reports[103–105], our analysis also shows that GFP expression is significantly correlated with folding energy around the RBS of GFP (Figure 2.3A).

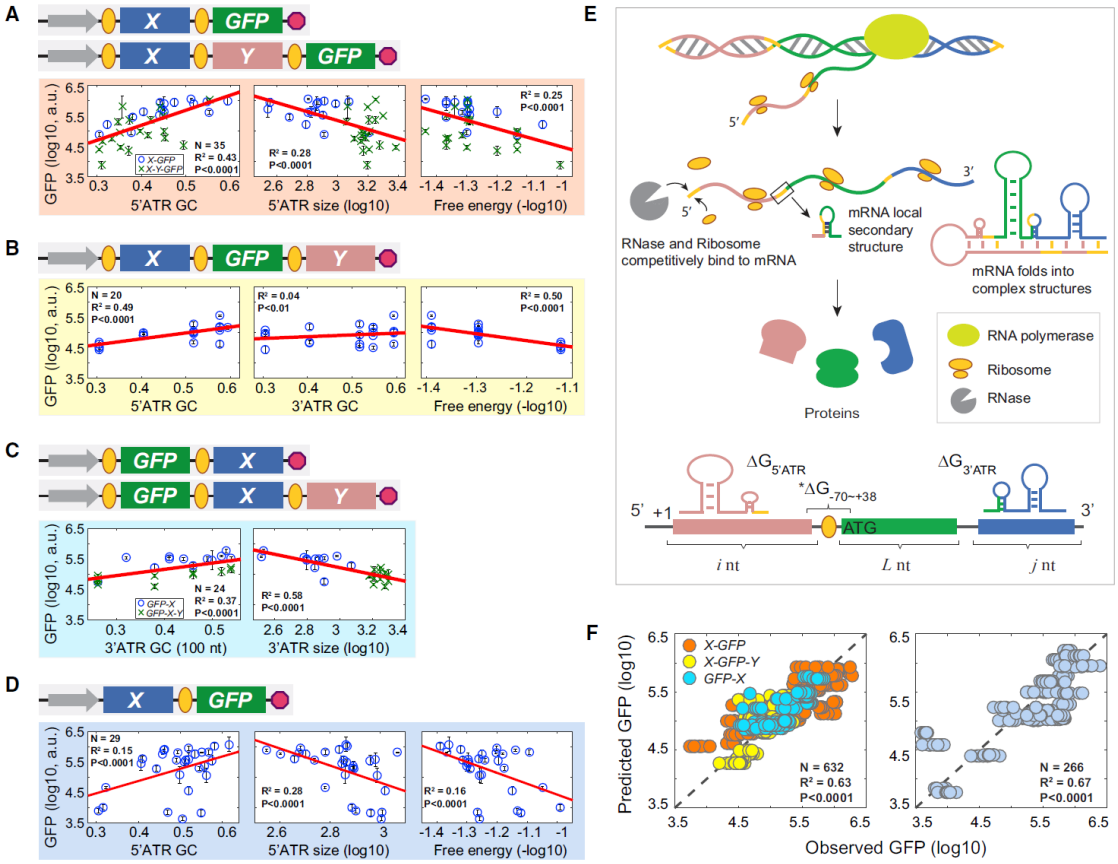


Figure 2.3 Quantitative Characterization of Adjacent Gene Regulation in Synthetic Operons (A) Scenario 1: GFP is arranged distal to the promoter. Top: Schematic representation of synthetic polycistronic gene circuits X-GFP. X and Y represent different gene names. Bottom: GFP expression is significantly affected by its 5' ATRs' GC content, size, and local folding free energy. 35 genetic circuits with one or two genes placed in front of GFP, which are labeled with different symbols in the regression results. The red lines are the linear regression results from the data. Error bars are the SD of eight measurements performed in three different days. (B) Scenario 2: GFP is placed in the middle of the three-gene operons (X-GFP-Y). GFP expression is significantly correlated with its 5' and 3'ATR GC content and local folding free energy. 20 circuits with different X and Y gene combinations were constructed. (C) Scenario 3: GFP is placed proximal to promoter (GFP-X). GFP expression is significantly affected by its 3'ATR GC content and size. 24 circuits with different 3'ATRs were constructed, and different symbols are used to indicate bi- or tricistronic constructs in the regression results. (D) Investigation of noncoding ATR regulation on GFP expression. X gene would not be expressed owing to a lack of RBS. GFP expression is significantly correlated with 5' ATR GC content, size, and local folding free energy. 29 circuits with different X genes were constructed. (E) A comprehensive model for ATR regulation on protein

expression. Top: Co-transcriptional translation and degradation. After RBS is transcribed, RNase and ribosome competitively bind to mRNA to initiate translation or degradation. Generally, gene expression is influenced by overall stability and local secondary structure. Bottom: Illustration of the five variables in the model: $\Delta G_{5'ATR}$, $\Delta G_{3'ATR\_100}$, $\Delta G_{-70\sim+38}$, and transcriptional sizes (i, j). -70 and +38 correspond to the position of the start codon (AUG) of the gene of interest. (F) Left: Experimentally observed GFP expressions are plotted against the GFP values predicted by the coding ATR model with the five statistically significant energetic terms and fitted coefficients. If the model predicted values and experimentally observed values agreed perfectly (R^2 = 100%), all the data points would fall on the dotted diagonal line of the squares. N is the total measurements for the 79 circuits. Dots with different colors indicate the data source from the three scenarios in (A–C). Right: Experimentally measured GFP fluorescence is plotted against the GFP expression predicted by the noncoding ATR model with the three statistically significant energetic terms (ΔG50ATR, i, and ΔG-70~+38).

Next, GFP was placed in the middle of the operon, and the sequence between the stop codon of GFP and the transcriptional terminator is denoted as 3' ATR. We found that 5' ATR GC content (positive impact) and local mRNA folding free energy (negative impact) have the most significant impacts on GFP expression, and 3' ATR GC content has a small contribution to GFP variations in this case (Figure 2.3B). Finally, circuits with GFP engineered proximally to the promoter were also constructed and investigated to probe the relationship between GFP expression and its 3' ATR. Similarly, results show that 3' ATR GC content and size have a positive and negative relationship with GFP fluorescence, respectively (Figure 2.3C). Sliding window analysis further revealed that the GC content of the first 100 nt of 3' ATR has the highest fitting efficiency, suggesting the rear 100 nt is important for GFP expression (Figure 2.4B).

Noncoding DNA sequences make up about 12% of the bacterial genome and play important roles in the regulation of gene expression and metabolism[106,107]. To investigate whether noncoding sequences would similarly affect adjacent gene expression in synthetic operons, we

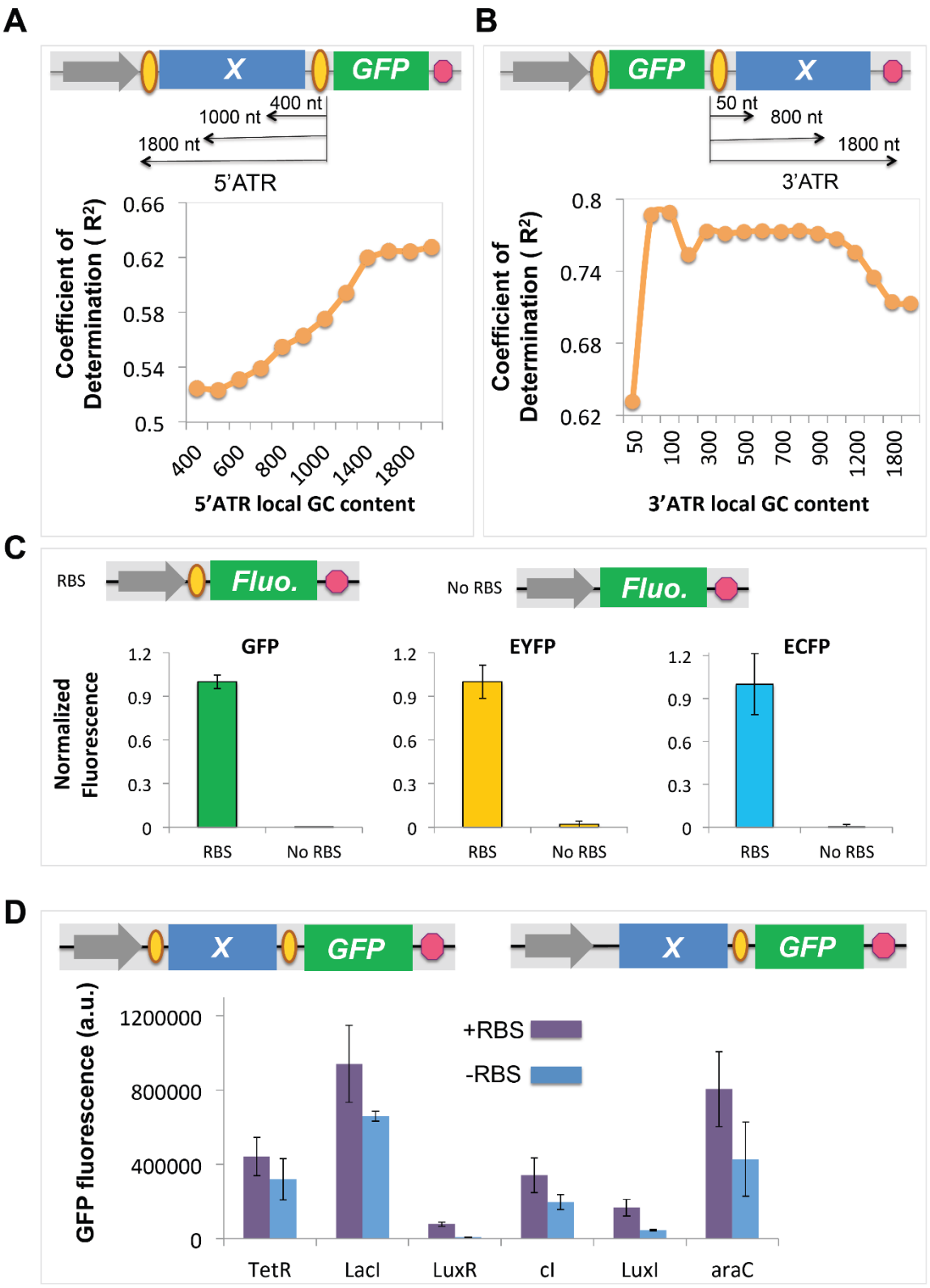


Figure 2.4 Sliding Window Analysis for Local GC Content and Model Fitting; Gene Expression Comparison for Circuits with and without RBSs (A) Top: Schematic representation of constructs

with GFP distal to the constitutive promoter. The black lines with arrows indicate 5' ATRs with different lengths. Bottom: GC content of 5' ATRs with different lengths from 400 nucleotides to the whole transcriptional region are calculated and then fitted to the model. The coefficients of determination (R^2) are compared to 5' ATRs with different lengths. Linear model results show that GC content with the whole 5' ATR has the highest fitting efficiency. X represents different genes used in the circuit. (B) Top: Schematic representation of constructs with GFP proximal to the constitutive promoter. The black lines with arrows indicate 3' ATRs with different lengths. Bottom: GC content of 3'ATRs with different lengths from 50 nucleotides to the whole transcriptional region are calculated and then fitted to the model. The coefficients of determination (R2) are compared to 3'ATRs with different lengths. Model fitting results show that GC content of the first 100 nucleotides 3' ATR has the highest fitting efficiency. (C) Top: Schematic representations of synthetic gene circuits with RBS or without RBS. Three fluorescent proteins GFP, enhanced yellow fluorescent protein (EYFP), and enhanced cyan fluorescent protein (ECFP) were chosen. Bottom: Results indicated that there are minimal fluorescence expressions for all the three circuits without RBS. Rectangle with filled color represents fluorescent genes (Fluo.). Cyan fluorescence was measured by plate reader (excitation: 405 nm; emission: 485 nm). Experimental data are replicated three times with total twelve data points, and bars represent the standard deviation of the mean. (D) Schematic representations of synthetic bi-cistronic gene circuits. GFP reporter was expressed downstream of a coding ATR (with RBS) or noncoding ATR (without RBS), respectively. Higher GFP expression was observed for circuits with the same genes (*TetR*, or *LacI*, or *LuxR*, or *cI*, or *LuxI*, or *AraC*) with RBS than those without RBS. Fluorescence was measured by flow cytometry. Error bar represents standard deviation of eight biological replicates. Gray arrow: constitutive promoter; Orange oval: ribosome binding site; Red hexagon: transcriptional terminator.

engineered 32 synthetic circuits with 32 genes, which are placed immediately downstream of the promoter without RBS to greatly limit their translation (Figure 2.4C). Our results showed a strong relationship between GFP expression and noncoding 5'ATR GC content, size, and local mRNA folding energy (Figure 2.3D) Higher GFP expression was observed for circuits with the same genes with RBS than those without RBS (Figure 2.4D), suggesting the RBS of 5' ATR may be important for mRNA stabilization and expression efficacy.

Altogether, these results offer direct evidence that adjacent coding and noncoding DNA fragments affect gene expression in synthetic operons, and ATR GC content has a positive

correlation with GFP expression while ATR size and local free energy are both negatively correlated.

2.2.3 Comprehensive model of ATR regulation

Our results revealed that gene expression in operons is affected by the sequence features of its adjacent genes and local mRNA secondary structures. The explicit mechanism of these effects remains elusive. We employed the same promoter, RBS, vector, and host cell for all the circuits to minimize the impact of transcription on protein expression variation. And there is a lack of complicated post-translational modifications in *E. coli*, so we believe that the ATR alters the secondary or tertiary structures of mRNA locally and/or globally, which perturbs the GFP mRNA translation and degradation process (Figure 2.3E). The GC content of 5' and 3' ATRs has a positive relationship with GFP expression (Figure 2.3). After the RBS is transcribed, ribosome and RNase competitively bind to mRNA[108,109]. So we infer that a GC-rich 5' and 3' ATR, which is likely to have a more stable secondary structure[110,111], could help stabilize the GFP transcript and decrease the risk of degradation by RNase, and thus result in higher GFP expression. On the other hand, the 5' and 3' ATR sizes are negatively correlated with GFP expression (Figure 2.3). Longer ATR may lead to lower mRNA stability due to the increased probability of elongation pausing and degradation of RNase. Moreover, the local mRNA folding energy near GFP's RBS (-70 to +38 nt) is believed to have an impact on the translation initiation of GFP. Overall, our statistics analysis revealed that 5' ATR GC content is the most important variable in the regression models for the X-GFP circuit (Figure 2.3A, partial R^2 =0.44) and *X-GFP-Y* (Figure 2.3B, partial R^2 = 0.51), whereas 3' ATR size has a bigger role in the model of *GFP-X* (Figure 2.3C, partial R^2 = 0.58). This result suggests that gene expression may be more easily modulated by the GC content of its 5' ATR and the size of 3' ATR.

To explore the possible mechanistic basis of ATR regulation and make quantitative predictions, we developed a comprehensive linear model integrating the three scenarios in Figures 2.3A-C. The biophysical model was based on previous pioneer work characterizing the relationship between free energy changes and protein translation initiation[112–115]. We next

developed a comprehensive model to explore the possible mechanistic basis of ATR regulation. The model builds on measurements of sequence-dependent energetic changes during polycistronic mRNA folding and translation. The energetic changes correspond to the translation efficiency and protein abundance (c, equation 1).

$$c \propto \exp(-\Sigma \beta_x \Delta G_{x,}), x = 1,2,3,\ldots \quad (1)$$

where ΔG is the energy term and b is the scaling coefficient[112]. For a given gene in an operon, the size of 5' and 3' ATRs is denoted as i nt and j nt, respectively (Figure 2.3E). The minimumfree energy of the local GFP mRNA secondary structure around the RBS is $\Delta G_{-70\sim+38}$. The entire folding energy for 5' ATR is $\Delta G_{5'ATR}$. The GC content of the first 100-nt 3' ATR has the highest fitting efficacy for GFP expression (Figures 2.3C and 2.4B), and it is known that GC content is correlated with the thermodynamic parameter ΔG[116,117], so we only calculated the free energy of the first 100 nt of 3' ATR ($\Delta G_{3'ATR\_100}$). Thus, the sum of the energy changes can be quantified to assess the abundance of a given gene expression (equation 2):

$$-\Sigma \beta_x \Delta G_x = \beta_0 + \beta_1 \cdot \Delta G_{5'ATR} + \beta_2 \cdot \Delta G_{3'ATR\_100} + \beta_3 \cdot i \cdot Gm + \beta_4 \cdot j \cdot Gm + \beta_5 \cdot \Delta G_{-70\sim+38} \quad (2)$$

The folding energy of $\Delta G_{5'ATR}$, $\Delta G_{3'ATR\_100}$, and $\Delta G_{-70\sim+38}$ is totally sequence dependent, and Gm is an average energy cost for synthesizing a nucleotide, which here for simplicity we assume is a constant. Although all the five variables are contained in the model, some variables may be unnecessary for a specific gene organization in a circuit. For example, in the noncoding ATR cases with X-GFP organization (Figure 2.3D), the j and $\Delta G_{3'ATR\_100}$ terms are constant values, owing to a lack of 3'ATRs.

The comprehensive model combined the three different scenarios with GFP placed at different positions in a polycistronic gene circuit (Figures 2.3A-C). To verify whether the five variables are necessary for the best prediction of the model, we performed stepwise regression to test the significance of each variable through adding or removing one of the variables step by step (the significance level for variable entry or stay is 0.05). From the sequence of generated models, the selected model is chosen based on the lowest Akaike information criterion. Our results indicated that all five variables are necessary for the coding ATR model integrating the three scenarios in Figures 2.3A-C, and the comprehensive model explains 63% of GFP variations

(Figure 2.3F, left). The noncoding ATR model with the three statistically significant variables $\Delta G_{5'ATR}$, i, and $\Delta G_{-70\sim+38}$ explains 67% of GFP variations (Figure 2.3F, right). With the comprehensive model, we can evaluate the influence of the adjacent transcriptional sequences on the expression of a certain gene in the operon, which provides a guide for circuit design and optimization during circuit engineering.

2.2.4 Protein expression metric guided logic circuit design

To illustrate how the metric could be used to guide circuit design, synthetic AND logic gate was designed and tested. The gate is composed of a hybrid promoter *pLux/tet*, which has one LuxR-AHL and one TetR binding site. GFP is the output. Maximized GFP expression is achieved in presence of two inputs AHL and aTc (Figure 2.5A), where AHL binds with LuxR protein to activate *pLux/tet* transcription and aTc can block TetR repression to *pLux/tet*. LuxR and TetR are constitutively expressed from the same promoter.

There are two possible ways to assemble this circuit, one is LuxR-TetR (LT) combination, and the other is TetR-LuxR (TL). The GC content of *LuxR* (30.3% GC, 781 bp) is lower than *TetR* (40.4% GC, 685 bp). So in AND-gate LT, TetR expression is lowered by its 5'-low-GC-content neighbor while the impact of LuxR to TetR expression in logic TL is minor because the size of 3' ATR is a more significant factor than GC content. We then calculated the equation for each circuit design and fed it into our model; the results indicate that LuxR expression in TL decreases by 4.4% compared with gate LT, however, TetR expression increases by 93.6% in circuit TL (Table 2.1). Therefore, we infer that the basal GFP expression in circuit LT would be greater than in TL, whereas TL would harbor more dynamic responses with induction of aTc because of higher TetR expression. An ordinary differential equation (ODE) model was then developed to simulate GFP expression based on the normalized LuxR and TetR production rate changes in the LT and TL gates. By tuning the relative production rates of LuxR and TetR according to the comprehensive regression model, we can predict GFP dynamics under induction of AHL and aTc (Figures 2.5B and 2.6A, solid lines). It can be seen that, after normalization, experimental dose-response results, shown as colored circles, are consistent with ODE model predictions for all aTc

concentrations. Basal expression of *pLux/tet* in circuit LT is significantly higher than in circuit TL (Figures 2.5B and 2.6A, data points with error bar). Moreover, the maximum GFP fluorescence is also higher in circuit LT, owing to decreased LuxR expression in gate TL. In addition, the sensitivity to AHL (concentration for half-maximal activation of GFP, $K_{0.5}$) is improved 2.4- to 64.5-fold in circuit TL compared with LT for different concentrations of aTc. And the nonlinearity (Hill coefficient) is generally increased 2- to 5-fold with high concentrations of aTc induction. These data are in accordance with the model calculations that TetR expression is relatively increased in circuit TL than in LT, which suppresses the basal expression of *pLux/tet* and improves the sensitivity and nonlinearity of the promoter to AHL and aTc.

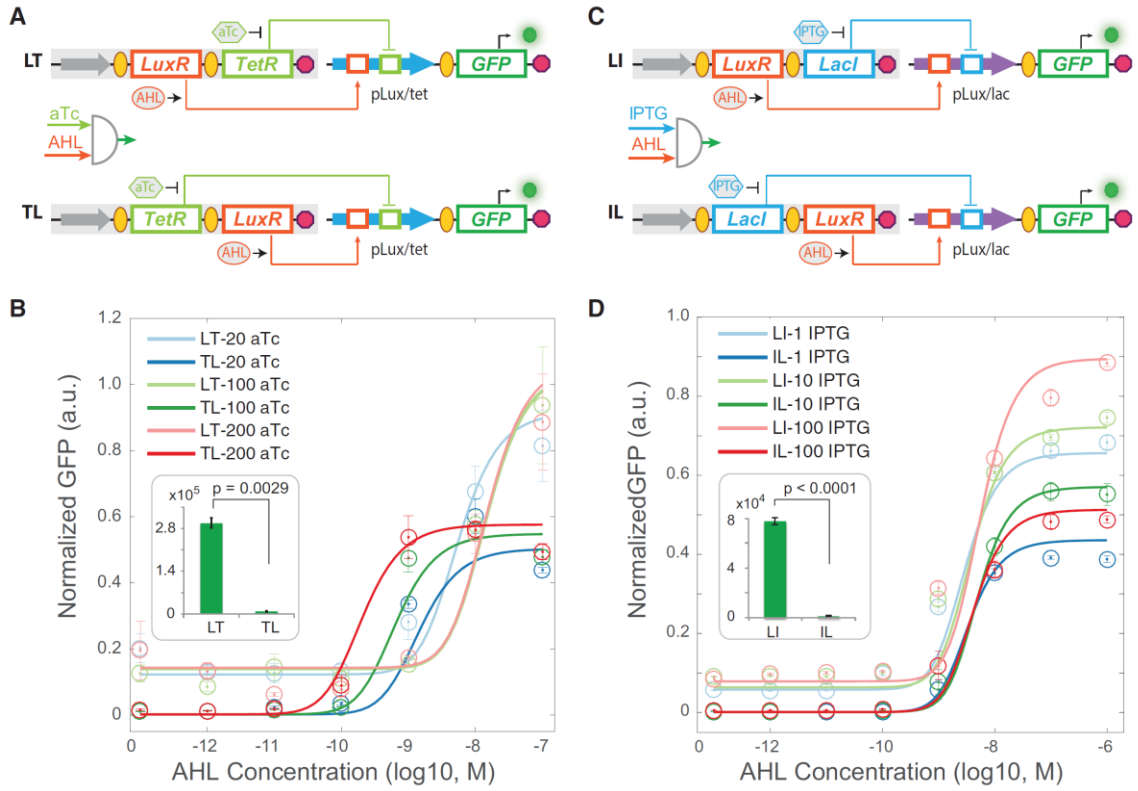


Figure 2.5 Model-Guided Circuit Design for Synthetic Logic Gates (A) Two designs for *pLux/tet*–AND logic gate. A constitutive promoter (gray arrow) drives LuxR (orange rectangle) and TetR (green rectangle) expression. *pLux/tet* is highly activated in the presence of both AHL and aTc. LT and TL represent the order of LuxR and TetR positions in the operon. LuxR can bind with AHL (gray oval) to activate *pLux/tet* promoter (blue arrow), while aTc (green hexagon) can block TetR inhibition to *pLux/tet* promoter. Lines with arrowheads indicate activation, and lines with T bars indicate inhibition. (B) Dose-response curves for different concentrations of AHL and aTc. The

solid lines are from ODE model simulations based on the calculated relative changes of LuxR and TetR concentrations in LT and TL from our linear comprehensive model. Data points with error bars are experimental results, showing good match with model predictions. The inset diagram is the basal expression of GFP for design of LT and TL. Color curves are inductions with different aTc concentrations (20 ng/mL, 100 ng/mL, and 200 ng/mL). (C) Two designs for *pLux/lac*–AND logic gate. A constitutive promoter drives LuxR and LacI expression. *pLux/lac* (purple arrow) is highly activated in the presence of both AHL (N-(b-ketocaproyl)-L-homoserine lactone) and IPTG (isopropyl b-D-1-thiogalactopyranoside, blue hexagon). LuxR can bind with AHL to activate *pLux/lac* promoter, while IPTG can block LacI inhibition to *pLux/lac* promoter. LI and IL represent the order of LuxR and LacI positions in the operon. (D) Dose-response curves for different concentrations of AHL and IPTG. The solid lines are model simulations based on the calculated relative changes of LuxR and LacI concentrations in LI and IL from our linear comprehensive model. Experimental results (data points with error bar) show good match with model predictions. Color curves are inductions with different IPTG concentrations (1 mM, 10 mM, and 100 mM). Inset diagram is the basal expression of GFP for design of LI and IL. Data represent the mean ± SE of three replicates. p values were calculated using Student's t test.

To further validate the metric's utility, another two AND-gate gene circuits (LI and IL) with the position of the genes switched (LuxR and LacI) were designed (Figure 2.5C). Hybrid promoter *pLux/lac* was used to indicate the relative concentrations of LuxR and LacI produced from the operon. LacI (53.3%, 1,153 bp) has a high GC content, which may increase LuxR expression. Our model calculations showed that LuxR expression increases by 74.3% and LacI increases by 38.1% in circuit IL compared with LI. Since promoter *pLux/lac* has two LacI-binding sites (one is in the region between -35 and -10, and the other is downstream of -10 element), so the overall LacI inhibition efficiency is increased ~76.2% considering the importance of spacing between the -35 and -10 elements to RNA polymerase binding (Table 2.1). Therefore, the basal GFP expression of logic IL would be lowered compared with LI. The ODE model also indicates higher GFP expression in gate LI (Figures 2.5D and 2.6B, solid lines). Experimental results confirmed that the basal expression for circuit LI is ~54-fold higher than IL, and GFP expression under each induction is higher in gate LI, which is consistent with the ODE model results (Figures 2.5D and 2.6B).

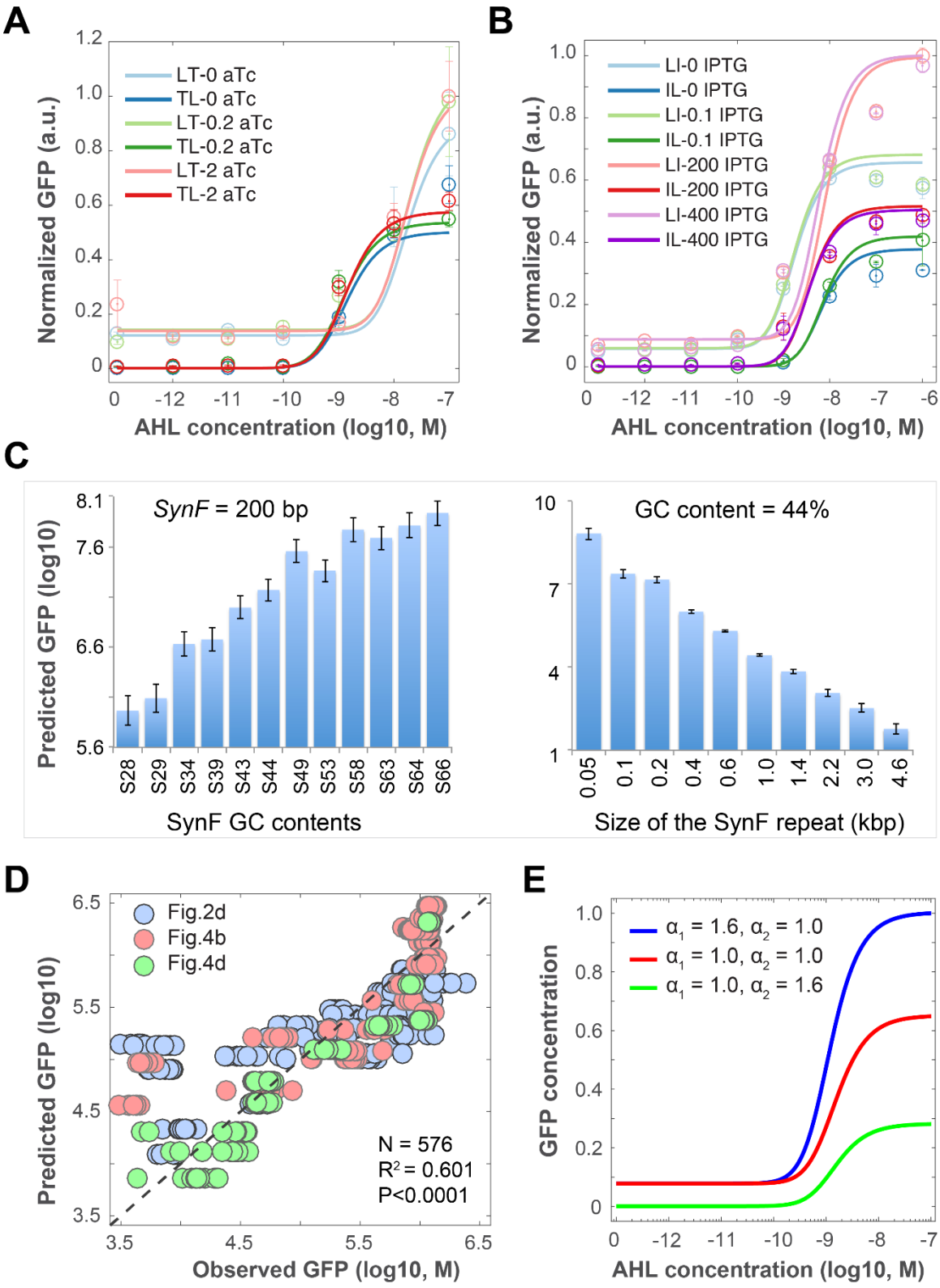


*Figure 2.*6 Model Simulation and Experimental Validation of GFP Dynamics for Synthetic Logic gates; GFP Expression Prediction Using Synthetic Fragments and Model Simulation for Different

Production Rates of LuxR and TetR in Circuit LT and TL (A) Dose responsive curves for different concentrations of AHL and aTc. The solid lines are from ODE model simulations based on the calculated relative changes of LuxR and TetR concentrations in LT and TL from our linear comprehensive model. Data points with error bar are experimental results, showing good match with model predictions. Color curves are inductions with different aTc concentrations (0 ng/ml, 0.2 ng/ml, and 2 ng/ml). Data represent the mean ± s.d. of three replicates. (B) Dose responsive curves for different concentrations of AHL and IPTG. The solid lines are model simulations based on the calculated relative changes of LuxR and LacI concentrations in LI and IL from our linear comprehensive model. Experimental results (data point with error bar) show good match with model predictions. Color curves are inductions with different IPTG concentrations (0, 0.1 μM, 200 μM and 400 μM). Data represent the mean ± s.d. of three replicates. (C) Left: Model predicted GFP expression under regulation of synthetic fragments with constant size but varying GC content from 28% to 66%. GFP predictions (mean and standard deviation) were calculated by XLSTAT based on the comprehensive noncoding ATR model (Figure 2.3F). Right: Model predicted GFP expression under regulation of synthetic fragments with constant GC content but varying sizes from 50 bp to 4600 bp. (D) Refined non-coding ATR model with the data in Figure 2.7B and 2.7D. The total data points (N) is 576 (266 from Figure 2B; 190 from Figure 2.7B; and 120 from Figure 2.7D). (E) α1 and α2 are the production rates of LuxR and TetR protein, respectively. Three scenarios with different values of α1 and α2 are plotted. When α1 > α2, more LuxR protein is produced, resulting in higher GFP expression. When α1 < α2, more TetR protein is produced, leading to lower GFP expression. Through tuning the relative size of α1 and α2, we can predict GFP dynamics under induction of AHL and aTc.

Taken together, the two sets of AND logic gates are an example of applying our comprehensive model-based tool to evaluate each gene's relative expression level in synthetic AND-gate gene circuits, and verify that ATRs' features and local mRNA stability changes in an operon-based gene network affect protein expression and circuit performance, including basal level, sensitivity, and nonlinearity. Furthermore, the tool could serve as a much-needed quantitative guide for rational design and optimization of gene expression for large genetic circuits.

2.2.5 Tuning gene expression with synthetic 5' ATRs

In general, the minimum free energy of RNA folding has a negative correlation with GC

Table 2.1 Model Evaluation for Each Gene's Expression in the AND Logic Gate

Coefficients of the comprehensive model								
	5'ATR Size	**3'ATR Size**	**Free energy ($\Delta G_{-70\to+38}$)**	**$\Delta G_{5'ATR}$**	**$\Delta G_{3'ATR}$ (100nt)**	**Intercept**		
	-3.10443	-1.15864	-1.1206	-1.80151	-0.28049	11.4635		
AND-Gates								
CP-LuxR-LacI(LacI)	815	8	-16.9	-121.6	-0.00001			
CP-LuxR-LacI(LuxR)	8	1187	-70	-0.05	-16.6			
						Predicted Expression	**Relative expression (10^(IL-LI))**	**Overall efficiency**
CP-LuxR-LacI(LacI)	2.911157609	0.903089987	-1.227886705	-2.084933575	5	5.109207328		
CP-LuxR-LacI(LuxR)	0.903089987	3.074450719	-1.84509804	1.301029996	-1.220108088	5.163765204		
CP-LacI-LuxR(LacI)	8	815	-70	-0.05	-7.1		1.38132904	2.762658081
CP-LacI-LuxR(LuxR)	1187	8	-15	-365.5	-0.00001		1.743840175	1.743840175
CP-LacI-LuxR(LacI)	0.903089987	2.911157609	-1.84509804	1.301029996	-0.851258349	5.24950447		
CP-LacI-LuxR(LuxR)	3.074450719	0.903089987	-1.176091259	-2.562887381	5	5.405271883		
CP-LuxR-TetR(TetR)	815	8	-12.5	-121.6	-0.00001			
CP-LuxR-TetR(LuxR)	8	719	-70	-0.05	-11.6			
							(10^(TL-LT))	
CP-LuxR-TetR(TetR)	2.911157609	0.903089987	-1.096910013	-2.084933575	5	4.962434847		
CP-LuxR-TetR(LuxR)	0.903089987	2.85672889	-1.84509804	1.301029996	-1.064457989	5.372368128		
CP-TetR-LuxR(TetR)	8	815	-70	-0.05	-7.1		1.936732422	1.936732422
CP-TetR-LuxR(LuxR)	719	8	-13.7	-152.5	-0.00001		0.956758533	0.956758533
CP-TetR-LuxR(TetR)	0.903089987	2.911157609	-1.84509804	1.301029996	-0.851258349	5.24950447		
CP-TetR-LuxR(LuxR)	2.85672889	0.903089987	-1.136720567	-2.183269844	5	5.353170472		

content[116,117]. Next, we sought to use synthetic noncoding DNA fragments, with the same size but varying GC content or the same GC content but varying sizes, to fine-tune gene expression in synthetic circuits. We first synthesized six short DNA fragments (with a constant size of 200 bp) with varying GC content from 28% to 53%, which were inserted downstream of the LuxR gene but upstream of GFP in the two-gene operon (*Promoter-LuxR-Synthetic fragment-GFP*). According to our model, synthetic fragments with varying GC content could tune GFP expression (Figure 2.6C).

Experimental results show that GFP expression is continuously increased for synthetic fragments with increasing GC content from 28% to 53% (Figure 2.7A). Low-GC-content fragments downregulated GFP expression about 25-fold. Microscopy results further confirmed flow cytometry data and visualized a gradual increase of fluorescence intensity with increasing GC content ATRs (Figure 2.7B). Using this strategy, we further synthesized 13 DNA fragments as 5' ATRs with varying GC content but with a constant size (200 bp) and placed downstream of the promoter (Figure 2.7C). Results indicate that synthetic short DNA sequences have a substantial impact on GFP expression: low- GC-content ATRs largely decrease expression of neighbors (up to 366-fold) and exhibit a gradually increasing pattern from 28% to 48%, while high GC content

(48%–67%) ATRs drive GFP expression to a level comparable with the control (without synthetic fragments). It is possible that GFP achieves its maximum expression when the upstream ATR mRNA piece has a relatively stable structure.

To further verify the role of ATR regulation, we varied the size of 5' ATR through shortening and adding a common sequence[118]. Using S44 (GC, 44%; size, 200 bp) in Figure 2.7C as the seed sequence, we shortened it to 100 bp and 50 bp, and lengthened it from 400 bp (combined with two pieces of S44) to 4,600 bp (combined with 23 pieces of S44), and all ten fragments have the same GC content (44%, Figure 2.7D). Model analysis and flow cytometry results show that GFP fluorescence intensity gradually decreases with increasing 5' ATR sizes (Figures 2.6C and 2.7D). We also used the data to further refine our comprehensive noncoding model and found three variables $\Delta G_{5'ATR}$, i, and $\Delta G_{-70\sim+38}$ are still required for the best fitting efficacy and explains 60.1% of GFP variations (Figure 2.6D). The refined model further expands the variables' range (GC, 28% to 67%; size, 50 to 4,600 bp) and could provide more accurate predictions. Taken together, we demonstrate that synthetic noncoding 5' ATRs with designed GC content and sizes can be used to accurately tune gene expression and achieve expression levels spanning more than 300-fold.

2.2.5 Tuning gene expression with synthetic 5' ATRs

Finally, we illustrated the application of synthetic ATRs to modulate the nonlinear bistable potential of synthetic toggle switches. As illustrated in Figure 2.8A, LacI protein could inhibit TetR by binding the *pLac* promoter, while TetR could bind *pTet* to block LacI expression, forming a mutually inhibitory network. Here, we positioned 200 bp synthetic ATRs with 28% and 67% GC content upstream of RBS-TetR module to tune TetR production (T_S28 and T_S67). According to our analysis above, low-GC content 5' ATR can downregulate TetR expression, while high GC content can keep TetR at a high level.

Flow cytometry was employed to analyze the initial states of the toggle switches with ATR insertions. As shown in Figure 2.8B, T_WT initially shows bimodal distribution, GFP-ON and GFP-OFF populations, resulting from gene expression noise in a relatively balanced system. In

contrast, both T_S28 and T_S67 exhibited unimodal distributions. Synthetic ATR S28 decreased TetR expression leading to higher LacI and GFP expression, whereas the fragment with 67% GC content showed a lower GFP expression than T_S28 and slightly lower than the high GFP population cells in T_WT (Figure 2.8B). The results indicate that the synthetic ATRs can tune the initial steady states of toggle switches and modulate the population from bimodal to unimodal distributions.

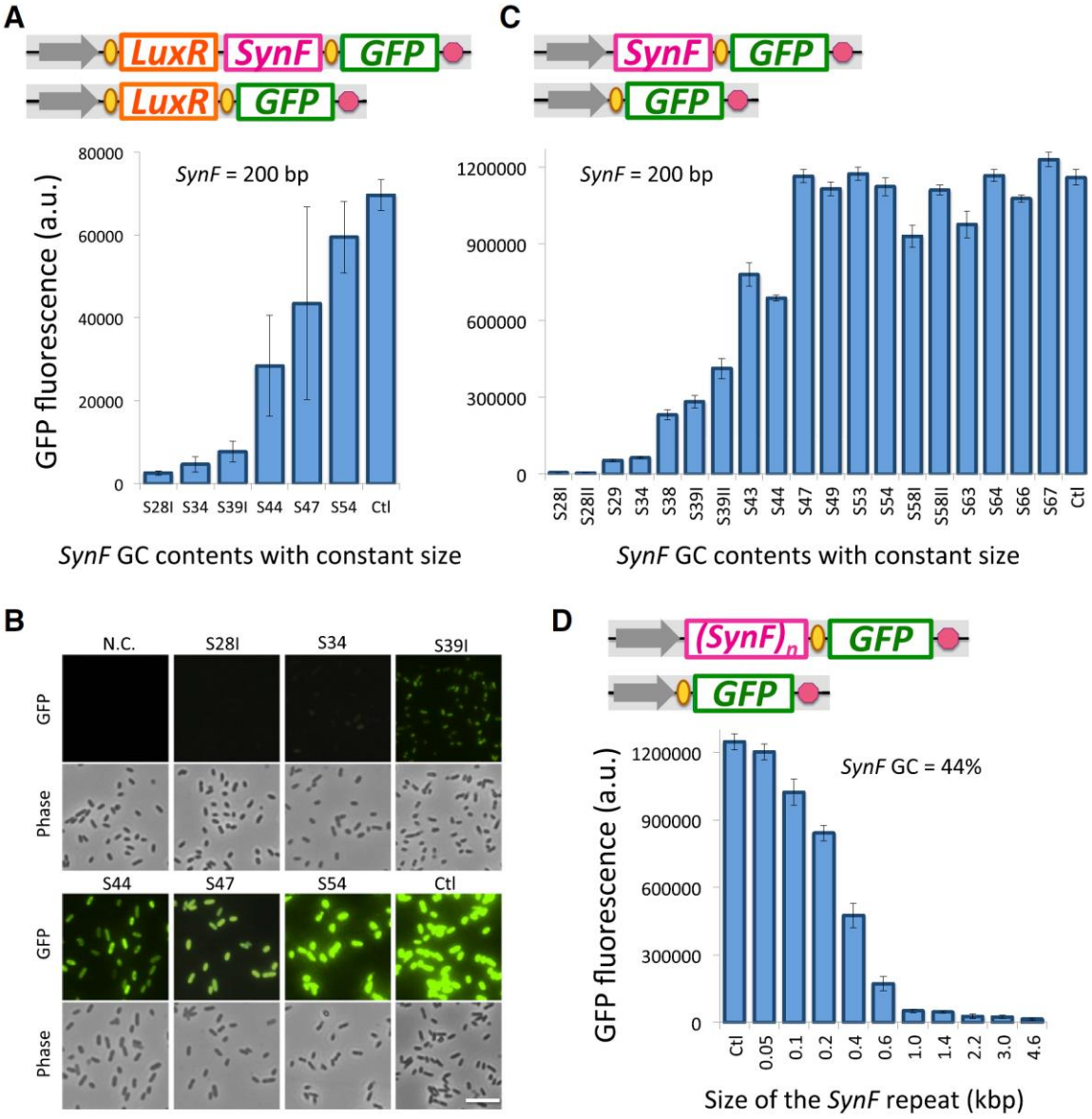


Figure 2.7 Tuning Gene Expression with Synthetic 5' ATRs (A) Synthetic 5' ATRs (SynF) to tune GFP expression for circuit *CP-LuxR-GFP*. 200 bp ATRs were inserted between LuxR and GFP

genes to tune GFP expression, and the control (Ctl) was constructed without an ATR insert. Flow cytometry results indicate that GFP fluorescence increases with gradually increasing 5' ATR GC content from 28% to 54%. (B) Microscopy results for GFP fluorescence for the constructs in (A). Scale bar, 5 mm. Magnification, 40x. (C) Synthetic 5' ATRs (SynF) with different GC content to tune GFP expression for circuit CP-GFP. All the SynF are the same size (200 bp) and are inserted upstream of GFP gene (top). Flow cytometry results of GFP fluorescence for 5' ATRs with GC content from 28% to 67% (bottom). (D) Circuits with different sizes of 5' ATR (through shortening and adding a common sequence S44; GC, 44%; size, 200 bp) were constructed to tune GFP expression. Flow cytometry results show that GFP fluorescence intensity gradually decreases with increasing size of 5'ATR. Error bars are mean ± SD of at least ten measurements performed on three different days.

To achieve a quantitative understanding of the ATR's regulation on bistability, we performed bifurcation analysis from the same mathematical model as the classical toggle switch[15]. We found that the production rate of TetR has a considerable effect on bistability and the bistable region. A small production rate, corresponding to low-GC ATR, has a small bistable region, whereas an increase in the production rate leads to a larger bistable region (Figure 2.8C). Experimentally, hysteresis of the three toggles was tested to verify the model analysis. The results indicate that all three toggles exhibited hysteresis, and T_WT harbors the broadest bistable region (Figures 2.8D-F). Moreover, consistent with model analysis, the bistable regions gradually decreased from T_WT to T_S67 to T_S28. Collectively, these results validate a novel strategy of using synthetic ATRs to tune the initial steady states and bistability of gene networks. Furthermore, this example demonstrates the feasibility of bridging ATR regulation with mathematical modeling to quantitatively understand and tune gene network dynamics.

2.3 Discussion

Circuit engineering is the first step for synthetic biologists to achieve designed functionalities with synthetic gene circuits. A successful synthetic gene circuit depends on full characterization of the biological components and the interactions that emerge between modules when assembled into a complete gene network[96,119–121]. Development of a reliable tool to predict

protein expression in the circuit has wide applications in biotechnology. For example, RBS Calculator is a well-developed design tool to predict and control translation initiation and protein expression in bacteria[91,112].

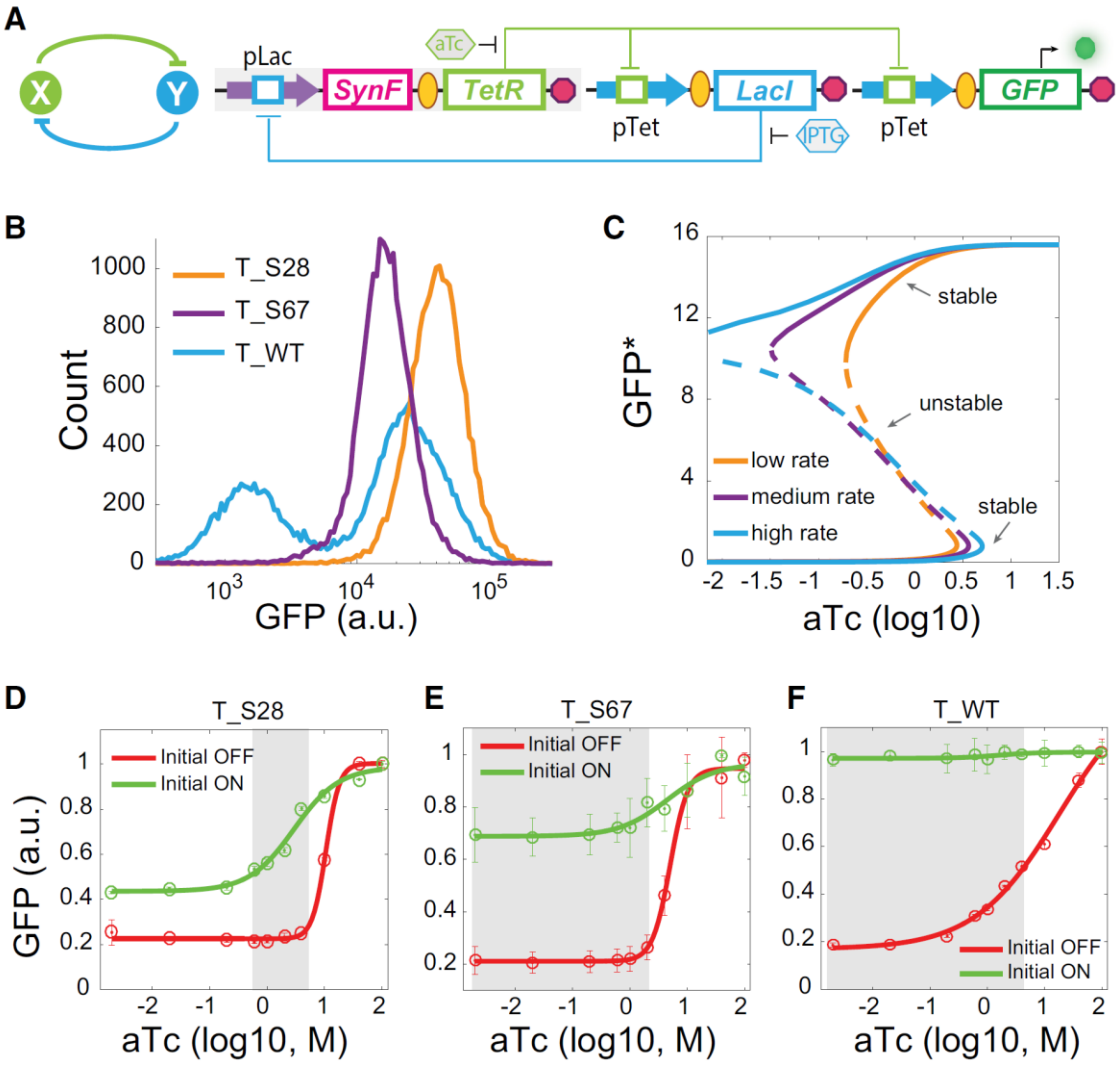


Figure 2.8 Using Synthetic ATRs to Modulate Bistability of Toggle Switches (A) Left: Abstract diagramof toggle switch topology, where X and Y mutually inhibit each other. Right: Molecular implementation of the toggle switch. LacI inhibits TetR by binding the pLac promoter, while TetR binds pTet to block LacI expression, forming a mutually inhibitory network. Inducers IPTG and aTc (hexagon) can relieve LacI and TetR inhibition, respectively. GFP serves as the readout of the pTet promoter. Synthetic ATRs (SynF) were arranged right upstream of the TetR gene. (B) Initial steady states for the three toggles. Toggle without ATR insertion (T_WT) shows bimodal distribution (GFP-OFF and GFP-ON), while T_S28 (ATR with 28% GC content, 200 bp) shows higher GFP expression and T_S67 (ATR with 67% GC content, 200 bp) shows lower GFP

expression than the GFP-ON population of T_WT. (C) Bifurcation analysis for GFP (LacI) expression with different TetR production rates under induction of varying concentrations of aTc. A low production rate for TetR, corresponding to T_S28, has the smallest bistable region, while a high rate (corresponding to T_WT) has the broadest bistable region. Solid lines represent stable steady-state solutions and dotted lines are unstable steady state solutions. GFP* is the computed GFP abundance from the model. (D–F) Hysteresis results for toggles (D) T_S28, (E) T_S67, and (F) T_WT under induction of varying concentrations of aTc. Red lines indicate the initial OFF cells with basal GFP expression, while green lines indicate the initial ON cells with high GFP expression. Data represent the mean ± SD of three replicates. The gray area is the presumed bistable region for each circuit.

Here, we systematically investigated how adjacent transcriptional regions affect protein expression in synthetic operon-based gene circuits. Through placing the GFP at different positions (proximal, middle, and distal) to the promoter, we developed a new protein expression metric that takes into account the features of adjacent transcriptional regions, including GC content, size, and stability of mRNA folding near RBS (Figure 2.3). The metric was established from about 120 gene circuits, which to our knowledge represents one of the largest databases of operon-based synthetic gene circuits in one study so far. This metric explains 63% and 67% of GFP variations in the coding ATR and noncoding ATR polycistronic gene circuits, respectively. Moreover, our experimental results also demonstrated the metric's predictions of gene expression changes and induced nonlinear dynamic responses in different genetic contexts (Figures 2.5, 2.7, and 2.8), suggesting the model's utility in guiding circuit design. Most ATRs in the circuits were 500–2,000 bp, and the maximum is 2,422 bp, which may undermine the contribution of ATR size to GFP variation. Moreover, because of the limitation in sample size and available gene resources, the collected data are not perfectly normally distributed, especially for circuits with GFP in the middle (*X-GFP-Y*), which may compromise the robustness of the model.

Consistent with previous results that gene position in operons can affect gene expression[97,98], our results further demonstrate that gene position (corresponding to change of ATR) significantly altered gene network dynamics, including basal expression, system sensitivity, and nonlinearity, which has profound impacts for nonlinear dynamic systems. Such an adjacent

gene regulation effect has been generally neglected during construction of synthetic gene networks.

Although it is relatively well established that gene expression is influenced by the local context, holistic understanding of architectural rules governing polycistronic gene circuits remains largely unexplored. Compared with previous gene expression tuning strategies or insulation strategies, such as RBS Calculator, bicistronic design with translation of a short leader peptide, or a designed DNA sequence surrounding the start codon (mostly less than 100 bp)[91,112,122–124], our work places more emphasis on whether and how polycistronic operon organization (X-GFP, X-GFP-Y, and GFP-X) and different adjacent genes (size ranging from 313 to 2,362 bp, and GC content ranging from 30.3% to 60.4%) affect protein expression in operon-based gene circuits. Furthermore, we validated that the usage of designed synthetic DNA fragments with either different GC content (28%–67%) or size (50–4,600 bp) as 5' ATRs tuned gene expression and modulated bistable regions of genetic toggle switches. The synthetic ATRs have a wide variable interval, therefore making them potentially applicable to a broad range of scientific and engineering tasks. Such a gene expression tuning strategy also avoids the production of unwanted peptides and hence reduces potential metabolic burden. We also observed that circuits having different ATRs have an impact on the time that cells reach stationary phase with similar optical density (Figure 2.9), suggesting that ATRs could be used as a means to ''program'' the metabolic load and fitness of a cell simultaneously.

Our results show that the context dependency of gene expression is not just limited to the RBS region but also includes characteristics of the whole operon. This ''global'' effect in polycistronic operons could be quantified by a biophysical model, which explains nearly two-thirds of protein expression variations across all the circuits with different configurations. The quantitative relationship between adjacent transcriptional regions and gene expression regulation in polycistronic circuits helps to evaluate each gene's relative expression levels in a circuit and predict circuit outputs, which would save experimentalist's time and resources to screen and test combinations of modules, and thus should greatly facilitate optimization of circuit design and accelerate the engineering of complex gene networks.

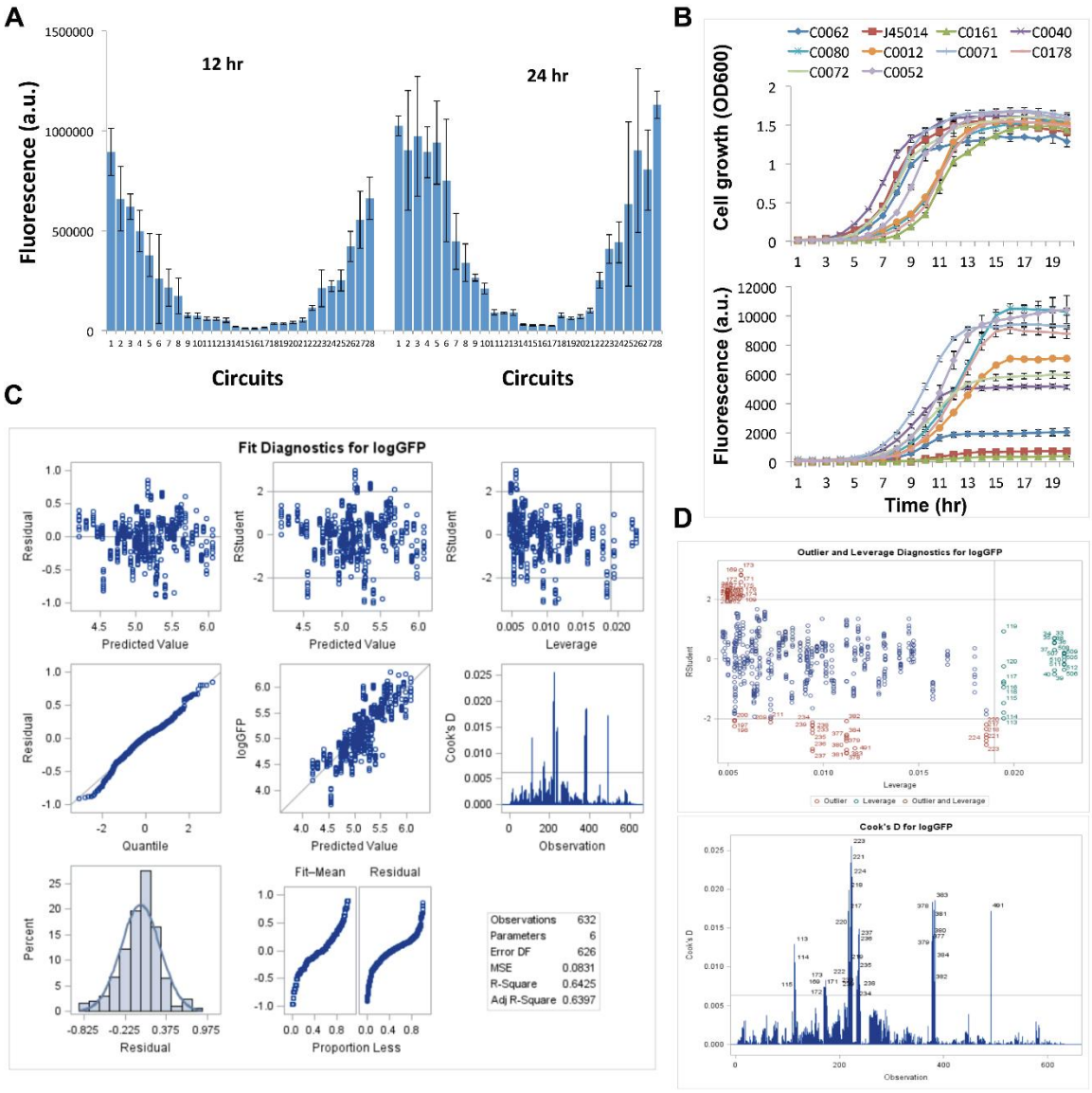


Figure 2.9 GFP Expression at 12 and 24 Hours; Cell Growth Rates for *X-GFP* Circuits; Fit Diagnostics for the Comprehensive Coding-ATR Model (A) GFP fluorescence of 28 gene circuits in Figure 2 was plotted at 12 and 24 hr. Similar fluorescence expression pattern was observed at 12 and 24 hr, and fluorescence was stronger at 24 hr. For stable protein expression, only 24 hr data are shown in this work unless specified. (B) Ten different gene circuits with different fluorescence expression levels (high, medium and low) are selected to test their growth rates under the same condition. Top: cell growth curves. All the samples reached the stationary phase after ~12 hr. Bottom: time course of the fluorescence for the ten selected circuits. Fluorescence becomes stable before 16 hr. Although the gene expression in the circuit influenced the time of cells going to the exponential phase, all the samples went to stationary phase with similar optical density (OD) value after ~12 hours. Cells with J45014 circuit (red curve) reach steady-state OD

about 4 hours earlier than cells with C0161 (dark green), while both show similar GFP expression levels. On the other hand, cells with C0012 circuit (yellow) and cells with C0161 arrive stationary phase almost simultaneously, however, their GFP expression is remarkably different. These results suggest that the time to reach steady-state OD for each strain has little explanatory effect on the fluorescence differences. OD and fluorescence were measured by plate reader with 96-well plates. Data indicates mean ± SD of three independent replicates. (C) Fit diagnostics the comprehensive coding-ATR model in Figure 2.3F. The Predicted value-Residual plot indicates that there is no apparent trend for the residuals, and the data is roughly normally distributed (Quantile-Residual plot and histogram), and the variables in the model explain most variation in the response variable from the residual-fit result (Fit-Mean and Residual). Leverage-RStudent plot and Cook's D value indicate there are some outliers and high-leverage observations, which may influence the model. Overall, the generated model has a good fitting of the experimental data. (D) Outlier and leverage diagnostics for the response (GFP). High-leverage data points and outliers are labeled out. Of the outliers, most of them are corresponding to a specific circuit, such as outliers 217~224 corresponding to the tricistronic circuit (*promoter-luxR-appY-GFP*, has 8 data points). Observations with high leverage such as 505~512 are corresponding to the circuit *promoter-GFPZif23_GCN4*. Moreover, some outliers are also high-leverage observations.

A central goal of synthetic biology is to develop genetic circuits to program cell behaviors in a predictable way. With the increasing complexity of integrated multi-layer circuits, organization of specific bio-components and circuitry structure design become extremely important for functionality[50,121,125]. The tool we provide here could serve as a much-needed quantitative guide for rational design and optimization of gene expression for large genetic circuits.

2.4 Materials and experimental methods

2.4.1 Strains, media, and culture conditions

All cloning experiments and fluorescent measurements were performed in *Escherichia coli* DH10B (Invitrogen). Synthetic toggle switches (T_S28, T_S67 and T_WT) were tested in *E. coli* K-12 MG1655 strain with lacI-/-[95]. Cells were cultured in liquid or solid Luria-Bertani (LB) broth medium with 100 mg/ml ampicillin at 37 °C. Chemicals AHL (N-(b-etocaproyl)-L-homoserine

lactone), IPTG (isopropyl b-D-1-thiogalactopyranoside), and aTc (anhydrotetracycline) were dissolved in ddH2O and diluted into indicated working concentrations. Cultures were shaken in 5 mL and/or 15 mL tubes at 220 rotations per minute (r.p.m).

2.4.2 Plasmid construction

Most genes are obtained from iGEM Registry (http://parts.igem.org/Main_Page). These genes are often used in synthetic biology projects, including transcriptional factors, quorum-sensing components, and other functional genes. Plasmids were constructed using standard molecular biology techniques and all genetic circuits were assembled based on standardized BioBrick methods. As an example, construct *Promoter-TetR-GFP* is composed of five BioBrick standard biological parts: BBa_J23104 (constitutive promoter, CP), BBa_B0034 (ribosome binding site, RBS), BBa_C0040 (tetR), BBa_E0040 (green fluorescent protein, GFP) and BBa_B0015 (transcriptional terminator). To produce RBS-TetR module, plasmid containing *TetR* was digested by *XbaI* and *PstI* as the insert fragment while RBS vector was cut by *SpeI* and *PstI*. Both fragment and vector were separated on 1% TAE agarose gel electrophoresis and purified using PureLink gel extraction Kit (Invitrogen). Purified fragment and vector were then ligated by T4 DNA ligase (New England Biolabs, NEB). The ligation products were further transformed into *E. coli* DH10B and plated on LB agar plate with 100 mg/ml ampicillin for screening. Finally, plasmids extracted by GenElute HP MiniPrep Kit (SIGMAALDRICH) were confirmed through gel electrophoresis (digested by *EcoRI* and *PstI*) and DNA Sequencing (Biodesign sequencing Lab, ASU). Similar steps were carried out for subsequent rounds of cloning to assemble the whole construct. Also, 17 transcriptional factors with varying GC content and sizes used in Figure 2.3D were amplified from E. coli genome with designed primers. Synthetic sequences were randomly generated with the same length (200 bp) but various GC contents (28%-67%). Sequences with RBS-features (AGGAGG) were redesigned to exclude its translation potential. All synthetic sequences and primers were synthesized as custom DNA oligos or gBlocks gene fragments from Integrated DNA Technologies (IDT). To express consistently in the cell, all constructs were finally subcloned into pSB1A3 vector prior to the test.

2.4.3 Minimum free energy calculation

All minimum free energy (MFE) of mRNAs were computed on Nucleic Acid Package (NUPACK) web server[126]. Specifically, we chose Serra and Turner parameter set as the RNA energy parameter and set 37°C, 1.0 M Na^+ and 0 M $Mg2^+$ to be the prediction algorithm[113]. $\Delta G_{5'ATR}$ and $\Delta G_{3'ATR\_100}$ were calculated from sequence including ATR (with or without RBS), and the two scar sequences introduced during cloning process. $\Delta G_{-70-+38}$ is obtained from 70 nt upstream sequence and 38 nt downstream around ATG (+1) codon of GFP gene.

2.4.4 RT-qPCR

Total RNA was extracted from three individual cell cultures (1.5 mL exponentially growing cell cultures, fresh cultures) for each construct in Figure 2.1B using Trizol (Invitron). DNase I (NEB) was used to remove traces of genomic DNA and then the total RNA was further purified using purelink RNA Mini Kit (Life technologies), and the eluted total RNA was quantified using BioTek's Synergy H1multi-mode Reader. cDNA was synthesized from RNA using an iScript cDNA synthesis kit and random primers (Bio-Rad). The reaction volume is 20 uL and ~1 ug RNA were used for reaction. Concentrations of cDNA are then quantified by qPCR using iTaq Universal SYBR Green Supermix (Bio-Rad) with the iQ5 Real-Time PCR detection system (Bio-Rad). Prokaryotic 16S rRNA was employed as endogenous control. Primers (IDT) used for amplifying 16S rRNA: 5'- AATGCCACGGTGAATACGTT-3' (rrnB, forward, starting at the 1361st nucleotide), and 5'- ACAAAGTGGTAAGCGCCCT-3' (rrnB, reverse, starting at the 1475th nucleotide) (Limet al., 2011). Two pairs of primers were designed to amplify GFP are P1: 5'-CAGTGGAGAGGGTGAAGGTGA-3' (forward, starting at the 87th nucleotide); and P2: 5'-CTGTACATAACCTTCGGGCAT-3' (reverse, starting at the 283th nucleotide); P3: 5'-AGACACGTGCTGAAGTCAAG-3' (forward, starting at the 320th nucleotide); and P4: 5'-TCTGCTAGTTGAACGCTTCCAT-3' (reverse, starting at the 539th nucleotide). qPCR result is analyzed using Bio-rad CFX Manager software version 3.1. Each sample was performed with two replicates for both 16S rRNA and GFP cDNAs, and gene expression was normalized to 16S rRNA. Delta Ct values were calculated ($Ct^{target} - Ct^{16S}$) and compared with the biological control

(*Constitutive promoter-RBS-GFP*) to calculate the relative GFP mRNA concentrations. The minimum information for publication of quantitative real-time PCR (MIQE) is also provided in Table 2.2.

Table 2.2 Minimum Information for Publication of Quantitative Real-Time PCR (MIQE)

Experimental Design	
Definition of experimental and control groups	Experimental: GFP mRNA levels for the circuits (*X-GFP*) in Fig. 1; Control: GFP mRNA levels for the circuit without inserted *X*. See text for more details.
Number within each group	Experimental: 5; Control: 1.
Assay carried out by the core or investigator' s laboratory?	Reverse transcription and qPCR carried out in the synthetic biology lab of SBHSE in Arizona State University.
Acknowledgement of authors' contributions	See Author Contributions statement
Sample	
Description	Exponentially growing *E. coli* cell cultures (fresh)
Volume/mass of sample processed	1.5 mL
Microdissection or macrodissection	Not applicable
Processing procedure	Trizol (Invitrogen)
If frozen - how and how quickly?	No frozen
If fixed - with what, how quickly?	No fixed
Sample storage conditions and duration (especially for FFPE samples)	Fresh sample, without storage
Procedure and/or instrumentation	Purelink Total RNA Purification protocol
Name of kit and details of any modifications	Purelink Total RNA Mini Kit (life technologies), Cat. #: 12183018A
Source of additional reagents used	Trizol, Cat. #: 15596026; 2-Mercaptoethanol, Cat. #: M3148
Details of DNase or RNase treatment	DNase NEB, Cat. #: M0303S
Contamination assessment (DNA or RNA)	PCR to amplify the *GFP* fragment for total-RNA before and after DNase treatment.
Nucleic acid quantification	Spectrophotometry
Instrument and method	BioTek's Synergy H1 multi-mode Reader
Purity (A260/A280)	1.91~2.13
RNA integrity method/instrument	Running agarose gel to check RNA quality and degradation
RIN/RQI or Cq of 3' and 5' transcripts	Not applicable
Inhibition testing (Cq dilutions, spike or other)	Not applicable
Reverse Transcription	
Complete reaction conditions	25°C 5 min for priming, then 46°C 20 min for reverse transcription, and finally 95°C 1 min for inactivation
Amount of RNA and reaction volume	~1 μg RNA, 20 μL reaction volume
Priming oligonucleotide (if using GSP) and concentration	Random primers
Reverse transcriptase and concentration	Reverse Transcription Supermix
Manufacturer of reagents and catalogue numbers	iScript cDNA synthesis kit, Bio-Rad , Cat. # 1708841
Storage conditions of cDNA	-80 degree refrigerator
qPCR Target Information	
Target gene	GFP (16S rRNA was used as a reference gene)
Sequence accession number	Not applicable

Location of amplicon	+87 ~ +283 (P1:P2); +320 ~ +539 (P3:P4) of GFP
Amplicon length	197 bp for P1:P2; 219 bp for P3:P4.
In silico specificity screen (BLAST, and so on)	Designed using Primer-BLAST, NCBI
Pseudogenes, retropseudogenes, or other homologs?	None
Location of each primer by exon or intron (if applicable)	Not applicable
What splice variants are targeted?	No splice variant
qPCR Oligonucleotides	
Primer sequences	**16S rRNA Forward:** *5'-GAATGCCACGGTGAATACGTT-3'* **16S rRNA Reverse:** *5'-CACAAAGTGGTAAGCGCCCT-3';* **GFP-P1-Forward:** *5'-CAGTGGAGAGGGTGAAGGTGA-3';* **GFP-P2-Reverse:** *5'-CCTGTACATAACCTTCGGGCAT-3';* **GFP-P3-Forward:** *5'-AGACACGTGCTGAAGTCAAG-3';* **GFP-P4-Reverse:** *5'-TCTGCTAGTTGAACGCTTCCAT-3'*
Location and identify of any modifications	None
Manufacturer of oligonucleotides	IDT
qPCR Protocol	
Complete Reaction Conditions	Initial denaturation 95°C for 2 mins, followed by 40 cycles of 95°C 5 s (denaturation), and 60°C for 30 s (annealing, extension and fluorescence reading).
Reaction volume and amount of cDNA/DNA	10 µL reaction volume; 0.1 µL cDNA
Primer, (probe), Mg++ and dNTP concentrations	Detailed information can be found from iTaq Universal SYBR Green Supermix, Bio-rad, Cat. #: 172-5120
Polymerase identity and concentration	
Buffer/kit identity and manufacturer	
Additives (SYBR Green I, DMSO, etc.)	SYBR Green I
Manufacturer of plates/tubes and catalog number	Bio-rad, Cat.#: HSP3801
Complete thermocycling parameters	95°C for 2 mins; 40 cycles of 95°C 5 s; 60°C for 30 s
Reaction setup (manual/robotic)	Manual
Manufacturer of qPCR instrument	iQ5 Real-Time PCR detection system (CFX384TM, Bio-Rad).
qPCR Validation	
Specificity (gel, sequence, melt, or digest)	Primers were designed specifically for GFP and 16S rRNA
For SYBR Green I, Cq of the NTC	40
PCR efficiency calculated from slope	93.82%
Calibration curves with slope and intercept	Cq values for serial 1:10 dilutions: 23.1907079; 26.1641088; 29.6333439; 33.1484379; 37.0967272 Slope: -3.4796; Intercept: 22.887
R^2 of standard curve	0.99757
Linear dynamic range	Dilutions spanned four orders of magnitude
Data Analysis	
qPCR analysis program (source, version)	Bio-rad CFX Manager software Version 3.1

Method of Cq determination	Auto threshold
Outlier identification and disposition	None
Justification of number and choice of references genes	16S rRNA has a consistent expression in log phase *E. coli* cells and thus is used as a reference.
Description of normalization method	Each sample was normalized to the endogenous 16S rRNA reference, followed by normalization to the control group without gene insert before GFP gene (Fig. 1b).
Number and concordance of biological replicates	Three biological replicates for each sample
Number and stage (RT or qPCR) of technical replicates	Two technical replicates
Repeatability (intra-assay variation)	See Fig. 1c and Supplementary Figure for standard deviation between biological replicates
Statistical methods for result significance	None

2.4.5 Flow cytometry measurements

All confirmed constructs were re-transformed into DH10B strain. Single colonies were picked and cultured in 4 mL LB medium (100 mg/ml ampicillin) for 24 hr at 37°C for testing. Flow cytometry measurements were performed using Accuri C6 flow cytometer (Becton Dickinson) and all samples were analyzed at twelve hour and 24-hour time points, and the two time points showed similar GFP expression pattern (Figure 2.9A). GFP excitation: 488 nm, and emission: 530 ± 15 nm. All data were collected in a log mode. Data files were further analyzed by MATLAB (MathWorks). All the fluorescence data are collected by flow cytometry unless specified, and the fluorescence was not normalized against cell density because we measured the fluorescence of single cells, instead of the population, so the fluorescence value is not directly correlated with population density. 20,000 individual cells were analyzed for each sample at a slow flow rate.

2.4.6 Hysteresis experiment

All synthetic toggle switch plasmids (T_S28, T_S67 and T_WT) were transformed into K-12 MG1655 strain with lacI-/-, and cells cultured overnight in LB medium. T_WT plasmid has been used in previous study[96]. We prepared two pre-cultures with two initially different stable steady states, i.e., low GFP state (OFF) without inductions and high GFP state (ON) induced with enough aTc. The two cells were then inoculated into media containing an aTc concentration range so that cells with different initial conditions were grown in identical conditions. Specifically, for OFF-ON experiment, samples were diluted evenly into 5 ml polypropylene round-bottom tubes

(Falcon) and induced with different amounts of aTc. Fluorescence was then measured at 6, 8 and 21 hr time points to monitor the fluorescence level. In our experiment, we found the intensity of fluorescence became stable after ~8 hr induction. For the ON-OFF experiment, cells were induced with 40 ng/ml aTc initially to prepare the initial ON cells and fluorescence was measured at 8 hr to ensure they were fully induced. The initial ON cells were then collected by low-speed centrifugation, washed once to remove the inducer, resuspended with LB medium, diluted, and transferred into fresh medium with various aTc concentrations at 1:100 ratio. Flow cytometry measurement was performed for each sample after 6, 10 and 18 hr culturing, respectively. Data shown in Figure 2.8 are 18 hr results.

2.4.7 Sample preparation and microscopy

Single colonies were picked and grew at 37°C in liquid LB medium. After 24 hours, 1 mL cells were collected and spun down at 2500 g for 5 min, washed with 1x phosphate buffer solution (PBS), and resuspended by 200 mL 1xPBS. 10 uL of concentrated cell solution was placed on glass microscope slides and images were captured with a Nikon Ti-Eclipse inverted microscope (magnification 40x). GFP was visualized with an excitation at 472 nm and emission at 520/35 nm using a Semrock band-pass filter. The exposure time for each sample is kept the same.

2.4.8 Growth curve assay

Ten different gene circuits with different fluorescence expression levels (high, medium and low) are selected to test their growth rates under the same condition. Single colonies harboring circuit plasmid were picked up and diluted into 4 mL LB medium, from which 300 mL were transferred into 96-well sterile plate. A negative control with only LB medium was also prepared. Optical density (OD600) and fluorescence (excitation: 485 nm; emission: 530 nm) were measured every 30 minutes by plate reader (BioTek) over 20 hours with shaking platform and temperature control (37°C). Three random colonies were picked up and triplicate wells were measured for each sample. Our results indicated that gene expression in the circuit influenced

the time of cells going to the exponential phase, but all the samples went to stationary phase with similar OD value after ~12 hr (Figure 2.9B). For stable protein expression, we chose the 24 hr data point in this study unless specified.

2.5 Quantification and statistical analysis

2.5.1 Statistical analysis and comprehensive model development

To investigate the correlation between GFP expression and sequence characteristics in different circuits with different genes and organizations, we performed multiple linear regression analysis using the classical statistical software SAS 9.4. Here, we mainly focused on five different independent variables including 5'- and 3'-ATR GC content (variable is computed as a percentage), 5'- and 3'- size (variable is computed as segment length), and ΔG-70-+38, all of which can be computed from the DNA sequence in each circuit. The dependent variable is GFP fluorescence measured by flow cytometry, which was transformed to log scale during analysis. Eight data points collected in three days were used for regression analysis in Figures 1 and 2, and twelve data points were collected in three different days for the 17 transcriptional factors insertion as non-coding ATRs, and all of the collected data points are imported to SAS for analysis.

All the information of the five variables is calculated from the specific DNA sequence. The 5'ATR includes the sequence from the scar right after the promoter to the scar right before the RBS of GFP. And the 3'ATR includes the sequence from the scar right after the GFP to the scar right before the terminator. The scar sequence is generated from the molecular cloning using biobrick modules, and the size is 6 or 8 nucleotides. GC content and size of ATRs are calculated using the web server Endmemo (http://www.endmemo.com/index.php). $\Delta G_{-70-+38}$ were computed using NUPACK web tool (http://www.nupack.org). Since the ΔG are negative values, log transformations were performed to the absolute value of ΔG, and then set to negative value. To build a comprehensive model for all the scenarios in Figures 2.3A-C (*GFP-X*, *X-GFP-Y*, and *X-GFP*), we introduced dummy values for some variables in some regression analyses for analytical convenience. For example, construct *GFP-X* (Figure 2.3C) has no varied 5' ATR (only a RBS and

scar sequence), and its GC content value is set to 0.04 instead of 0. Similarly, $\Delta G_{5'ATR}$ is set to -0.05 for constructs without 5'ATR, and $\Delta G_{3'ATR\_100}$ is set to -0.00001 for constructs without 3'ATR. These dummy values do not significantly influence model fitting efficiency.

We first use scatter plot to display the relationship between GFP and each of the variables we are interested, without any data transformation. As shown in Figure 2.10, the data has a large variability ranging from 21,000 to 1900,000 (arbitrary unit), and the fit without transformation is weakly linear and heteroscedastic. It would be problematic to use linear data for regression because of the inconstant variance from the data. However, the log is a variance stabilizing transformation, and it clearly reduced changes in variability of the data along the x-axis (Figures 2.3A-D). Furthermore, transformed data conforms to a nearly normal distribution (Figure 2.9C), more easily enabling us to perform multiple regression analysis to find a quantitative estimation of the relationship between GFP and the other three or five variables together.

To explore possible mechanistic basis of ATR regulation, we developed a comprehensive linear model based on the sequence dependent energetic changes during the polycistronic mRNA folding and translation and the costs of protein biosynthesis. The biophysical model was based on previous pioneer work characterizing the relationship between free energy changes and protein translation initiation[112,114,127]. We calculated the free energies for 5' ATR and the first 100 nucleotides of 3' ATR ($\Delta G_{5'ATR}$ and $\Delta G_{3'ATR\_100}$) using NUPACK. Since all the energy terms are negative values, absolute values were first acquired for each of them and then set to negative values for data analysis. The constant Gm is set to 1, and for cases of non-coding ATRs, the coefficients for j and $\Delta G_{3'ATR\_100}$ are set to 0, owing to a lack of 3' ATRs.

To find the linear comprehensive coding-ATR model having the best prediction of dependent variable from the independent variables, we performed stepwise regression with the five variables: $\Delta G_{5'ATR}$, $\Delta G_{3'ATR\_100}$, 5' ATR size, 3' ATR size and $\Delta G_{-70-+38}$. Stepwise regression is an automated tool for model selection through adding the most significant variable or removing the least significant variable as needed for each step (the significance level for variable entry or stay is 0.05). From the sequence of generated models, the selected model is chosen based on the lowest Akaike information criterion. Results showed that all the five variables are statistically

significant for the best prediction of GFP expression in the comprehensive coding-ATR model, and explains 63% of GFP variations. It is necessary to note that the negative correlation between protein abundance (c) and the sum of energetic terms ($\Sigma\beta_x\Delta G_x$,) in the equation is already reflected in the coefficients of each term.

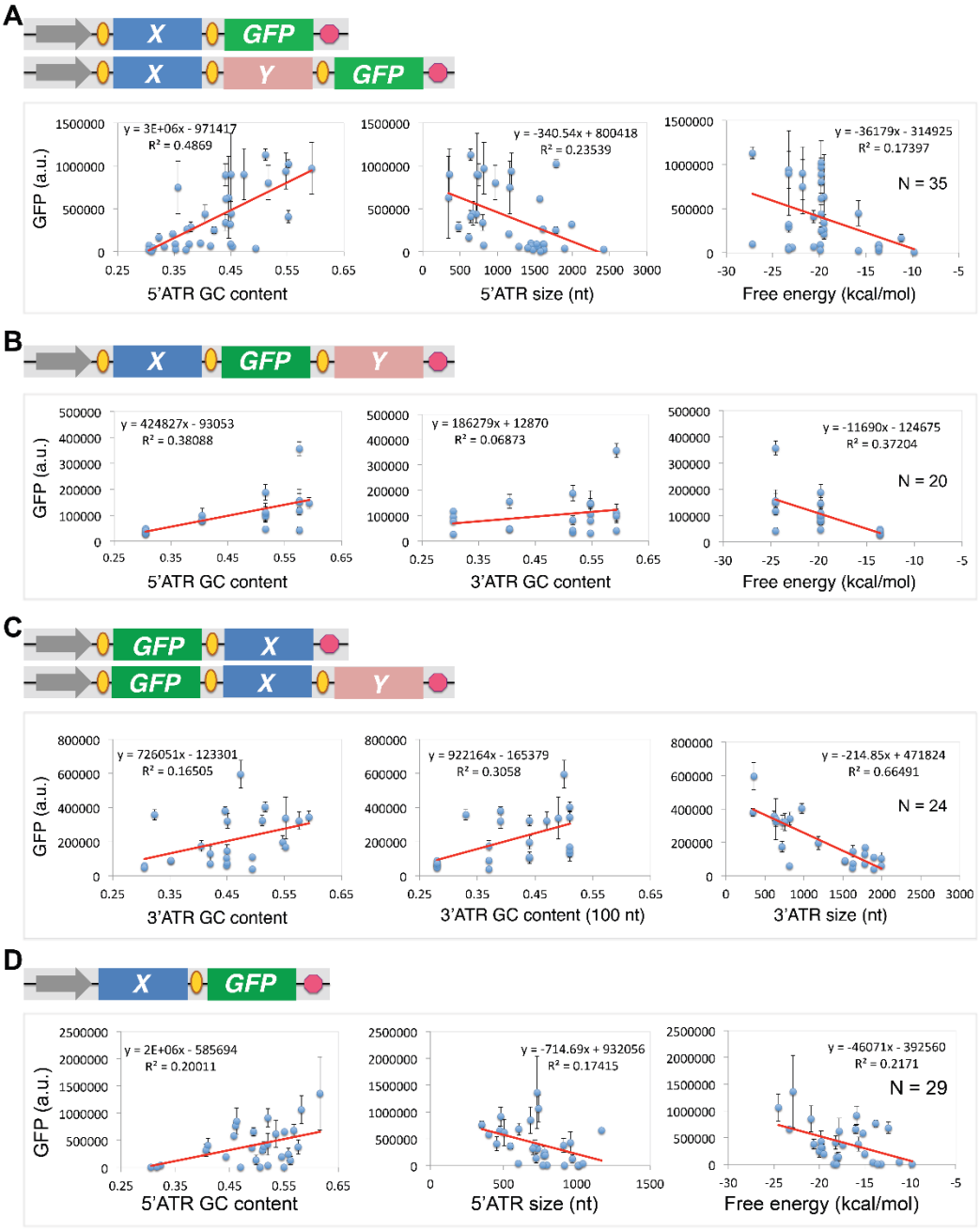


Figure 2.10 Linear Plots with GFP and Variables in Figure 2.3A-D (A) Linear plots for the scenario 1 in Figure 2.3. Top: Schematic representation of synthetic polycistronic gene circuits *X-GFP*.

Bottom: Scatter plots of GFP and 5' ATR GC content, 5' ATR size, and Free energy. Red line is the fitted regression result. Equation and R^2 are also displayed for each fit. (B) Linear plots for the scenario 2 in Figure 2.3. Top: Schematic representation of synthetic polycistronic gene circuits *X-GFP-Y*. Bottom: Scatter plots of GFP and 5' ATR GC content, 3' ATR GC content, and Free energy. (C) Linear plots for the scenario 3 in Figure 2.3. Top: Schematic representation of synthetic polycistronic gene circuits *GFP-X*. Bottom: Scatter plots of GFP and 3' ATR GC content, 3' ATR GC content (100 nt), and Free energy. N is the number of circuits. (D) Linear plots for non-coding ATRs results in Figure 2.3D. Top: Schematic representation of synthetic polycistronic gene circuits with non-coding X genes (29 circuits in total). Bottom: Scatter plots of GFP and 5' ATR GC content, 5' ATR size, and Free energy. Red line is the fitted regression result. Equation and R^2 are also displayed for each fit.

The fitting diagnostics indicated that there is no apparent trend for the residuals, and the data is roughly normally distributed, and the variables in the model explain most variation in the response variable from the residual-fit result (Figure 2.9C). The predicted value by observed GFP plot (Predicted Value - logGFP) reveals a reasonably successful model for explaining the variation in GFP for most of the circuits (Figure 2.3F, left panel and Figure 2.9C). The predicted responses (logGFP value) are calculated according to the generated linear regression model, with the corresponding inputs from each circuit. And a plot of predicted GFP against experimentally observed GFP values are then generated to evaluate and visualize the model-fitting efficacy (Figure 2.3F). If the model predicted values and observed values agreed perfectly (R^2 = 100%), all the data points would fall on the dotted diagonal line of the squares. However, several outliers in the combined model are also observed and some observations with high leverages might also be overly influencing the fit result (Figure 2.9D). Of the outliers, most of them are corresponding to specific circuits, such as outliers 217~224 corresponding to the tricistronic circuit (*promoter-luxR-appY-GFP*, has 8 data points). Observations with high leverages such as 505~512 are corresponding to the circuit *promoter-GFP-Zif23_GCN4*. Moreover, some outliers are also high-leverage observations. Given the data sample size (N = 632), the original data collection, and the overall data-fitting efficacy, we here didn't exclude the outliers or data with very high leverages (although that would improve the model-fitting efficacy).

Similar analysis was also applied to the data with non-coding ATR, and results showed that 5'ATR size and folding energy $\Delta G_{5'ATR}$, local mRNA folding energy $\Delta G_{-70-+38}$ are crucial for the best prediction of GFP expression in the comprehensive non-coding ATR model. Furthermore, the three variables together explain two-thirds of GFP variations in those synthetic circuits (Figure 2.3F, right panel). The model and coefficients were also validated by another statistical software XLSTAT (version 2017.4). Based on the comprehensive non-coding model (Figure 2.3F), we then employed XLSTAT to predict GFP expression (mean and standard deviation) in circuits regulated with synthetic ATRs having either different GC contents or different in Figures 2.7C and 2.7D. Model predicted GFP (Figure 2.6C) has a similar expression trend with experimental results (Figures 2.7C and 2.7D).

We also performed k-fold cross validation to further assess our model performance (k = 10). The entire dataset was randomly partitioned into a training dataset, a validation dataset, and a testing dataset. The model was built based on the training dataset (50% of the original data) and then validated on the other 25% dataset, and finally was used to assess the performance on the testing dataset (25% of the original data). The selection method is stepwise, selection criterion is Schwarz' Bayesian Criterion (SBC) and stop criterion is Akaike's Information Criterion (AIC). We performed 10 times of the 10-fold validation and found that the coefficients for each variable and intercept as well as R^2 are very close to the above comprehensive model. Moreover, the standard deviation for the square root of mean squared error (RMSE) from the 10 repeats of 10-fold validation is very small (0.0064 for coding ATR, and 0.0128 for non-coding ATR), suggesting the model we built has a decent prediction accuracy and consistency.

In summary, we have demonstrated that the coding and non-coding adjacent transcription regions have remarkable effects on regulating GFP expressions in synthetic operon-based gene circuits (Figure 2.3). Furthermore, we can use a general biophysical model with sequence-dependent energetic changes to quantify the ATR regulation on gene expression. In this study, we mainly investigated five factors involved in ATR regulation: 5' and 3' ATRs free energies $\Delta G_{5'ATR}$ and $\Delta G_{3'ATR\_100}$, transcriptional sizes and the mRNA folding energy near the GFP starting codon. It is possible that there are some other unknown or uncharacterized factors

influencing GFP expression, such as the codon degeneracy for the coding ATRs. Furthermore, there may have some special local secondary or higher structures in some ATRs, which may impact the degradation or translation of GFP.

2.5.2 Deterministic model construction and prediction for the logic gate

In the four logic gates, GFP expression depends on the relative concentrations of activator (LuxR) and repressor (TetR or LacI) produced from a constitutive promoter. AHL binds with LuxR protein to activate *pLux/tet* transcription and aTc can block TetR repression to *pLux/tet*. Since the two sets of logic gates (LT/TL and LI/IL) are constructed similarly and described by the same deterministic equations, we here only explain the technical details for the gate LT. The model was built based on our previous work[50]. we derived the following ordinary differential equations for intercellular concentrations of LuxR (U), TetR (R) and GFP (G):

$$\frac{dU}{DT} = (k_0 + \alpha_1) - d_1 \times U \quad (1)$$

$$\frac{dR}{dt} = (k_0 + \alpha_2) - d_2 \times R \quad (2)$$

$$\frac{dG}{dt} = \left(c_1 + \frac{K_1 C}{C + K_n}\right) \times \frac{1}{K_r^{nt} + (R \times F)^{nt}(R \times F)} - d_3 \times G \quad (3)$$

$$f = \frac{AHL^{ni}}{AHL^{ni} + K_i^{ni}} \quad (4)$$

$$C = \frac{(f \times U)^2}{K_d} \quad (5)$$

$$F = \frac{1}{K_t^{nr} + ATC^{nt}} \quad (6)$$

The first two equations describe the concentrations of LuxR and TetR, both of which are driven by a constitutive promoter at a constant level (k_0). α_1 and α_2 are constants used to describe the relative changes of LuxR and TetR production, owing to the position changes in the And-gate circuit. d_1 and d_2 are the degradation rates for the LuxR and TetR protein, respectively. The third equation describes the concentration of GFP, which is determined by the relative concentrations of LuxR and TetR. LuxR binds to AHL molecules and forms the active LuxR monomers in the form of (LuxR-AHL), when the AHL concentration reaches a certain threshold (quorum-sensing mechanism). So the fraction of LuxR monomers (f) bound by AHL can be described by Equation

4, where ni is the binding cooperativity (Hill coefficient) between LuxR and AHL, and Ki represents the dissociation constant between LuxR and AHL. LuxR needs to form a dimer to bind the promoter and activate transcription, so the concentration of the functional LuxR dimer (C) that binds to the hybrid promoter pLux/tet and activates its transcription can be described by Equation 5, where K_d is the is the dissociation constant for dimerization. Thus, GFP expression driven by LuxR and inducer AHL is represented by the first part of Equation 3. C_1 is the basal mRNA expression without LuxR protein; K_1 is the production rate; and K_n is the dissociation constant between C and *pLux/tet* promoter. TetR protein can bind and inhibit GFP transcription, and the inhibition can be repressed by inducer aTc. So high GFP expression is achieved in presence of high doses of aTc, and vice versa (Equation 6). The second part of Equation 3 describes TetR inhibition to GFP expression, under induction of aTc. And the third part of Equation 3 is the degradation of GFP.

The three ordinary differential equations were used to model the two sets of AND-gate circuits: LT and TL, LI and IL. For each of the two sets, most parameters should be the same except a1, a2, c1, and Ki. Based on the parameter used in our previous studies[50], we used the following parameters in our simulations: $k_0 = 1.0$, $d_1 = 0.2$, $d_2 = 0.2$, $d_3 = 0.2$, $c_1 = 0.002$ (for TL) or 0.08 (for LT), $K_1 = 1.7$, $K_n = 4.4$, $K_d = 13$, $K_t = 400$, $K_r = 3.2$, ni = 1.2, nt = 2, ni = 1.2, nr = 2. For circuits LI and IL, $c_1 = 0.002$ (for IL) or 0.05 (for LI), $K_t = 1000$, and the other parameters are the same.

From our comprehensive linear model, we calculated that LT has more LuxR than TetR production (Table 2.1), so the basal expression c_1 is set to a bigger value in LT model. K_i has little effect on the shape of the GFP dynamic curves, but determines the AHL concentration producing half conversion of LuxR monomers into LuxR-HSL complexes (half GFP activation). So the K_i value in the model is acquired from the experimental data. Through changing the relative expression of LuxR and TetR (i.e. α_1 and α_2), we can modulate GFP production dynamics (Figure 2.6E). To predict the GFP responses in circuit TL with AHL and aTc inductions, we use the parameter α_1 and α_2 in LT as a control to tune the parameter α_1 and α_2 in TL. According to the linear model calculations, the production rate for LuxR in LT and TL almost doesn't change, but

production rate of TetR in TL increases by ~93% (Table 2.1). For example, we set the production rates for LuxR and TetR in circuit LT to 1.0 ($k_0 + \alpha_1$) and 0.6 ($k_0 + \alpha_2$), respectively. So in the circuit TL, the two rates should be 1.0 ($k_0 + \alpha_1$) and 1.15 ($k_0 + \alpha_2$) based on calculations. For different doses of aTc induction, we allowed ~10% parameter variations for α_1 and α_2. We found that the model simulations have a good match with our experimental data. The parameters for α_1 and α_2 in TL and LT under different doses of aTc are listed below:

Circuit	aTc (0 ng/ml)	aTc (0.2 ng/ml)	aTc (2 ng/ml)	aTc (20 ng/ml)	aTc (100 ng/ml)	aTc (200 ng/ml)
LT	$\alpha_1 = 0$ $\alpha_2 = -0.3$	$\alpha_1 = 0$ $\alpha_2 = -0.4$	$\alpha_1 = 0$ $\alpha_2 = -0.38$	$\alpha_1 = 0$ $\alpha_2 = -0.3$	$\alpha_1 = 0$ $\alpha_2 = -0.35$	$\alpha_1 = 0$ $\alpha_2 = -0.25$
TL	$\alpha_1 = 0.1$ $\alpha_2 = 0.1$	$\alpha_1 = 0.1$ $\alpha_2 = 0.05$	$\alpha_1 = 0.1$ $\alpha_2 = 0$	$\alpha_1 = 0.1$ $\alpha_2 = 0.1$	$\alpha_1 = 0.1$ $\alpha_2 = 0.1$	$\alpha_1 = 0.1$ $\alpha_2 = 0.25$

2.5.3 Bifurcation analysis for the synthetic toggle switches

For the toggle switch model in Figure 2.8, we used the same mathematical model and most parameters in the Gardner et al paper[15]. Here we think the synthetic ATRs mainly influenced the TetR production rate, with low rate corresponding to T_S28 ($\alpha_1 = 400$, $\beta = 2.7$), medium rate corresponding to T_S67 ($\alpha_1 = 600$, $\beta = 3.0$), and high rate corresponding to T_WT ($\alpha_1 = 1000$, $\beta = 3.245$). All the other parameters are set the same as in Gardner et al paper. Bifurcation analyses are performed using XPP-AUTO software (www.math.pitt.edu).

CHAPTER 3

APPLICATIONS OF MACHINE LEARNING TECHNIQUES IN GENETIC CIRCUIT DESIGN

3.1 Introduction

Gene circuit engineering is a popular methodology in synthetic biology with real-world applications in biomanufacturing and biosensing[128–130]. The main goal is to construct synthetic gene circuits with biological components such as genes, noncoding RNA elements, promoters, and other small modules. Designed synthetic circuits can in turn mimic biological behaviors and even implement novel functions within or outside of living cells. This bottom-up approach has been widely used for applications in areas such as pharmaceutical development, fuel production, metabolic engineering, genome engineering, and biomedical applications[76,86,131,132].

There are typically two main strategies to engineer a gene circuit: monocistronic and polycistronic. The former construction method ensures independent expression of each gene driven by its private promoters, while the latter architecture, often referring to operon, simultaneously transcribes each gene to the same mRNA under a single promoter but follows separate translation process to produce the needed proteins[90,92,133]. The operon-based construction largely exists in prokaryotes, and it can be commonly found in eukaryotes and viruses. Since the polycistronic architecture requires fewer biological components, it could facilitate circuitry construction; therefore, it is widely used in gene circuit engineering[94,96,134].

Presently, an increasing number of studies are being conducted to investigate how the neighboring regions of a certain gene affect its expression in the polycistronic gene circuit. The results obtained so far indicate that the transcriptional distance and gene's position in the synthetic operon have significant impact on gene expression[97,98]. A recent study demonstrated that the features of adjacent transcriptional region (ATR) including GC content, size and local RNA free energy around ribosome binding site (RBS) have strong correlations with protein expression outcome[135]. The authors previously built a linear regression model that took into account several ATR features to predict the specific gene expression as well as the dynamics of complex gene circuits. However, building this type of biological model requires tremendous amount of data, and the resulting model had relatively low prediction accuracy[135]. Therefore,

powerful analytical models are needed to accurately predict the behavior of complex and synthetically engineered biological circuits.

Machine learning (ML) methods have a tremendous potential in performing complex data analysis in investigation of synthetic biological systems. The transcription and translation processes, for instance, are more complex in eukaryotes than in prokaryotes[136–138]. Eukaryotic hosts play an important role since they are able to produce larger proteins that require post-translational modification[139]. In terms of application of eukaryotes, several ML methods have been utilized to design regulatory regions with less possibility of unforeseen interactions, and also to optimize gene expression[140–144].

To date, ML methods are being utilized to discover how gene sequences map to biological functions; however, there is yet little work done to directly understand the involved biological pathways. In one such study, biologists rely on ML to engineer proteins using accelerated directed evolution. Random forests (RF) were found to be robust and computationally efficient on smaller datasets (i.e., fewer than 104 training examples), including the sample datasets often encountered in protein engineering research projects[145–148]. Deep learning draws much attention nowadays in multi-interdisciplinary research[149]. For instance, it has been discovered that deep learning is useful for building predictive models to understand genotypes' contribution to gene expression[150].

In this study, several ML methods were utilized to further investigate the ATR influence on gene expression in polycistronic gene circuits. Specifically, two distinct types of experiments were conducted. First, a regression model was built to predict gene expression, which yielded R^2 scores of 0.97 and 0.95 using RF and ANN, respectively, compared to the best score of 0.63 previously obtained using linear regression[135]. Second, the generated decision tree classifier models further confirmed the hypothesis regarding the influences of attributes of the neighboring genes on protein expression. Additionally, another classifier was created using GC content of each gene and GFP fluorescent outcomes to predict the synthetic gene circuit patterns. Both RF and ANN classifier models achieved 100% accuracy while identifying varying patterns of synthetic gene circuits. The models built are important tools that can help biologists to select influential

attribute sequences in synthetic gene circuit design with fewer trial and error experimental attempts.

The rest of the paper is organized as follows. Section 2 provides a brief background on synthetic gene circuit design and the biological data collection methodology. Section 3 presents details regarding the utilized machine learning models, data reparation, and the obtained results for both the regression and classification models. And finally, Section 4 is the Conclusions and Future Directions.

3.2 Synthetic gene circuit design

Specific details of the experimental protocols needed for engineering synthetic gene circuits and data collection can be found elsewhere[135]. Briefly, a synthetic gene circuit consists of various biological components, such as a promoter to initiate transcription process, a terminator to determine the end of transcription process, a ribosome binding site for ribosome anchoring to start the protein translation, and finally, the genes of interest (Figure 3.1).

In the experiments conducted in this work, the high-copy plasmid method was used to express all synthetic gene circuits. The cloning experiments were followed by the standard molecular biology techniques, and all synthetic circuits were assembled based on standardized biobrick cloning methods. The GFP fluorescence of each generated circuit was measured using a flow cytometer. A total of 20,000 individual cells were analyzed for each sample at the slow flow rate, and eight replicate results for each circuit were collected for further analysis. In summary, 79 synthetic gene circuits were constructed with five different gene patterns (sequences) using two or three genes: *X-GFP*, *X-Y-GFP*, *X-GFP-Y*, *GFP-X* and *GFP-X-Y* (Figure 3.1). The resulting fluorescence values were collected for further machine-learning analysis. To reiterate, in synthetic biology, GFP is commonly used as a reporter protein for the easy measurement of its expression level and the circuit dynamic performance.

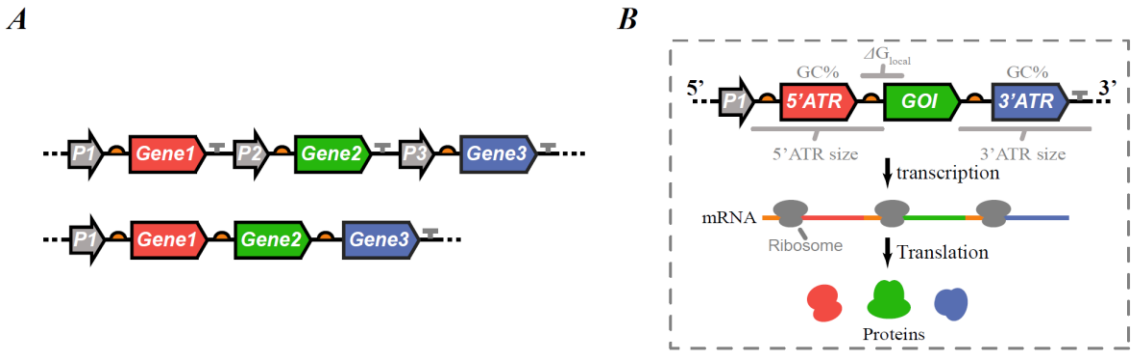


Figure 3.1 Synthetic Circuit Engineering Strategies and the Attributes Affect Polycistronic Gene Expression (A) Schematic showing the strategies of construction synthetic gene circuit. (Top) Monocistronic organization. Each gene's expression is independently initiated by their own promoters. (Bottom) Polycistronic organization. Transcription process is driven by one promoter to produce a single mRNA. The gray arrows represent the promoters; the orange ovals represent the ribosome binding site (RBS); the gray Ts represent the transcriptional terminators; the color boxes represent different genes. (B) Schematic illustrating the attributes affect polycistronic gene expression. A single mRNA is generated through transcription process following by separate translation processes to produce their own protein products. Attributes of adjacent transcriptional region (ATR) including 5' and 3' GC content, 5' and 3' size and local RNA free energy around RBS have strong correlations with protein expression outcome.

It must be noted that gene expression is influenced by overall stability and organization of biological components. Therefore, in some of the synthetic circuits, genes were placed in front of GFP, and in others, GFP were placed in the middle of the three-gene operons (e.g. *X-GFP-Y*) or proximal to promoter (e.g. *GFP-X*). In synthetic biology, the organization of synthetic operon facilitates construction of genetic cascades and decreases the number of biological components (such as the promoters and terminators) required for complex genetic circuits. The specific parameters which could significantly impact the functionality of synthetic gene circuits were *GC-content*, which is the percentage of nitrogenous bases in a DNA or RNA molecule that are either guanine or cytosine, and ATR sizes, positions, and degrees of stability such as *Rear_T_Size*, *Rear_T_dG*, *Front_T_Size*, *Rear_F300_dG, Front_GC_content,* etc.[135].

3.3 Machine learning experiments

A total of 8 experiments were conducted for each of the 79 synthetic gene circuits to obtain GFP Fluorescence values. Hence, a total of 632 rows of data were available for the regression models using both two-gene and three-gene patterns. Other than the GFP values and the circuit patterns, 9 attributes were used as shown in Table 3.1. The classification models relied on the GC content of the gene parts X, Y, and G, as well as GFP fluorescence values to predict the patterns of the gene circuit. Therefore, only three-gene patterns (X-GFP-Y, GFP-X-Y, and X-Y-GFP) were used for classifiers, resulting in a smaller subset of 408 rows of input. The following sections discuss the methods of data preparation, the conducted machine learning experiments including both the regression and classification models, and finally, the obtained results.

Table 3.1 Synthetic Circuit Design Attributes

Attributes	Description
dG	Local RNA structure free energy of GFP RBS
Rear_GC_content	3' ATR DNA GC content
Rear_100BP_GC	First 100 base pair (BP) of 3' ATR DNA GC content
Front_T_dG	5' ATR DNA free energy
Front_T_Size	5' ATR DNA total BP
Rear_T_dG	3' ATR DNA free energy
Rear_T_Size	3' ATR DNA total BP
Rear_F100_dG	First 100 BP of 3' ATR DNA free energy
Rear_F300_dG	First 300 BP of 3' ATR DNA free energy

3.3.1 One-hot encoding and standard scaling

To create the training dataset, first the categorical attributes were transformed into numerical data using the one-hot encoding technique[151]. The patterns indicating the position of gene segments were represented as vectors of all zeroes with the exception of a single ‘1’ to signify the position. Unlike the simple method of using digits to represent categorical attribute values, the one-hot encoding method has the advantage that the distances between digitized values are all the same in terms of number of bits.

ML models are sensitive to range and distribution of numerical attribute values[151]. Standard scaling was used to normalize numerical attributes, so they all have similar distribution

ranges, means, and standard deviations. For the ANN classification model, the distribution of the input data was normalized to the range [0,1] to further facilitate convergence.

3.3.2 Log-normal distribution

The original GFP fluorescence values collected in this study had a skewed distribution, but most ML techniques operate under the assumption of normal distribution of numerical data[151]. To comply with the requirement, log GFP values were used instead of the original GFP values[152]. As depicted in Figure 3.2, the log values are closer to Gaussian distribution and allow numerical techniques to build more accurate prediction models. The log-normal distribution is commonly used in all manners of numerical data analysis[153,154].

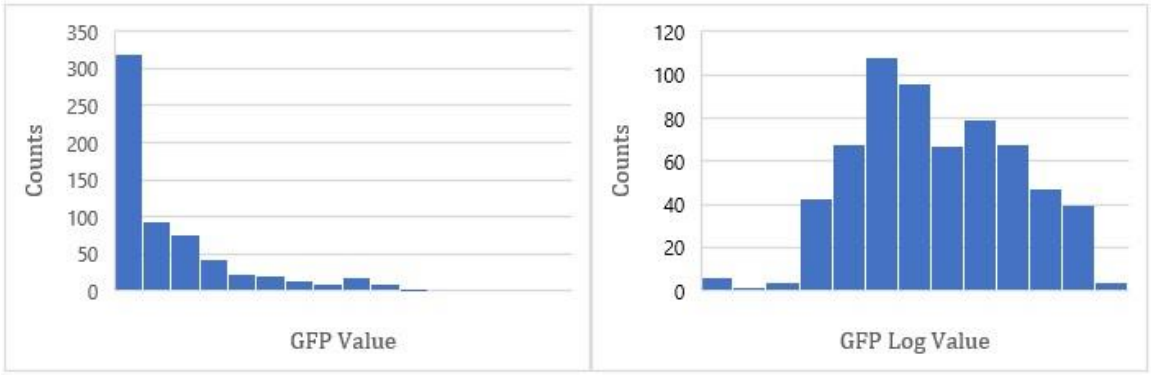


Figure 3.2 Untreated (left) or Log-transformed (right) GFP Distribution. X axis represents equal GFP value intervals and Y axis represents number of values fall in each interval. The pattern of GFP value with log transformation is closer to a normal distribution

3.3.3 PCA analysis and random splitting

Principal Component Analysis (PCA) was applied to better understand how the attributes contributed to the variance present in the dataset. For the regression model, it was shown that the dataset required at least 5 attributes that accounted for 95% of the total variance. To statistically validate the performance of the ML models, datasets in all experiments were randomly split into (80%, 20%) for training and testing purposes, respectively.

3.3.4 The machine learning Experiments

Two types of models were built to investigate various aspects of gene circuit design: First, the regression models were built using various algorithms such as Linear Regression, Decision Tree Regressor, Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Networks (ANN); and Second the classification models included RF, SVM with linear kernel and polynomial kernels, Stochastic Gradient Descent (SGD), K-means Clustering, and ANN. Since the dataset was not large, 5-fold cross-validation was applied to evaluate the overall performance of the model with random shuffling.

A total of six ANN models were built using different hyperparameters such as choice of activation functions, batch normalization, and dropout regularization (Table 3.2). The obtained results indicated that the rectified linear activation function (ReLU) performed much better largely because it does not saturate for positive values[151]. Further, it has been shown that if all hidden layers use the scaled version of the exponential linear unit (SeLU), the network will self-normalize, which solves the vanishing/exploding gradients problem. One pre-condition for using this activation function is that all attributes must all be normalized first[155]. Batch Normalization (BN) requires adding an operation before or after the activation function of each hidden layer, zero-centering and normalizing the input, and then scaling and shifting the results; therefore, it allows the models to learn the optimal scale and mean of each of the layer's inputs[113]. Adaptive moment estimation (Adam) combines Momentum optimization and RMSProp together to keep track of an exponentially decaying average of past gradients as well as that of the past squared gradients[156]. The utilized dropout technique has proven to be highly successful as a regularization technique[157]. In the ensuing experiments, the hyperparameter *p* or the dropout rate was set to 0.2.

3.3.5 Results

In the regression experiments, the following algorithms were utilized: linear regression, decision tree, RF, and SVM. The R^2 scores (Coefficient of Determination) and Root Mean Square Error (RMSE) values were computed to evaluate the performance of each algorithm. Among the algorithms, both the decision tree regressor and RF regressor achieved the R^2 score at 0.97. Using

5-fold cross-validation, the highest average R^2 score of 0.94 was obtained. The deep learning models achieved the R^2 score of 0.95 compared to the best score of 0.63 obtained in an earlier study (Figure 3.3).

Table 3.2 Six ANN Models that Use Adam Optimizer

	Activation Function	**Batch Normalization**	**Regularization**
Model 1	ReLU	No	
Model 2	SeLU	No	
Model 3	ReLU	Yes	
Model 4	SeLU	Yes	
Model 5	ReLU	No	Dropout
Model 6	SeLU	No	Dropout

Among ML algorithms that allow direct examination of the mechanisms of the models, decision trees best provided an in-depth understanding of the design attributes (i.e., via the generated decision rules) and their impact on gene circuit design. Five different decision trees, each with different random seeds, were generated and the five top-most used features in each case were recorded for inspection. Further analysis of the recorded features help biologist to better understand the significance of the chosen attributes in order to design more accurate and beneficial future experiments. Interestingly, the top 5 significant gene circuit design features identified in this work are in complete accordance with those identified by biologists based on the knowledge of the subject matter, namely, *Front_T_size*, *Front_T_dG*, *dG*, *Rear_F100_dG*, and *Rear_T_Size*.

In the second set of experiments, the following classification models were constructed to predict the patterns of synthetic gene circuits: RF, SVM with linear and polynomial (degree 3) kernels, SGD, and K-means clustering. For each classifier, the overall accuracy, confusion matrix, sensitivity, and specificity measures were computed to accurately assess the performance of the models. Both the RF and ANN models reached 100% accuracy in predicting gene circuit

patterns (Figure 3.4). In the experiments conducted with ANN's, some minor performance fluctuations were observed during different statistical runs of the learning models. By observing the learning curves (not shown here due to space limitations), random fluctuations occurred before the final convergence. However, the overall performances of all models are very robust, ranging in values from 0.96 to 1.00 in accuracy for the classification models, and 0.93 to 0.95 in R^2 score for regression models.

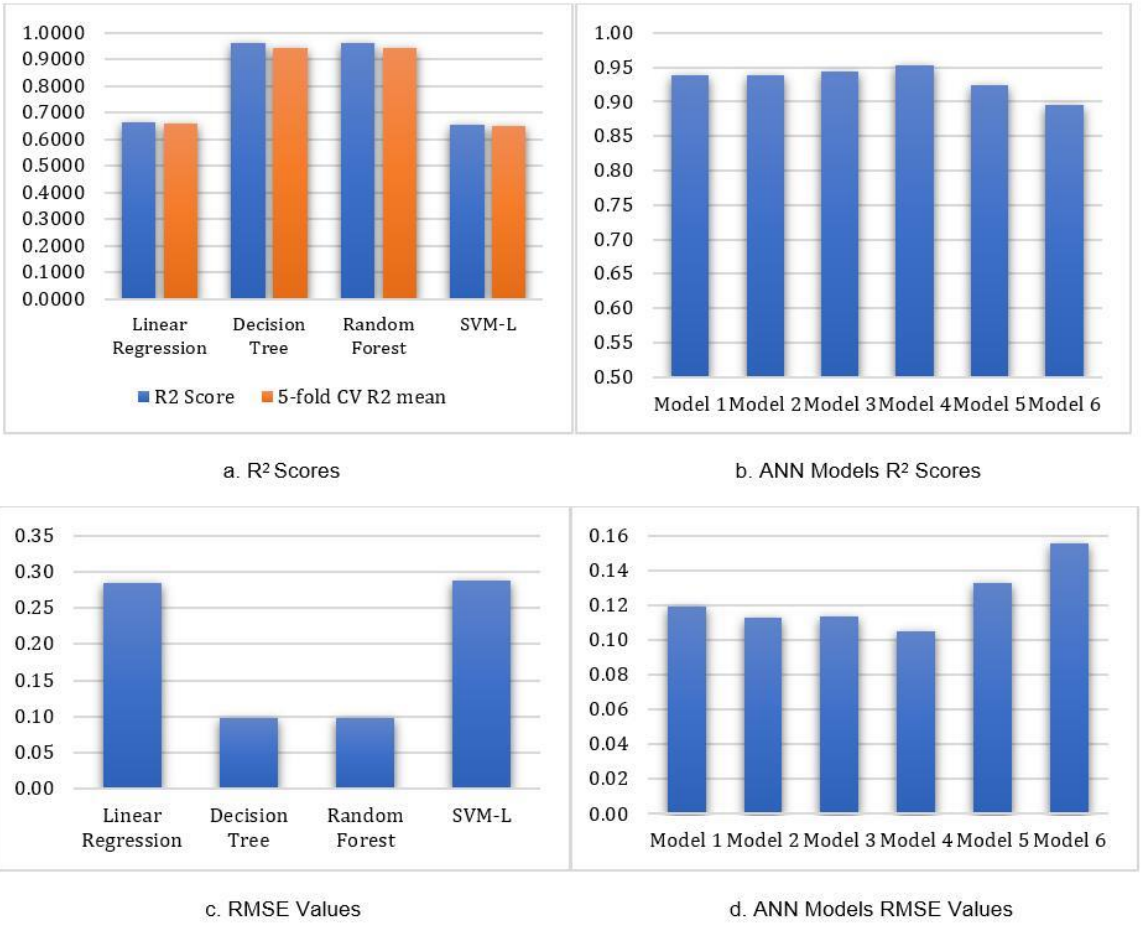
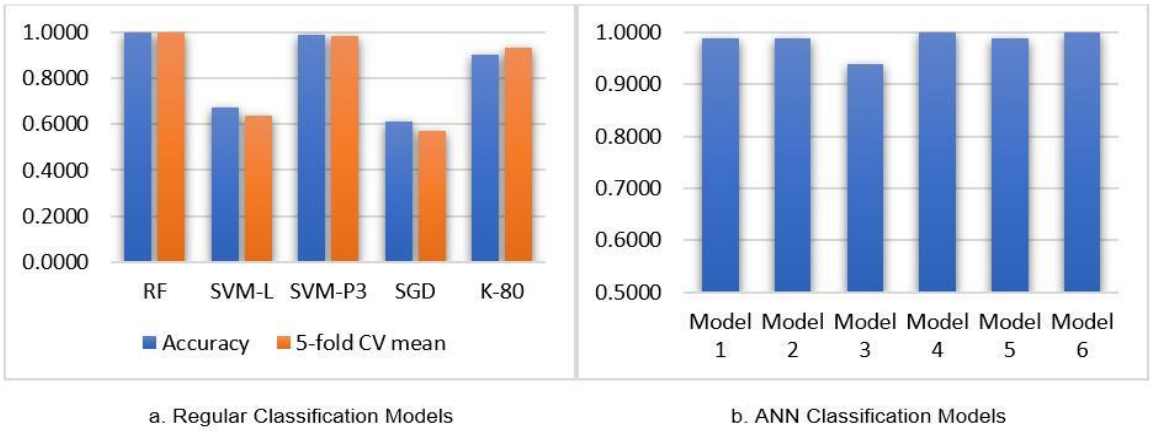


Figure 3.3 R^2 Scores and RMSE of Regression Models. For a & b, Y axis represents R^2 score (Coefficient of Determination) of each tested algorithm. It shows among regular algorithms, Decision Tree and Random Forest reached much higher R^2 score (>0.9) than the other two and 5 out of 6 ANN models reached R^2 score above 0.9. The second bar in a represent the mean of R^2 score of 5 folds cross-validation. For c & d, Y represent RMSE value of each tested algorithm. They present the coherent results where Decision Tree and Random Forest have the lowest value among traditional algorithms and ANN models have reached comparatively low values.



a. Regular Classification Models

b. ANN Classification Models

Figure 3.4 Performance of Various Classification Mode. Y axis represents overall prediction accuracy of each algorithm. The orange bars in the left figure represent the overall accuracy means of 5-folds cross-validation.

As observed in the computed confusion matrices (Table 3.3), the RF, SVM (polynomial kernel, degree 3), and K-means (*k*=80) classifiers learned the three gene circuit patterns of interest *G-X-Y*, *X-G-Y*, *X-Y-G* with highest degrees of accuracy. To further demonstrate the performance of the classifier models, Table 3.4 reports the computed specificity, sensitivity, overall accuracy values. Again, the top performing classifiers were RF, SVM-P3, and K-Means.

Table 3.3 Confusion Matrices for ML Models

Random Forest			
	GXY	**XGY**	**XYG**
GXY	17	0	0
XGY	0	39	0
XYG	0	0	26

SVM-L			
	GXY	**XGY**	**XYG**
GXY	0	7	10
XGY	0	31	8
XYG	0	2	24

SVM-P3			
	GXY	**XGY**	**XYG**
GXY	17	0	0
XGY	0	39	0
XYG	0	1	25

SGD			
	GXY	**XGY**	**XYG**
GXY	0	11	6
XGY	0	33	6
XYG	0	9	17

K-means_80			
	GXY	**XGY**	**XYG**
GXY	15	0	2
XGY	0	35	4
XYG	2	0	24

Table 3.4 Accuracy, Specificity and Sensitivity Scores for Each ML Model

Algorithms	**Random Forest**			**SVM-L**			**SVM-P3**			**SGD**			**K-means_80**		
Accuracy	100%			67%			98%			60%			90%		
	GXY	**XGY**	**XYG**	**GXY**	**XGY**	**XYG**	**GXY**	**XGY**	**XYG**	**GXY**	**XGY**	**XYG**	**GXY**	**XGY**	**XYG**
Specificity	1	1	1	0	0.79	0.92	1	0.97	1	0	0.62	0.58	0.88	1	0.8
Sensitivity	1	1	1	0	0.77	0.57	1	1	0.96	0	0.84	0.65	0.88	0.89	0.92

In summary, the generated decision trees, which are not included here due to space limitations, all indicated that the most influential attributes for design of synthetic gene circuits are those which directly control the size of ATR's as well as the position and order of the promoter in an operon such as *Rear_T_Size*, *Rear_T_dG*, *Front_T_Size*, *Front_GC_Content*, etc. This level of understanding of the model and its relationship with gene expression will facilitate future engineering of more complex synthetic gene circuits.

3.4 Conclusions and future directions

The results obtained in this work demonstrate the enormous potential of ML in improving the prediction accuracy of gene expression in complex synthetic gene circuit design compared to data-driven mathematical models. The key to the massive improvements of prediction is the ability to extract underlying connections among various attributes and the GFP fluorescence using ML methods instead of relying on traditional statistical techniques that create linear regression models. One obvious benefit of the experiments conducted in this work is that they show how biologists can investigate complex biological systems by focusing on mappings of experimental input data and the observed results. ML methods have been shown to facilitate gene expression prediction with high accuracy in polycistronic gene circuit. And that in turn offers the potential for systematic analysis of more complex gene circuits. Another interesting find in this study was that the constructed classifiers were able to determine the position of each gene and their significance in contributing to the GFP fluorescence intensity in polycistronic expression circuits. This is of exceptional benefit in design of complex gene networks to achieve specific expression levels or circuit dynamics without enormous trail-and-error experimental attempts.

In terms of future directions, it must be noted that due to the small size of the dataset, no extensive hyperparameter tunings needed to be performed simply because high accuracy models were successfully built. With more experimental data, there will be the opportunity to discover even more vital information about design of gene circuits commonly used in pharmaceutical industries, fuel production, metabolic engineering, genome engineering, and numerous biomedical applications. One of the advantages of ML methods is that the models can be

retrained and improved when new data becomes available. Biological experiments are time-consuming, so as new data becomes available and consolidated with old data, the effectiveness and robustness of the models could be extensively enhanced.

Although the data collected in this work were from prokaryotic experiments, similar methods can also be applied to other biological systems, such as eukaryotic organism, virus or even cell-free transcription-translation systems. Theoretically, using ML techniques to achieve higher prediction accuracy requires increasing number of experimental data inputs. Hence, these methods are tremendously adequate for investigating high-throughput biological analysis, including large-scale genomic sequencing or pharmaceutical drug discovery.

CHAPTER 4

PREDICTABLE CONTROL OF RNA LIFETIME USING ENGINEERED DEGRADATION-TUNING RNAS

4.1 Introduction

Precise regulation of gene expression at the level of transcription or translation plays a pivotal role in establishing basic cell function, ensuring appropriate responses to environmental cues, and even robust therapeutics and diagnostics[41,124,158–161]. Therefore, effective strategies are required to enable accurate and predictable control of the production and degradation of RNA and protein molecules[39,138,162]. In bacteria, such control has largely been achieved through engineering of the production of RNA (transcription) or protein (translation). Modulation of the -35 and -10 consensus elements has allowed for engineering of synthetic promoter libraries with a broad range of transcription efficiencies[30,163,164]. This mechanism-driven methodology has also been applied to develop tools to manipulate translation, where RNAs featuring low folding energy coupled with high affinity Shine Dalgarno (SD) sequences encourage efficient ribosome binding, thereby leading to accelerated translation rates[112,165]. Libraries of ribosome binding sites (RBSs) with varying strengths have been developed to predict and tune protein yields[112,166,167]. Other attempts have been made to control the production of gene products by developing synthetic transcriptional terminators[168–170], riboregulators[64,65,171,172], thermosensors[173], ribozymes[60], CRISPR activation and interference systems[71,174–176], switchable guide RNAs[177–179], engineering regions nearby open reading frames (ORFs)[135,180–183], and through optimization of codon usage[184,185].

RNA molecules in prokaryotes are typically unstable, with half-lives on the minute timescale, which allows cells to rapidly adapt to changes in the environment[186,187]. This rapid degradation is orchestrated by an ensemble of bacterial RNases that have been extensively studied[109,188]. In E. coli, which lacks 5’ → 3’ exonucleases, the vast majority of RNA degradation processes combine the actions of endonucleases and 3’ → 5’ exonucleases. Specifically, the RNase E endonuclease or the RNase III targets the underlying RNA molecule for primary cleavage followed by complete degradation via 3’ → 5’ exonucleases44. Previous studies have

discovered several naturally occurring 5’ UTRs, termed RNA stabilizers, or rationally designed synthetic DNA cassettes that can increase RNA half-life by forming 5’ secondary structures[110,189–191]. These 5’ hairpin structures have been shown to be able to control heterologous mRNA half-life and widely used to regulate recombinant protein expression without introducing stress to host cells49. However, most of the engineered 5’ stabilizing elements are designed and tested on an ad-hoc basis. Understanding of their functional structural features remains elusive.

Here, we explore the structural space and report a library of modular degradation-tuning RNAs (dtRNAs) that can be inserted at the 5’ end of a transcript of interest to manipulate its stability. Based on in silico analysis, these RNA modules can form secondary structures that impact RNA degradation without interfering with RBS context. We systematically characterize dtRNA structures and find that RNA stability is strongly correlated with structural features such as stem length and GC content, loop size, 5’ spacing sequence and the presence of RNase cleavage sites. Manipulation of these features yields a library of dtRNAs capable of tuning expression upwards by 5-fold or downwards by 8-fold, resulting in an overall dynamic range of 40-fold. Integrating dtRNAs with the highest stability enhancements into gene circuits enables us to tune the dynamics of a positive feedback loop and increase noncoding RNA levels for improved CRISPR interference. We further apply synthetic dtRNAs to cell-free systems and confirm their ability to increase gene expression in vitro. Lastly, we demonstrate the utility of dtRNAs by integrating them with a toehold switch sensor to implement improved paper-based viral diagnostics, illustrating the potential of dtRNAs for medical and biotechnological applications.

4.2 Results

4.2.1 Modulation of RNA stability by variants of the native *ompA* stabilizer

To verify the effectiveness of a naturally occurring RNA stabilizer, we inserted the 5’ UTR sequence from the *E. coli ompA* transcript between the promoter and RBS region to tune downstream GFP expression[110,135,189] (Figure 4.1A, right). The RNA sequence of the stabilizer forms secondary structures to stabilize the mRNA following transcription (Figure 4.1A, left). It can be seen in Figure 4.1b that the wild-type (WT) stabilizer does indeed increase GFP levels

moderately compared to a control (Ctrl) mRNA lacking the stabilizer sequence. Sequence analysis shows that the ompA stabilizer forms two hairpins (hairpin_1 and hairpin_2, blue structures in Figure 4.1A) and two single-stranded sequences between the two hairpins (ss1) and downstream of hairpin_2 (ss2) (Figure 4.2A). To investigate the contribution of these components to maintaining RNA stability, we designed and synthesized two variants of the ompA stabilizer: “Hp1” includes hairpin_1 and the first 7 nucleotides of ss1, and “Hp2” includes hairpin_2 and the first 7 nucleotides of ss2. Using a plate reader to measure GFP fluorescence after 16 hours of incubation, we first tested each cassette on a high-copy plasmid driven by a strong promoter but did not observe any significant fluorescence enhancements (Figure 4.2B). To alleviate potential saturation of the transcription and degradation process, each cassette was then inserted into the plasmid driven by a weak promoter. Interestingly, both “Hp1” and “Hp2” displayed greater GFP expression than the WT ompA sequence, with design “Hp1” providing about a 2-fold increase in GFP over the control (Figure 4.1B).

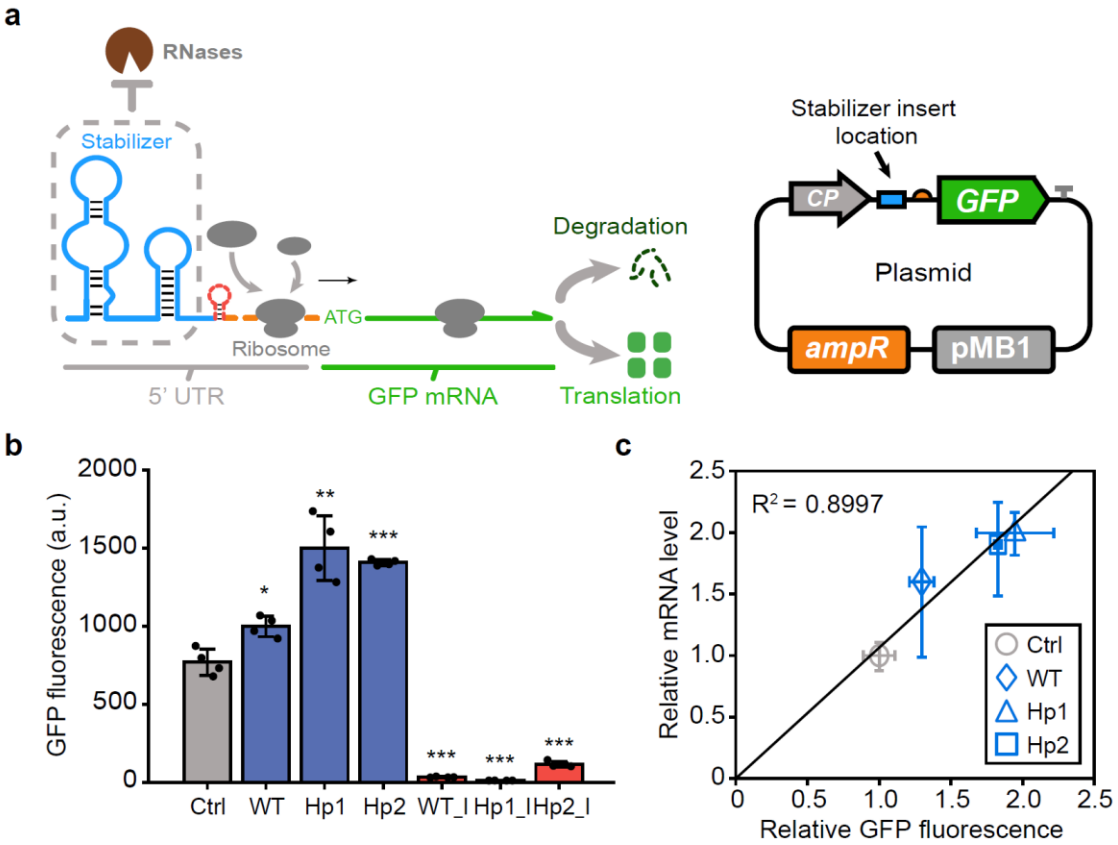


Figure 4.1 Modulation of RNA Stability by Native *ompA* Stabilizer Variants (A) Schematic showing the stabilizer protection mechanism and the plasmid constructed for fluorescence measurements. Engineered stabilizer variants are inserted between a constitutive promoter and the RBS to regulate GFP expression. Engineered stabilizer variants can form a hairpin structure (blue) to block RNase access. The structure depicted by a red dashed line indicates the small hairpin structure design nearby the RBS of WT_I, Hp1_I and Hp2_I. For the plasmid map, the gray arrow represents the constitutive promoter; the blue rectangle represents the RNA stabilizer; the orange oval represents the RBS; the green box represents GFP gene; the gray T represents the transcriptional terminator. (B) Plate reader measurements shows that GFP fluorescence is affected by engineered stabilizer variants. The designs adopt the whole (WT) or part (Hp1 and Hp2) of the native *ompA* stabilizer and exhibit GFP fluorescence enhancement (blue). Low GFP expression is observed for circuits WT_I, Hp1_I and Hp2_I with small hairpin structures nearby the RBS region (red bars). The gray bar represents the control circuit result (Ctrl). Error bars are the SD of four biological replicates. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ by student's t test. (C) Comparison between relative mRNA level and relative GFP fluorescence for circuit WT, Hp1 and Hp2. The result shows a strong correlation between these two factors ($R^2 = 0.8997$).

To explore the impact of extra secondary structures formed close to the RBS on GFP expression, another three stabilizer variants were designed and synthesized: "WT_I", "Hp1_I" and "Hp2_I" which, compared to above designs, form eight extra base pairs with their downstream sequence to establish a short hairpin structure near RBS (red structure in Figure 4.1A). These three designs showed weaker or no fluorescence (Figure 4.1B and 4.2C), demonstrating that RNA secondary structure can interfere with translation when it is too close to the RBS, as expected from previous reports[112,192].

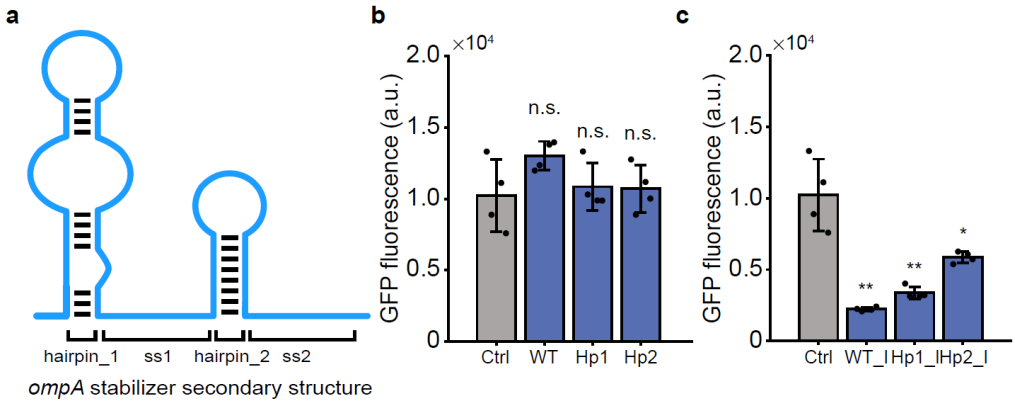


Figure 4.2 Structure of Naturally Occurring *ompA* Stabilizer and GFP Expression Is Triggered by A Strong Promoter (A) Schematic showing the structure of naturally occurring *ompA* stabilizer which comprises two hairpin structures, hairpin_1 and hairpin_2. Single-stranded nucleotide sequence one (ss1) is located between two hairpins and single-stranded nucleotide sequence two (ss2) lies downstream of hairpin_2. (B-C) GFP fluorescence measurement results for circuits transcription under a strong promoter. (B) Design WT, Hp1 and Hp2 exhibits comparable GFP fluorescence. (C) Each design with small structure formation nearby RBS region shows low GFP fluorescence levels. The data represents the mean ± SD of four biological replicates. n.s. (not significant) $p > 0.05$, * $p < 0.05$, ** $p < 0.01$ by student t test.

To rule out the possibility that the observed increase GFP fluorescence was due to enhanced translation rather than increased RNA stability, RT-qPCR experiments were carried out for Ctrl, WT, Hp1, and Hp2 to measure their RNA levels. Figure 4.1C shows that RNA level variations can explain about 90% of the change of their corresponding GFP fluorescence (R^2 = 0.8997), indicating that the observed fluorescence enhancements are attributed primarily to increased RNA levels. These results demonstrate the viability of using artificial upstream 5' UTR sequences to modulate RNA stability in our synthetic system. In addition, studying variants of naturally occurring RNA stabilizers helps distill two general principles for their effective design: use of hairpin structures and ensuring an appropriate distance between the hairpin and the RBS.

4.2.2 Identifying functional structural features of synthetic dtRNA

The general principles of dtRNA design and placement with respect to the RBS provide a foundation for identification of specific structural features that critically influence RNA stability. In silico analysis suggests stem length, stem GC content, loop size, 5' spacing sequence, and 3' insulation as primary candidate features to investigate (Figure 4.3A)

Figure 4.3B displays quantitative characterization of the impacts of stem GC content on RNA stability. Theoretically, stems with high GC content are more thermodynamically stable and could lead to stronger enhancements of RNA stability. Fifteen dtRNAs with the same secondary structure (6-nt loop and 12-bp stem) but varying stem GC content were designed and tested (Figure 4.3B). Fluorescence measurements show that structures with low GC content (less than

20%) nearly abolish the GFP expression enhancements, likely due to the unwinding of unstable AU rich hairpins removing their potential RNA-stabilizing effects. On the other hand, as the fraction of GC base pairs increases, GFP fluorescence increases concomitantly until it peaks at 66.7% GC content (8 out of 12 GC base pairs). With higher GC content, we observe diminished expression enhancement, presumably because RNA structures with GC-rich stem loop could act as transcriptional terminators, which stall RNA polymerases and cause the transcriptional complex to fall off and therefore lead to lower expression value[193,194]. This result quantifies the non-monotonic relationship between GC content and resulting RNA stability and also identifies that medium level (from 41.6% to 66.7% in our result) GC content is ideal for dtRNA structure to maximally enhance RNA stability.

To investigate the impact of stem length on RNA stability, another ten dtRNAs sharing the same loop sequence and optimal stem GC content but varying stem length were designed and tested (Figure. 4.3C). Fluorescence measurements show that structures with long stem lengths (30 bp) nearly eliminate RNA stability enhancement, possibly because even perfectly paired hairpins that are over 30 bp in length are likely to be targeted by RNase III to initiate RNA degradation process[188]. GFP fluorescence reaches its highest value for stem lengths of 12 bp. Further reductions in stem length lead to decreased hairpin stability and increased susceptibility to RNases as the stem is decreased down to 3 bp. These effects thus result in the non-monotonic relationship between stem length and the resulting RNA stability, where hairpins with 12 bp stem length show the top effect of RNA stability enhancement.

Finally, to identify the relationship between loop size and RNA stability, we designed and tested another set of twelve dtRNA structures containing optimal stem features but various loop sizes. In theory, tetraloops, which are hairpin loops of 4 nt, endow an RNA structure with strong thermal stability and make them highly nuclease resistant[195]. This effect is confirmed experimentally in Figure 4.3D where structures with loop sizes of around 4 nt (3 nt and 6 nt in our result) display the highest RNA stability enhancement. GFP fluorescence levels decrease with enlarging loop size, likely because large loops increase the possibility for RNase targeting and thereby weakening the RNA stability. Increasing loop sizes also increase the entropic cost

associated with hairpin formation, making the hairpin less thermodynamically stable. These results demonstrate a monotonically decreasing relationship between loop size and RNA stability and determines that a loop size of around 4 nt (3 nt to 6 nt) is ideal for RNA stability enhancement.

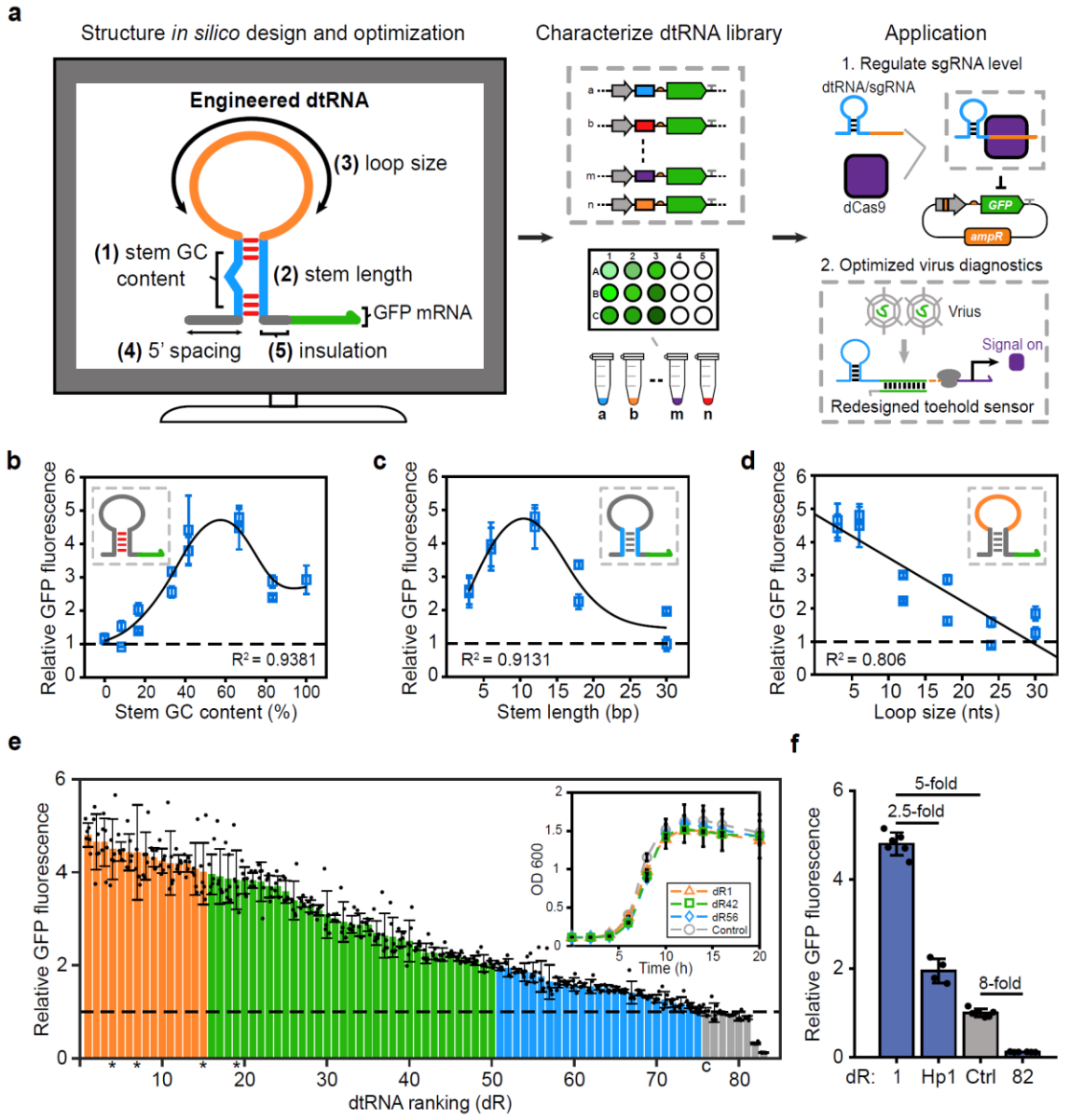


Figure 4.3 Identifying Functional Structural Features of Synthetic dtRNAs (A) Schematic showing the workflow for the present study. (B-D), Correlations between each structural feature and the relative GFP expression. For all designs, 3' insulation is achieved by insertion of ten single-stranded nucleotides downstream of the hairpin structure to minimize interference with the

downstream RBS. (B) Correlation between dtRNA stem GC content (0% to 100%) and the relative GFP fluorescence, R^2 = 0.9381; (C) Correlation between dtRNA stem length (3 bp to 30 bp) and the relative GFP fluorescence, R^2 = 0.9131; (D) Correlation between dtRNA loop size (3 nt to 30 nt) and the relative GFP fluorescence, R^2 = 0.806. The insets color-code the characterized structural features of dtRNA, and the green arrow represents GFP mRNA. The dash line represents the control fluorescence level. Error bars are the SD of six biological replicates. (E) Relative GFP fluorescence of synthetic dtRNA library. Orange bars represent designs with over 4-fold fluorescence enhancement; green bars represent designs with 2 to 4-fold enhancement; blue bars represent designs with 1-fold to 2-fold enhancement; gray bars represent designs with fluorescence lower than the control (c). Error bars are the SD of six biological replicates. Asterisks represent the dtRNAs used for in vitro measurement. Inset: Growth curve measurement results showing the OD 600 values for dR1, dR42, dR56 and control over 20 hours. Error bars are the SD of three biological replicates. (F) GFP fold difference among dtRNA structures with least and the most stable sequences, engineered stabilizer variant Hp1 (Figure 4.1B) and the control. Over 40-fold dynamic range is achieved through optimizing functional structural features of the dtRNAs.

Having designed the necessary structural features to enhance RNA stability, we next explored incorporating motifs to decrease RNA stability. We first attempted to insert the previously reported RNase E cleavage site (UCUUCC, 6-nt) into dtRNA structures[173,196]. No significant GFP fluorescence change was observed when cleavage sites were inserted into the stable hairpin (Figure 4.4A). However, GFP fluorescence was significantly reduced when introducing three cleavage sites into the relatively unstable large loop hairpin structure, demonstrating stabilizers with relatively “open” structures are easily targeted by RNases (Figure 4.4B). We next interrogated the impact of a 5’ spacing sequence on RNA stability reduction. Unlike a previous report that found that as little as a 5-nt single-stranded region at the 5’ end of the RNA could completely abolish the stabilizer function[190], we observed RNA stability enhancement for structure with 12-nt single-stranded sequence. Indeed, the stabilizing effect is completely abolished only when the 5’ single-stranded region reaches 18 nt in length (Figure 4.4C). We then combined these two features by inserting RNase E cleavage site into the 5’ spacing sequence to test if RNA stability can be further decreased. As expected, GFP fluorescence is decreased when the cleavage site is inserted 6 nt away from the hairpin structure,

and the fluorescence level is even further downregulated by about 8-fold below the control when two RNase E cleavage sites are inserted (Figure 4.4D).

We also investigated other features such as the presence of bulges within the stem and loop GC content and found that they have insignificant effects on RNA stability (Figure 4.5A and B). To investigate if dtRNAs can also be applied to genes with very different sequence composition, we select dtRNAs with varying stabilizing capabilities to regulate mRFP expression. Sequence comparison analysis shows only 3% coverage between GFP and mRFP gene suggesting mRFP reporter shares very different sequence composition with GFP. Following the same circuit construction, we insert each dtRNA to the upstream of mRFP to measure their effect on the reporter expression. Fluorescence measurement result shows that dtRNAs with top ranking in the library displayed higher relative mRFP fluorescence (Figure 4.5C). We further compare selected dtRNAs' mRFP performance to their GFP enhancement. The result also exhibits high correlation (R^2 = 0.8681), suggesting dtRNA performance is transferable to the other genes with different sequence compositions (Figure 4.5D). To further verify RNA stability enhancement is independent of genetic background, two dtRNA variants with top enhancement performance were measured with different promoters and RBSs (Figure 4.5E and F). Results from these studies confirm that dtRNAs can enhance RNA stability in a variety of genetic contexts.

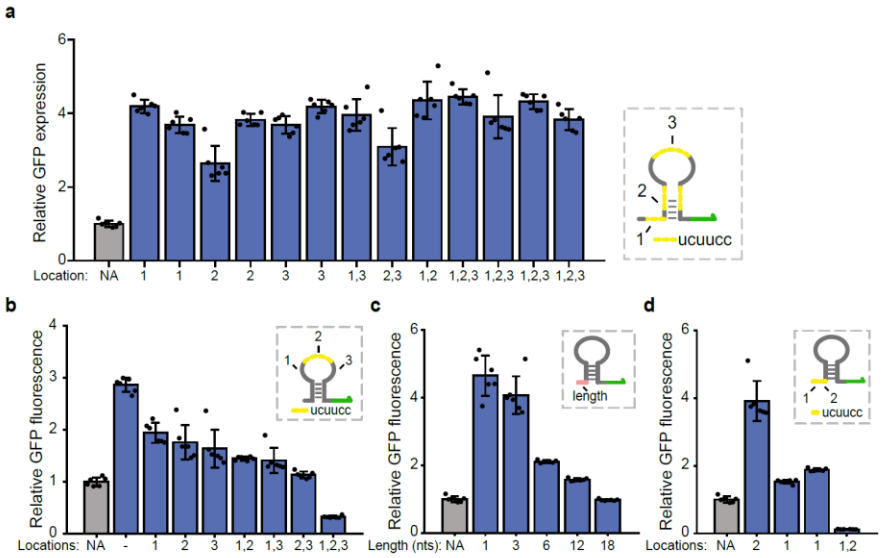


Figure 4.4 Fluorescence Measurements on dtRNAs with RNase E Cleavage Sites Engineered into Different Structural Regions. (A) Thirteen synthetic dtRNAs are designed with single or multiple RNase E cleavage sites (UCUUCC) engineered into different structural regions of dR1. The regions are marked for single or multiple RNase E cleavage sites insertion (right). Fluorescence measurement result shows that insertion of cleavage sites have insignificant effects on RNA stability. (B) Fluorescence measurement for dtRNAs with multiple RNase E cleavage sites inserting into 18-nt loop region. The inset shows the location for RNase E cleavage sites insertion. (C) Characterize the effect of dtRNA 5' spacing length on GFP expression. Five dtRNAs with 5' spacing lengths from 1-nt to 18-nt are designed to measurement their effect on GFP expression. The inset shows the location of dtRNA 5' spacing region (pink). (D) Fluorescence measurement of dtRNAs with RNase E cleavage sites engineered into 12-nt 5' spacing region. The inset shows the position of RNase E cleavage site (yellow). Error bars are the SD of six biological replicates.

To test the observed gene expression tuning can be attributed to RNA levels, RT-qPCR experiments were performed to measure RNA levels for selected dtRNAs with a range of GFP fluorescence enhancement levels. The results show a strong correlation between relative RNA level and relative GFP fluorescence (R^2 = 0.9406), suggesting GFP fluorescence variation is mainly due to the change of RNA levels (Figure 4.6A). Next, we designed additional dtRNAs with combined parameters of each feature and calculated their predicted relative GFP to investigate if dtRNA stabilizing capability can be predicted based on our feature design rules (a-i, Table 4.1). Fluorescence comparison shows a strong correlation between the predicted and observed GFP (R^2 = 0.5295, Figure 4.6B). We also increased the 5' spacing length of dtRNA with the highest predicted stabilizing capability among the new designs (f, Table 4.1). Same to our previous result that the stabilizing effect is nearly abolished with long 5' single-stranded regions (Figure 4.6C). These results demonstrate that dtRNA stabilizing ability can be roughly predicted based on each design rule.

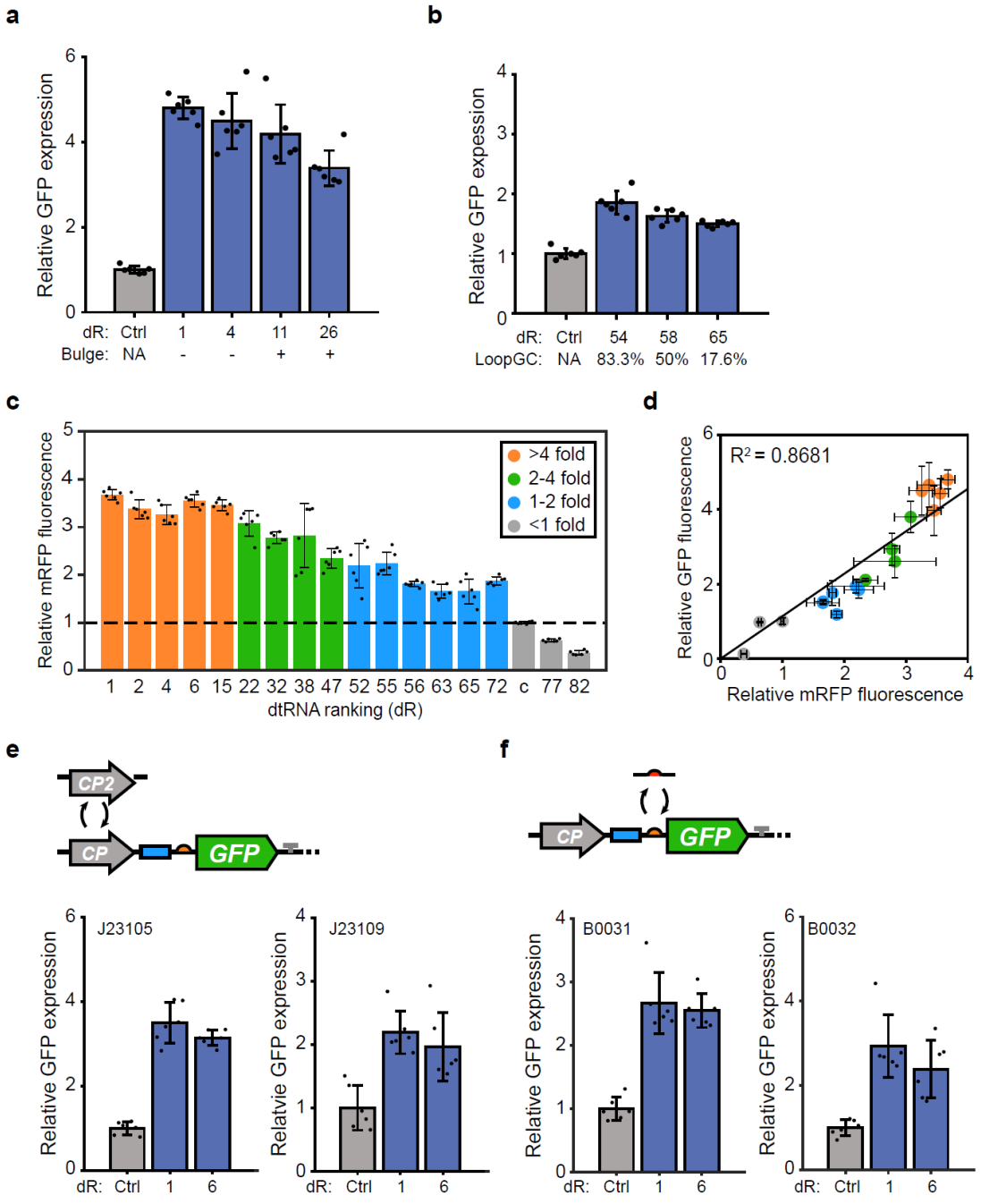


Figure 4.5 Introduction of Bulge and Loop GC Content Have Insignificant Effects on GFP Fluorescence and Commonality Test for dtRNA Regulation (A) Relative GFP expression of circuits regulated by dtRNAs with or without the bulge introduced in stem region. Three-nucleotide bulge was designed into stem region of dR1 and dR4 to be dR11 and dR26. There is no significant fluorescence difference among all designs indicating the introduction of bugle has little effect on GFP fluorescence enhancement. (B) Fluorescence measurements for designs with

the same stem feature but varying loop GC content. We maintained 18 nt loop size and designed structures with 83.3%, 50% and 17.6% loop GC content, respectively. The result indicates that loop GC content also has minor effect on GFP variations. (C) Relative mRFP fluorescence regulated by selected dtRNAs with varying stabilizing abilities. Colors of the bar represent the fold enhancement of each dtRNA on GFP reporter. (D) Comparison between relative mRFP fluorescence and relative GFP fluorescence of selected dtRNAs. The result exhibits high correlation (R^2 = 0.8681) between the report gene expression suggesting dtRNA performance is transferable to the other genes with different sequence composition. (E) Commonality test for circuits with different promoters (Top). Two promoters are selected (Biobrick number: J23105 and J23109) and engineered into the circuit with identical constructions. GFP fluorescence measurement result shows that dtRNAs are able to enhance GFP fluorescence by different promoters (Bottom). (F) Commonality test for circuits with different RBSs (Top). We further engineered circuits with different RBSs (Biobrick number: B0031 and B0032. GFP fluorescence measurement results show that synthetic dtRNAs can upregulate the GFP fluorescence with different RBSs (Bottom). Error bars of each figure are the SD of six biological replicates.

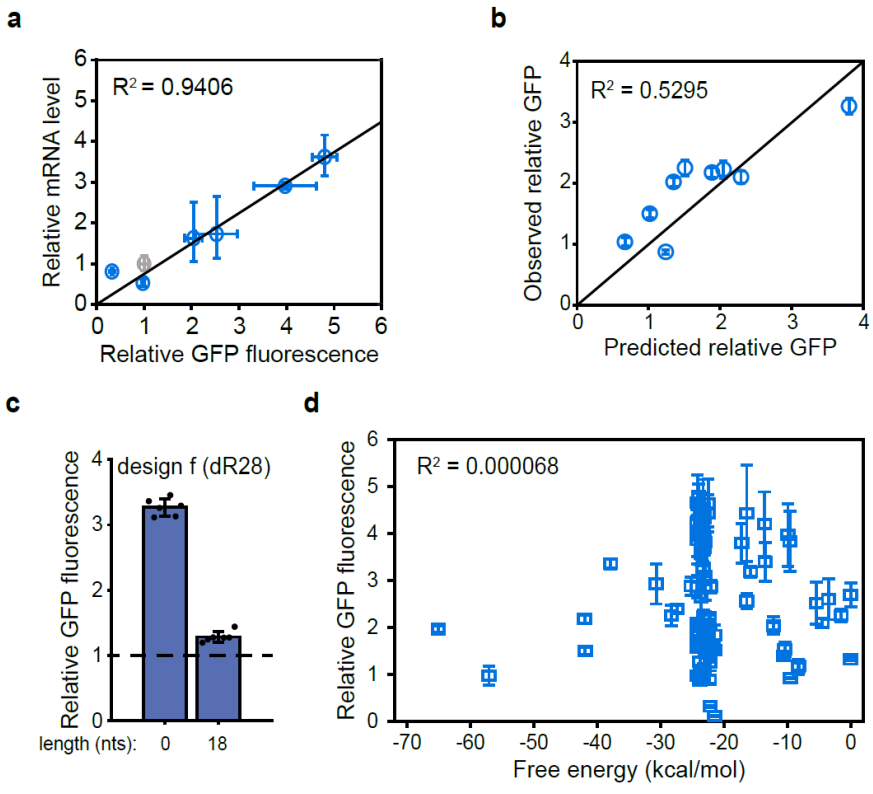


Figure 4.6 qPCR Measurements of Selected dtRNAs with Varying Stabilizing Efficiency and the Prediction of Additional Designed dtRNAs (A) qPCR measurement of relative RNA levels for

dtRNAs with diverse stabilizing efficiency. The result displays a strong correlation between relative RNA levels and relative GFP fluorescence (R^2 = 0.9406). Error bars of relative mRNA level are the SD of three biological replicates. (B) Relative fluorescence Comparison between predicted relative GFP and observed relative GFP of dtRNAs constructed followed by combined design rules (Table 4.1). The result demonstrates that dtRNA stabilizing efficiency can be roughly predicted followed our designed rules (R^2 = 0.5295). (C) Fluorescence measurement of dtRNA design f (Table 4.1) without (left) or with (right) 18 bp 5' spacing. Error bars are the SD of six biological replicates. (d) Scatter plot reveals that structure MFE is not significantly correlated with GFP fluorescence enhancement regulated by synthetic dtRNA library (R^2 = 0.000068).

In all, we systematically designed and tested a library of 82 synthetic dtRNAs and identified the functional structural features affecting RNA stability. Each dtRNA shares a single hairpin structure with an insulator sequence at the 3' end to prevent the interference between the stability hairpin and RBS region. Designed by tuning combinations of each features, dtRNAs enable quantitative control over gene expression with a wide dynamic range of 40-fold from the least to the most stable sequences (Figure 4.3E and F, dtRNA stability ranked 1 through 82, denoted dR 1-82). We also note that no significant correlation between the dtRNA minimum free energy (MFE) and GFP fluorescence was detected (Figure 4.6D), indicating that a combination of RNA sequence and structural features, rather than RNA folding alone, define transcript stability.

4.2.3 Modulation of gene circuit dynamics and noncoding RNA levels

As an initial test of the utility of dtRNAs, we selected two dtRNAs with the top GFP enhancement performance (dR1 and dR6) to incorporate into a LuxR/LuxI quorum sensing (QS) regulatory circuit to measure their impact on downstream GFP expression. We believe using top performed dtRNAs can lead to the prominent results for the redesigned system. It can be seen in Figure 4.7A that synthetic dtRNAs are only inserted in the 5' region upstream of the LuxR sequence to regulate LuxR expression (circuit C_dR1 and C_dR6). GFP fluorescence was measured to quantify the dose-response readout of each circuit. It can be seen in Figure 4.7B that as the 3OC6HSL induction increases, GFP fluorescence increases by C_dR1 and C_dR6 become more pronounced when compared against the circuit without dtRNA regulation (C_Ctrl),

suggesting synthetic dtRNAs are capable of stabilizing LuxR mRNA and thereby enhancing downstream GFP fluorescence in synthetic gene circuit (Figure 4.7B). Such stability enhancement is amplified in high induction cases because of increased transcript abundance.

Table 4.1 Information of Additional Constructed dtRNAs (design a-i)

dtRNA Design index	Stem GC content (%)	Stem length (bp)	Loop size (nt)	Predicted factor stem GC (α)	Predicted factor stem length (β)	Predicted factor loop size (γ)	Predicted relative GFP
a (dR43)	25	4	6	0.539	0.603	1	1.508
b (dR80)	25	20	6	0.539	0.496	1	1.24
c (dR50)	25	12	18	0.539	1	0.539	1.348
d (dR75)	25	20	18	0.539	0.496	0.539	0.669
e (dR48)	75	4	6	0.819	0.603	1	2.291
f (dR28)	75	12	6	0.819	1	1	3.8
g (dR46)	75	20	6	0.819	0.496	1	1.885
h (dR44)	75	12	18	0.819	1	0.539	2.048
i (dR64)	75	20	18	0.819	0.496	0.539	1.016

To explore this impact on nonlinear gene circuit dynamics, synthetic dtRNAs were inserted into a LuxR/LuxI QS-based positive feedback loop to tune the bistability of each circuit[50(p),197]. The constitutive promoter in circuits C_dR1 and C_dR6 was replaced with a pLux promoter such that LuxR gene can activate itself to form a positive feedback topology (circuit H_dR1 and H_dR6) (Figure 4.8A). Two weak dtRNAs (dR81 and dR82) were also inserted to

tune the behavior of positive feedback circuit (H_dR81 and H_dR82). We measured the robustness of history-dependent response (hysteresis), the hallmark of positive feedback topology, to determine the dynamics of each circuit[15,198]. A small bistable region is first observed for circuit H_Ctrl without dtRNA regulation (Figure 4.7C, purple lines). The bistable regions of circuit H_dR1 and H_dR6 regulated by dtRNA structures shifted to lower 3OC6HSL concentration because increased LuxR transcript stability and hence its protein abundance makes it easier for the system to switch to the ON state (Figure 4.7C and Figure 4.8B and C, green lines). We also observed enlarged bistable regions for circuits regulated by weak dtRNAs in higher drug concentration (H_dR81 and H_dR82, Figure 4.7C and Figure 4.8C, blue lines). To better explain our experimental data, we built a mathematical model for positive feedback circuit regulated by dtRNAs and performed two-parameter bifurcation analysis on the system. The result validates our data that dtRNAs with stronger stabilizing capability generate smaller bistable regions localizing in low drug concentration, while weaker dtRNAs regulation result in larger bistable region shifted to high drug concentration (Figure 4.7D). This experiment illustrates the feasibility of using synthetic dtRNAs to fine tune gene circuit dynamics.

To further explore the tunability of dtRNAs on noncoding RNA levels, we built a CRISPR interference system to control small guide RNA (sgRNA) levels by redesigning 5' sequence of sgRNA that targets a GFP promoter with dR1and dR6, and another two top-performed dtRNAs (dR15 and dR19). When transcribed from a weak promoter, each redesigned sgRNA can guide dCas9 to bind with the cognate promoter region to inhibit downstream GFP expression (Figure 4.7E). Stable sgRNAs are more likely to interact with dCas9 for GFP inhibition and fluorescence measurements confirm that GFP expression regulated by the redesigned sgRNAs is significantly lower, yielding about 22% to 36% decrease compared to the original sgRNA (sgRNA_WT) regulated GFP intensity. These results demonstrate that noncoding RNA levels can also be tuned by synthetic dtRNAs (Figure 4.7F).

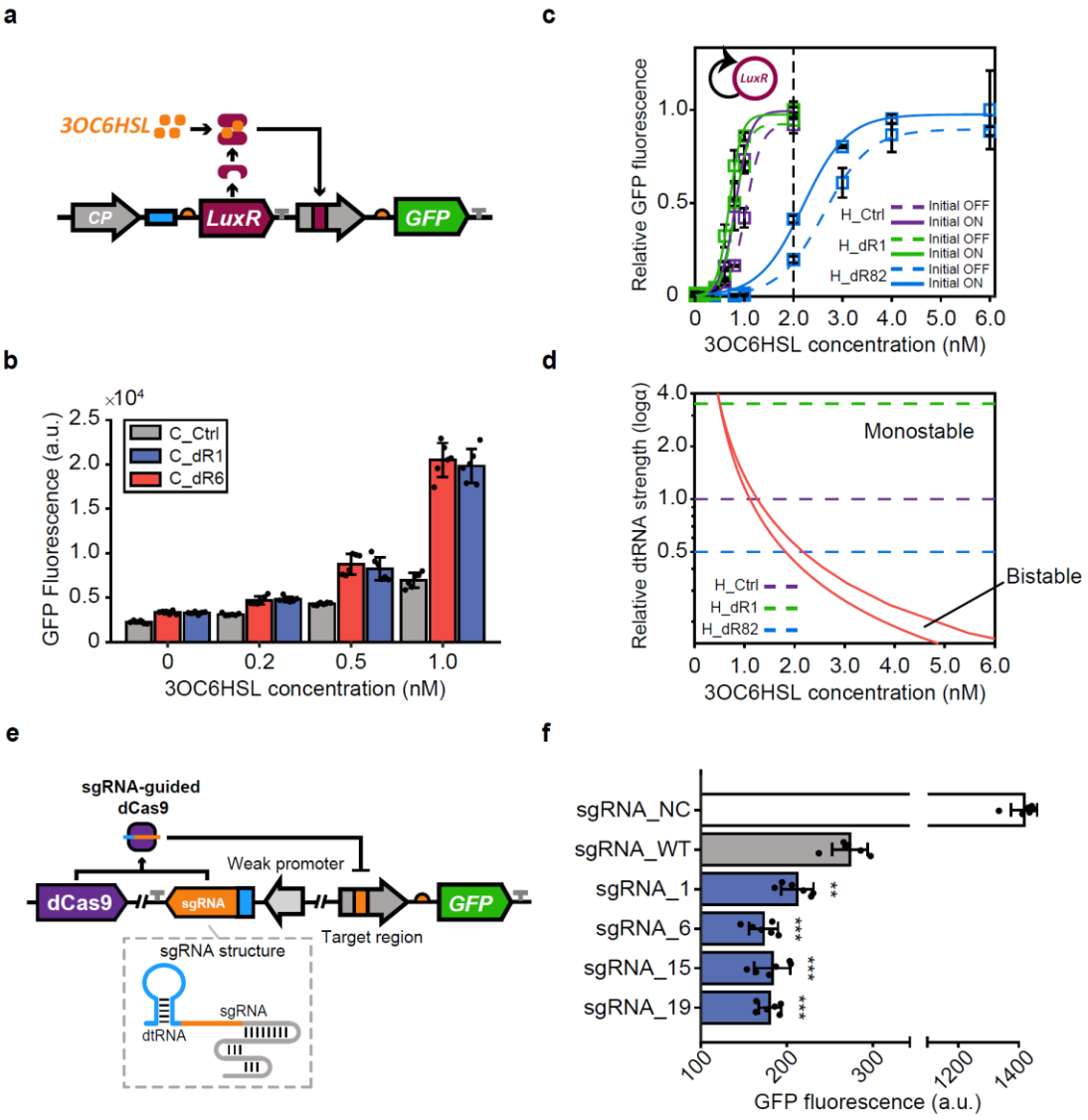


Figure 4.7 Using dtRNAs to Modulate Gene Circuit Dynamics and Noncoding RNA Levels in Synthetic Gene Circuits. (A) Schematic showing the construction of the LuxR/LuxI quorum sensing gene circuit where a constitutive promoter (gray arrow) triggers the expression of LuxR gene (purple rectangle). After being expressed, the LuxR protein dimerizes with 3OC6HSL (orange dots) and interacts with the pLux promoter to activate GFP gene expression (green rectangle). The blue rectangle represents the location of dtRNA insertion (dR1 and dR6). (B) Dose-response measurement results induced by various 3OC6HSL concentrations. Error bars are the SD of four biological replicates. (C) Hysteresis experiment results for the synthetic positive feedback loop. Various concentrations of 3OC6HSL are applied to induce each circuit. The purple lines indicate the result of initial OFF/ON experiment for the control circuit H_Ctrl; The green lines indicate the result for circuit H_dR1; The blue lines indicate the result of initial

OFF/ON experiment for circuit H_dR82. The zoomed in hysteresis result of 0 to 2 nM (dash line) 3OC6HSL concentration can be found in Figure 4.8B. The data represents the mean ± SD of three biological replicates. (D) Two-parameter bifurcation analysis result. The red lines mark the bifurcation between the monostability and bistability. The bistable region becomes smaller when shifted to lower drug concentration with the increasing of dtRNA strength. Parameter α are estimated based on our qPCR result. (E) Schematic showing CRISPRi regulation controlled by dtRNAs. Selected dtRNAs (dR1, dR6, dR15 and dR19) are integrated with sgRNA which can guide dCas9 to repress GFP expression. (F) Steady state fluorescence measurement for each CRISPRi system. All redesigned sgRNAs exhibit even lower GFP level compared to the original sgRNA (sgRNA_WT). sgRNA_NC represents the negative control result. The data represents the mean ± SD of six biological replicates. ** $p < 0.01$, *** $p < 0.001$ by student's t test.

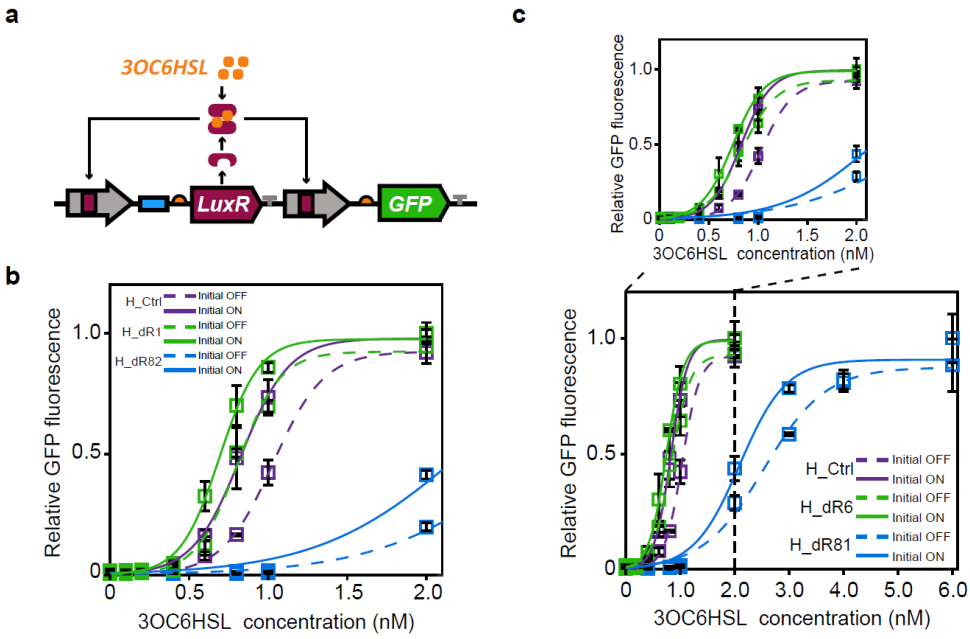


Figure 4.8 Hysteresis measurement for Engineered Positive Feedback Loop H_dR6 and H_dR82 Regulated by dtRNA (A) Schematic showing the construction of positive feedback loop, dtRNA is only inserted at 5' upstream of the LuxR gene. All genetic components are sharing the same colors as showed in Figure 4.7A. (B) The hysteresis result of Figure 4.7C regulated by dR1 and dR82 induced by 0 to 2 nM 3OC6HSL concentration. (C) Hysteresis results for synthetic positive feedback circuit regulated by dR6 and dR81. Various concentrations of 3OC6HSL are applied to induce the circuit. The purple solid and dash lines indicate the control initial on and initial off experiment results; The green solid and dash lines represent H_dR6 initial on and initial off experiment results. The blue solid and dash lines represent H_dR81 initial on and initial off experiment results. The top panel is the enlarged result induced by 0 to 2 nM 3OC6HSL concentration. The data represents the mean ± SD of three biological replicates.

4.2.4 *In vitro* regulation of gene expression by synthetic dtRNAs

Cell-free expression system is cell extract (or enzyme purified)-based tool that has been widely used in synthetic biology, metabolic engineering and in vitro diagnostics[41,83,199,200]. To test whether synthetic dtRNAs enable regulation of gene expression in cell-free expression system, we constructed four circuits with another set of top-performed dtRNAs (dR4, dR7, dR15 and dR19) to measure their impact on GFP expression in cell-free transcription-translation expression systems (Figure 4.9). For these experiments, triple guanines (GGG) were inserted at the 5' end of the dtRNAs to ensure strong transcription via T7 RNA polymerase.

We first performed measurements without the addition of RNase inhibitor to each reaction (- RNase inhibitor group). The result in Figure 4.9 (top) shows that GFP fluorescence of each circuit starts to increase shortly after the reaction begins, and it reaches a steady state after about an hour reaction (Figure 4.10A). Steady-state GFP fluorescence is much stronger for circuits regulated by synthetic dtRNAs, where dtRNA dR7 regulated circuit displays about a 10-fold fluorescence enhancement. Enhancement effects can also be detected for each reaction with RNase inhibitor treatment (Figure 4.9B, bottom and Figure 4.10B). In both cases, dtRNA significantly increased GFP fluorescence compared to control due to increased mRNA stability.

To better quantify gene expression enhancement due to RNA stability increases, we constructed a dynamic model to describe dtRNA regulated GFP expression enhancement in both scenarios (Figure 4.9B, solid lines). Since the cell free system provides abundant molecular machinery for transcription and translation, we chose to use a simplified model that includes only these two steps without nonlinear terms. We solved this simplified model analytically and fitted against experimental time course directly. Fitting results gave us a more quantitative view of the dtRNA's efficacy and are consistent with experimental observations. Using model-fitted parameters, we can calculate GFP accumulation rates over time in both scenarios, where circuits regulated by dtRNAs display much faster GFP accumulation rates compared to the control (Figure 4.9C). Theoretical derivations show that the time required for GFP accumulation rate to reach its maximum (peak of the curve) is only dependent on mRNA and protein degradation rates

Given that protein degradation rates remain constant for all scenarios, the right-shifted peaks of dtRNAs mathematically support decreased mRNA degradation rates.

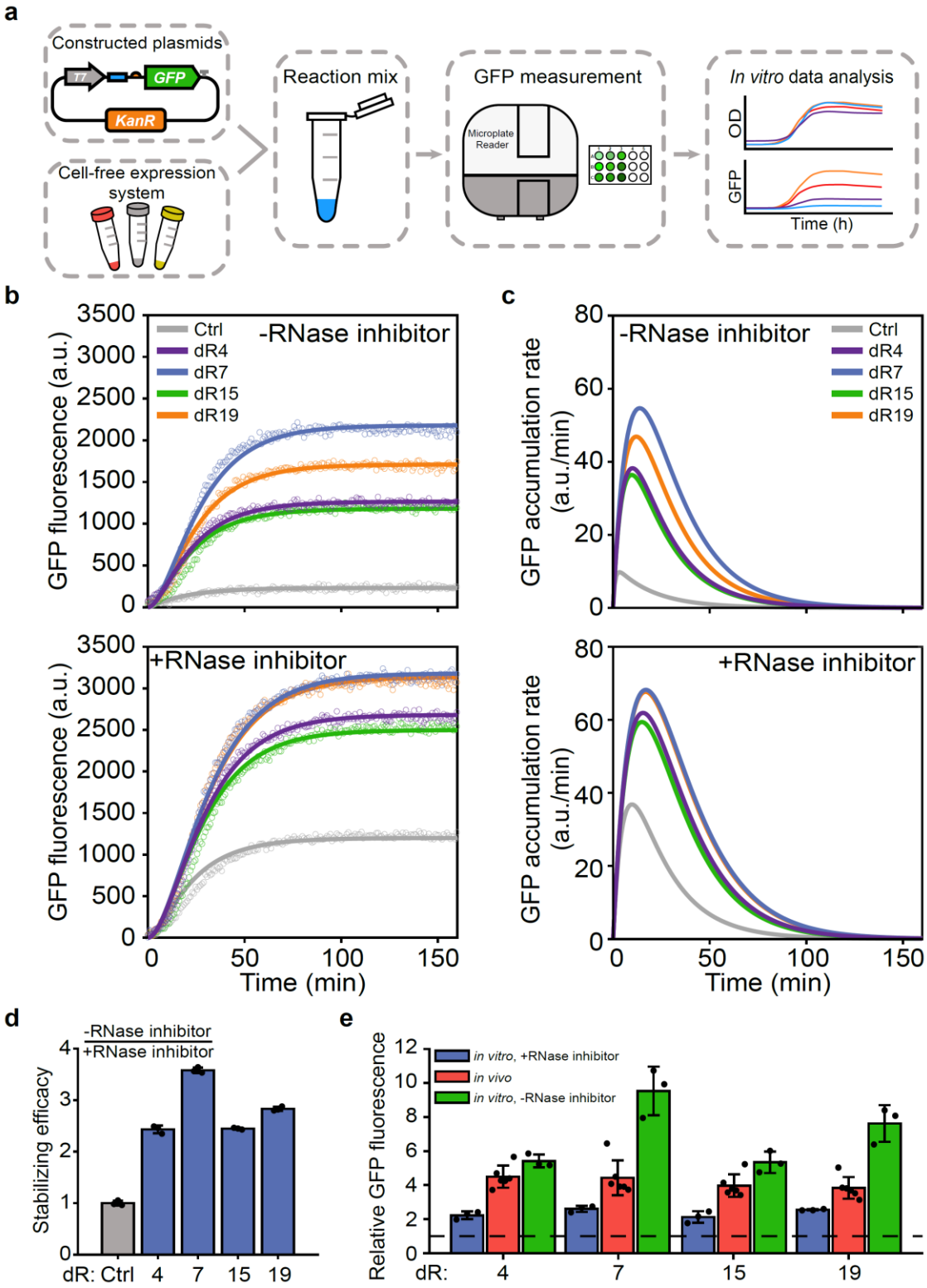


Figure 4.9 In vitro Regulation of Gene Expression via Synthetic dtRNAs (A) Schematic showing the in vitro gene expression measurements with synthetic dtRNAs (dR4, dR7, dR15 and dR19). (B) GFP expression measurement over time regulated by dtRNAs without (top)/with (bottom) RNase inhibitor treatment. Colored circles represent the observed mean GFP fluorescence of each design; solid lines represent model fitting results for each design. GFP fluorescence is measured every 50 seconds. (C) Model simulation of GFP accumulation rate regulated by dtRNAs without (top)/with (bottom) RNase inhibitor treatment. (D) Bar chart result shows the stabilizing efficacy of each dtRNA. Stabilizing efficacy is defined as the ratio between steady state GFP without RNase inhibitor and with RNase inhibitor treatment. The resultant values are further normalized against the control value. (E) Relative GFP fluorescence comparison among circuits regulated by the same dtRNAs *in vitro* and *in vivo*. Error bars are the SD of three biological replicates for in vitro measurement and six biological replicates for *in vivo* measurement.

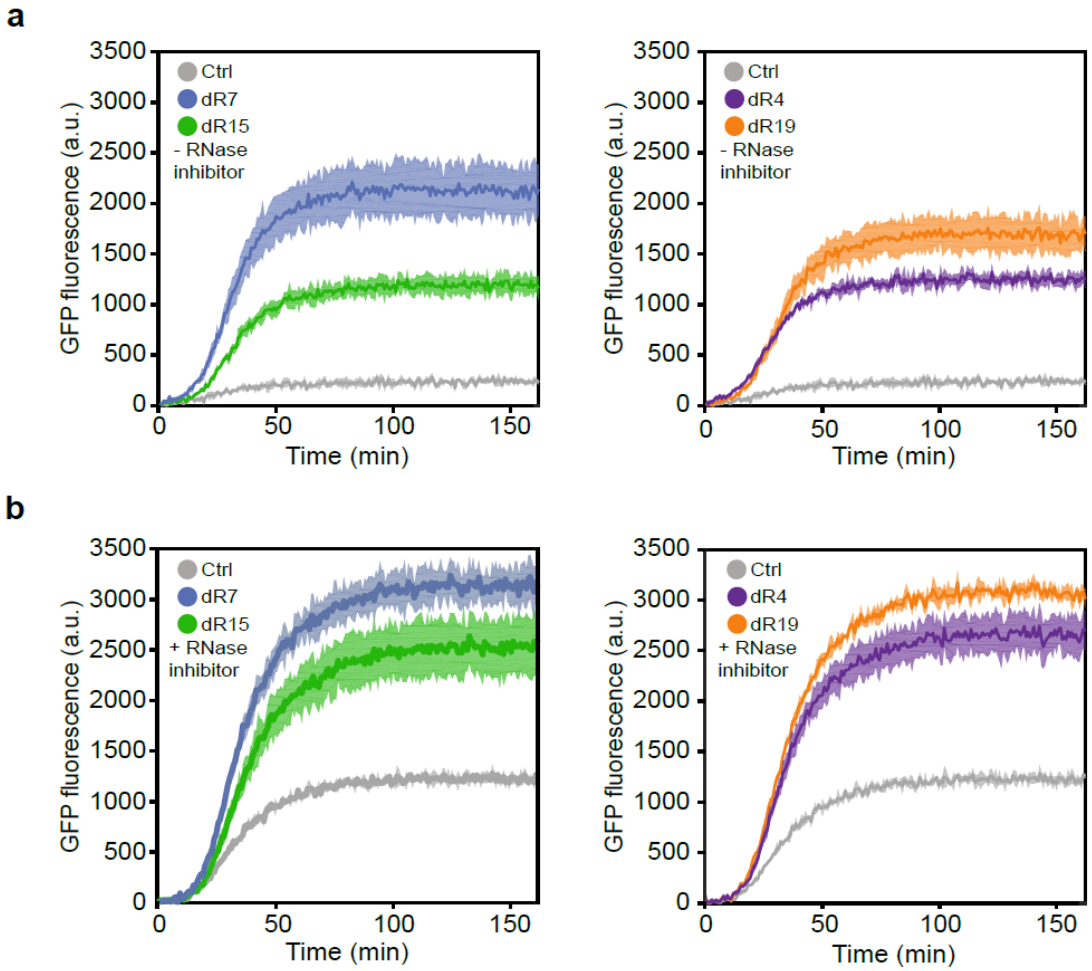


Figure 4.10 Details of *In vitro* Regulation of Gene Expression via Synthetic dtRNAs (A) GFP fluorescence measurement results of designs without RNase inhibitor treatment. (B) GFP fluorescence measurement results of designs with RNase inhibitor treatment. The gray curve

represents the mean fluorescence for circuit without dtRNA regulation (Ctrl). The purple, blue, green, and orange curves represent the mean fluorescence for circuits regulated by selected dtRNAs. The shallow area of each curve represents the SD of three biological replicates. GFP fluorescence is measured every 50 seconds.

Stabilizing efficacy, defined as the ratio between steady state GFP without RNase inhibitor and with RNase inhibitor treatment, measures robustness of dtRNAs in vitro against RNase activities, which could impact dtRNAs effectiveness (compare Figure 4.9B top and bottom). Figure 4.9D shows that all dtRNAs display over 2-fold stabilizing efficacy compared to the control. dR7 dtRNA yields the strongest enhancement at 3.6-fold, illustrating stability of dtRNAs even in the presence of RNase. Environmental dependence of dtRNA's stability enhancement potential is further quantified by comparing relative GFP intensities in live bacteria cells or in cell-free expression systems (Figure 4.9E). It can be seen that dtRNA's capability is most pronounced in complex background, i.e. *in vitro* without RNase inhibitor.

4.2.5 Improved viral diagnostics using hybrid dtRNA/toehold switch sensors

The toehold switch is a programable RNA device that can interact with a user-specified target RNA to activate translation of a protein of interest[65] and has been widely applied in areas including *in vitro* viral diagnostics[41,42], gene circuit engineering[83,171,201] and education[202]. Toehold switches feature a long single-stranded region known as a toehold at their 5' end that is designed to initiate binding with the target RNA. However, transcripts with excessive 5' single-stranded regions could be easily targeted and digested by RNases (Figure 4.4C and D). To address this limitation, we coupled toehold switches with dtRNAs to improve their performance in a diagnostic assay. These hybrid systems were constructed by inserting dtRNAs at the 5' end of an existing toehold switch designed for *in vitro* detection of norovirus in paper-based cell-free reactions (Figure 4.11A). Five hybrid systems were designed using the main structure of dtRNA with best performance in *in vitro* gene expression measurement (dR19, Figure 4.9B) with different combinations of 5' spacing and insulator sequences: dR19_1 (2-nt 5' spacing, 6-nt insulator), dR19_2 (2-nt 5' spacing, 10-nt insulator), dR19_3 (2-nt 5' spacing, 18-nt insulator), dR19_4 (6-nt

5’ spacing, 6-nt insulator) and dR19_5 (8-nt 5’ spacing, 6-nt insulator). The β-galactosidase (lacZ) **α** peptide (lacZ**α**) was used as the reporter as previously described[42]. This short peptide undergoes complementation with added β-galactosidase ω peptide during the *in vitro* translation reaction to generate an active enzyme and cleave a colorimetric reporter substrate.

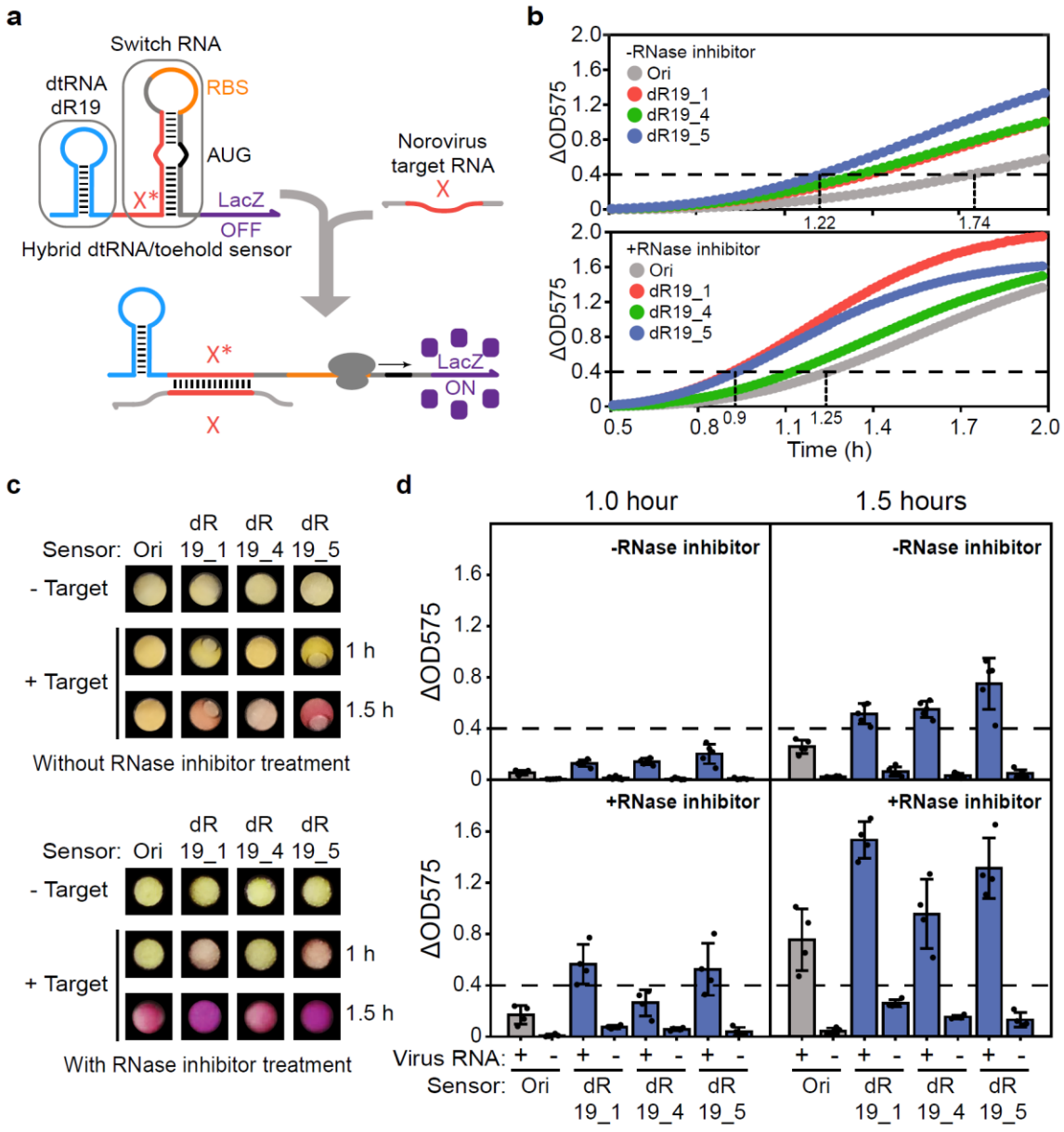


Figure 4.11 Redesigned Hybrid dtRNA/toehold Switch Sensors Improve the Performance of *in vitro* Paper-Based Viral Diagnostics (A) Schematic showing the structure of redesigned toehold switch sensors and their recognition of target RNAs. The synthetic dtRNA is integrated upstream of the sensor for stabilization. During viral RNA recognition, the target RNA with a sequence X is

recognized by the complementary X* region in the toehold switch. Binding through the single-stranded toehold region enables unwinding of the sensor hairpin to expose the RBS and start codon AUG for translation initiation. The synthetic dtRNA maintains its stable structure and protects the whole sensor transcript during the reaction. (B) Norovirus diagnostics results without (top)/ with (bottom) RNase inhibitor treatment. Each curve represents the average OD value of five reaction replicates. Blue dots represent sensor dR19_5; green dots represent sensor dR19_4; red dots represent sensor dR19_1; gray dots represent original sensor Ori. The details of each diagnostic result are shown in Figure 4.13C and D, Photographs and their corresponding diagnostic results for each sensor after 1- or 1.5-hour reactions with/without RNase inhibitor treatment, respectively. + represents the addition of synthetic norovirus RNA to the sensor. - represents the negative control. The dash line indicates the detection threshold for each device (ΔOD575 = 0.4). The data represents the mean ± SD of at least four biological replicates.

To test these hybrid sensors in paper-based diagnostic systems, synthetic norovirus RNA was introduced to paper-based devices containing cell-free reactions and DNA templates for transcription of the sensors without RNase inhibitor present. We observed that sensors with dtRNAs (dR19_1, dR19_4 and dR19_5) exhibited faster detection speed (1.22 hours, ΔOD575 = 0.4) without leaky expression, while the original sensor (Ori) without dtRNA only showed detectable signals after 1.74 hours of induction (Figure 4.11B, top and Figure 4.12A and B). Sensor dR19_2 and dR19_3 exhibited leaky expression and thus were not subjected to further experiments (Figure 4.12C). To test if the detection speed could be further improved, we proceeded to treat the paper-based device with RNase inhibitor for the second-round diagnostics. Remarkably, we found that all devices showed even faster detection speed against the group not treated with inhibitor, where signals of sensor dR19_1 and dR19_5 can be discerned within an hour (0.9 hour), indicating that the 5’ dtRNA structure can significantly improve the speed for viral diagnostics with RNase inhibitor treatment (Figure 4.11B, bottom). At the same time, however, higher expression leakage is also observed for each device, indicating the addition of RNase inhibitor, although it accelerates reaction speed, can also increase the likelihood of false positive results (Figure 4.12D). Further analysis demonstrates that non-inhibitor-treated sensor dR19_5 displays low expression leakage but faster diagnostic speed than the original sensor Ori in the presence of RNase inhibitor. Thus, hybrid sensors can exceed the performance of standard

toehold switch assays without requiring the addition of RNase inhibitor. From photographs and their corresponding diagnostic results, we confirm the improvement of viral diagnostics by using the hybrid dtRNA/toehold switch devices (Figure 4.11C and D). The details of each reaction can be found in Figure 4.13.

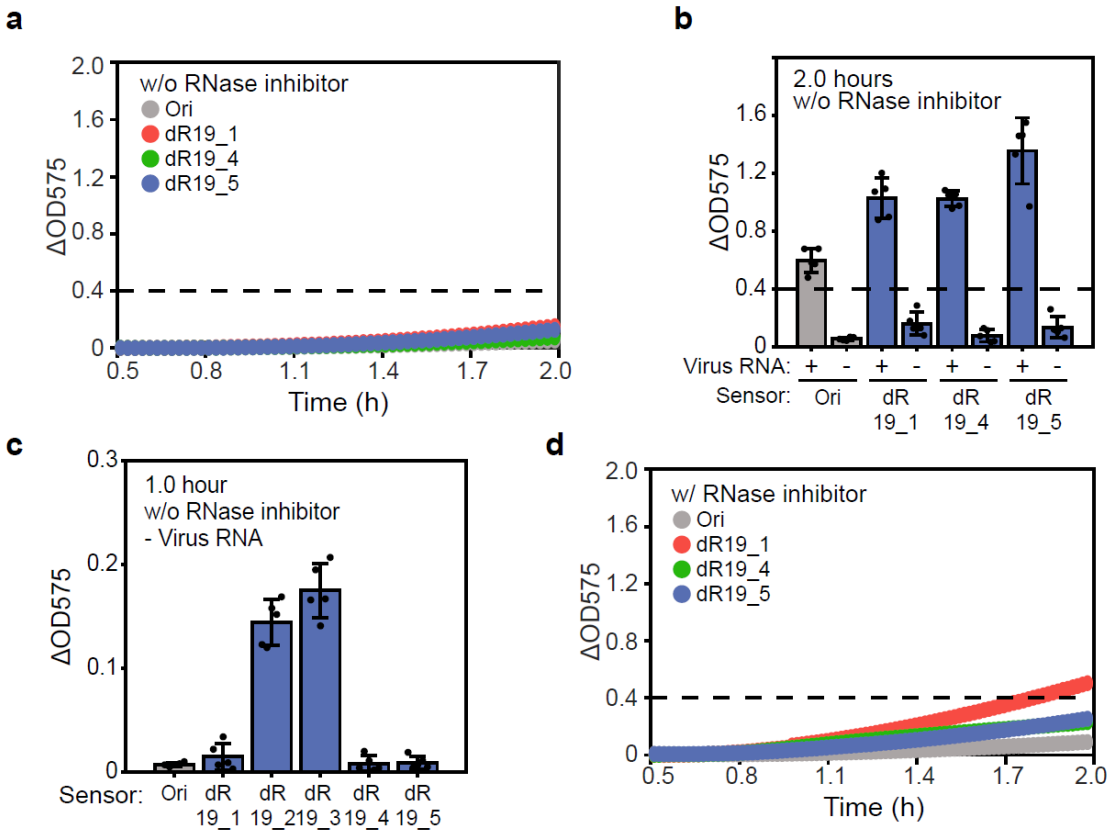


Figure 4.12 In vitro Norovirus Diagnostics 2-h Result and the Expression Leakage of Each Toehold Sensor (A) Expression leakage of sensors Ori, dR19_1 dR19_4 and dR19_5 without RNase inhibitor treatment. (B) Plate reader measurement shows 2-hour viral diagnostics result without RNase inhibitor treatment. "+" represents groups induced by synthetic norovirus RNA and "–" represents the negative control; The dash line indicates the detection threshold for each device (ΔOD575 = 0.4). The data represents the mean ± SD of five biological replicates. (C) Plate reader measurement shows device dR19_2 and dR19_3 exhibit high expression leakage. The data represents the mean ± SD of five biological replicates. (D) Expression leakage of sensors Ori, dR19_1 dR19_4 and dR19_5 with RNase inhibitor treatment.

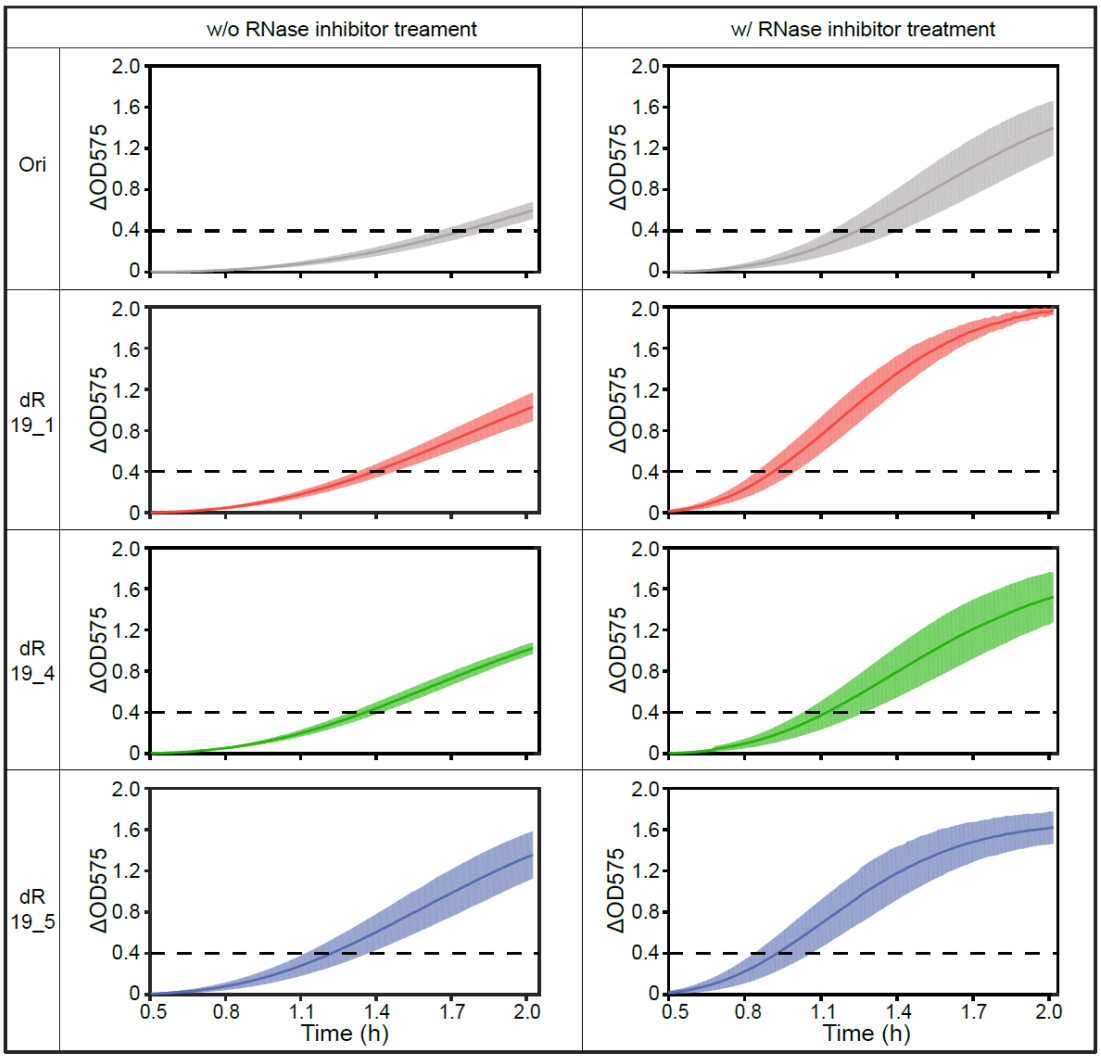


Figure 4.13 Norovirus Diagnostic Results for Sensor Ori, dR19_1, dR19_4 and dR19_5. The shadow area for each sensor represents the SD of at least four biological replicates.

4.3 Discussion

A great many methods have been developed to meet the increasing demand for control of gene expression. Naturally occurring RNA stabilizers or engineered 5' stability hairpins that thwart RNases activity hold potentials to directly control RNA half-life and have been broadly applied to regulate cellular RNA levels as well as heterologous protein yields[110,189–191]. In this study, we systematically identify the RNA structural features that influence stability, design a library of synthetic dtRNAs, and use them to tune gene expression levels in vivo and in vitro. We

demonstrate their application by using dtRNAs to increase the strength of CRISPR interference and to enhance the speed and stability of paper-based viral diagnostics.

Unlike previous studies of engineering 5' UTR sequence to manipulate translation process[112,192], our work, similar to previous synthetic 5' stabilizing elements[190,191,203], mainly focus on engineering stability hairpins to tune the RNA degradation process and hence control gene expression. Our results suggest that 5' UTR RNA secondary structure can be divided into three regions with differing effects on gene expression: the proximal module which is close to RBS, the intermediate module, and the distal module which localizes far from RBS. Secondary structures in the proximal module negatively contribute to the translation process thanks to the energy cost for unfolding the hairpin structure during translation initiation to ensure enough landing space for ribosome and RBS binding[192]. On the other hand, the distal module contributes more to the ability to block RNases from anchoring to the RNA molecules and therefore prevent the RNA molecule from being degraded. The intermediate module likely contributes to both processes. In this study, we manipulated functional features in the distal and intermediate modules to achieve a 40-fold dynamic range in tuning mRNA stability (Figure 4.3E).

In fact, mRNA degradation and translation are closely intertwined processes, only considering one to determine the final protein yield could overestimate dtRNA's capability. After being transcribed, mRNA is competitively targeted by RNases and ribosome subunits, where, in theory, a stable mRNA has higher chance for ribosome binding than unstable mRNA. Furthermore, highly translated genes can also be shielded by ribosomes that serve as the protector against RNases activities. This positive side-effect of enhanced RNA stability can be observed in our RT-qPCR results where RNA fold increase can account for over 94% but still not all GFP expression increases (Figure 4.6A). Therefore, stabilized RNAs could possess mildly higher translation rate than the unstable ones.

Our results show that a range of gene expression levels can be achieved by altering functional structural features of dtRNAs, demonstrating its potential as an alternative for precise gene expression regulation (Figure 4.3E). Compared to engineered synthetic promoter and RBS libraries, it is relatively easy to construct dtRNAs following our design rules in diverse engineering

scenarios. Besides, similar to the previous studies, our work also confirms that gene expression regulation by dtRNA modules exert little effect on cell growth, indicating RNA manipulations renders less burden for cell economy (Figure 4.3E, inset)[48,203,204].

We also successfully apply our dtRNA modules to upregulate gene expression in cell-free expression systems. An RNA-based device, the toehold switch sensor, is optimized with our dtRNAs for rapid paper-based viral diagnostics. Higher detection sensitivity with low expression leakage is achieved using the redesigned sensors, making them more compatible for potential field diagnostics. More importantly, dtRNA robustness against RNase activities suggests that they can also be used to enhance expression in crude-extract-based cell lysates, which are substantially cheaper to produce but have higher RNase levels[205,206]. Previous work has shown that native 5' UTR structures can be used to enhance gene expression in such cell-free reactions[207]. Overall, our work provides a purely RNA-based method to regulate gene expression *in vivo* and *in vitro* that can be used for a variety of different biotechnological applications.

4.4 Materials and methods

4.4.1 Strain, media and culture condition

All molecular cloning experiments were performed in *Escherichia Coli* DH10B (Invitrogen). Synthetic circuits (Figure 4.7) were tested in *E. coli* K-12 MG1655 with *lacI-/-*. Cells were grown at 37 °C in liquid and solid Luria-Bertani (LB) broth medium with 100 μg/mL ampicillin, or 50 μg/mL kanamycin, and were shaken in 5-mL or 15-mL tubes at 220 rotations per minute (rpm). Chemical 3OC6HSL was dissolved in ddH_2O and were further diluted to various working concentrations for dose-response and hysteresis measurements.

4.4.2 Plasmid construction

Most genes were obtained from iGEM Registry (http://parts.igem.org/Main_Page). Plasmids were constructed based on general molecular biology techniques and standardized Biobrick cloning methods as previously described[96]. For example, to assemble GFP gene (E0040) with a strong RBS (B0034), plasmids with GFP gene were digested with *xbaI* and *PstI* as

the cloning insert while plasmids containing RBS were digested with *SpeI* and *PstI* as the cloning vector. Digested plasmids were then separated on 1% TAE Agarose gel by gel electrophoresis. Gel bands with correct insert or vector size were selected and purified using the PureLink gel extraction Kit (Invitrogen). Gel extraction products with insert and vector were ligated by T4 DNA ligase (New England Biolabs, NEB) and transformed into E. coli DH10B. Transformed cells were plated on LB agar plates with 100 μg/mL ampicillin, or 50 μg/mL kanamycin for screening. In the end, plasmids extracted by GenElute HP MiniPrep Kit (SIGMA-ALDRICH) were confirmed through gel electrophoresis (digested by EcoRI and PstI) and Sanger DNA Sequencing (Biodesign Sequencing Core, ASU). Similar Biobrick cloning steps were taken for the following genetic components until the entire circuit has been constructed. All names and Biobrick number of genetic components can be found in Table 4.2.

For construction of the circuits with dtRNAs or sgRNAs, each structure was analyzed and designed by NUPACK design package[126] and their respective DNA oligos were synthesized by IDT. Biobrick *XbaI* and *PstI* cleavage sites were added at 5' or 3' end of the DNA oligos. DNA Oligos for the same dtRNA were diluted with ddH_2O and hetero duplexed on a heat block and were further ligated into the plasmids with the promoter digested by *XbaI* and *PstI*. The guide sequence of sgRNA or redesigned sgRNAs were designed and then synthesized by IDT. The sequence 5'-GCTA-3' and 5'-AAC-3' were added on sgRNA forward and reverse primers, respectively. DNA oligos for the same sgRNA were diluted by ddH2O, hetero duplexed on a heat block and ligated to the vector digested by *SapI* as previously described[208]. The rest of the cloning steps remain the same as the general gene circuit construction.

4.4.3 Plate reader OD and fluorescence measurements

All sequencing-confirmed gene circuits were transformed into E. coli DH10B. Single colonies were picked and cultured in 4 mL of LB medium with 100 μg/mL ampicillin. Cells were shaken until they were evenly distributed in the medium of which 300 μL were transferred into 96-

well plate for OD and fluorescence measurements. Optical density (OD600) and fluorescence (excitation: 485 nm; emission: 530 nm) were measured every 15 minutes at 37-degree under

Table 4.2 Information of iGEM Registry of Standard Biological Components and Commonly Used Genetic Parts

Biobrick number or gene name	Abbreviation in the paper	Gene description
BBa_J23104	CP	Constitutive promoter family member
BBa_J23116	CP	Constitutive promoter family member
BBa_J23105	CP	Constitutive promoter family member
BBa_J23109	CP	Constitutive promoter family member
BBa_R0062	pLux	LuxR activated promoter in concern with HSL
T7 promoter	T7	T7 polymerase specific promoter
BBa_B0034	RBS	Ribosome binding site
BBa_B0031	RBS	Ribosome binding site
BBa_B0032	RBS	Ribosome binding site
BBa_B0015	T	Transcriptional terminator used for engineering all the circuits
BBa_E0040	GFP	Green fluorescence protein used as the reporter
BBa_E1010	mRFP	Red fluorescence protein used as the reporter
BBa_C0062	LuxR	LuxR activator
BBa_pSB1A3	Vector	Plasmid backbone used for circuit cloning
BBa_pSB3k3	Vector	Plasmid backbone used for circuit cloning
pCOLODuet	Vector	Plasmid backbone used for *in vitro* experiments

continuous plate shaking (Synergy H1 Hybrid Reader, BioTek) at 220 rpm over 21 hr. For all the experiments, at least three random colonies were picked as biological replicates. For stable protein expression, we chose the 16-hour data point for further analysis in the study unless specified.

4.4.4 Flow cytometry measurements

We used Accuri C6 flow cytometer to perform the flow cytometry measurements (Becton Dickinson). Cultured samples were collected and run through the flow cytometer. For each sample, 20,000 individual cells were analyzed at the slow flow rate and the fluorescence intensity was not normalized with the cell density because it only measured single cell data. All the results were then collected in log mode and further analyzed by MATLAB (MathWorks).

4.4.5 RT-qPCR

For selected gene circuits, three biological replicates were used to quantify the mRNA levels. Total RNA was extracted from the 2 mL of cell culture using the Quick-RNA Fungal/Bacterial Miniprep Kit (Zymo Research). Purified RNA was treated in column with DNaseI (Zymo Research) to remove the extra DNA. Total RNA was eluted by nuclease-free water and the concentration quantified for the following experiments. cDNA was then synthesized from each RNA sample using iScript Reverse Transcription Supermix for RT-qPCR (Bio-Rad). For each 20-uL reaction, about 1 μg RNA was used for reverse transcription. qPCR was performed for each cDNA sample using iTaq Universal SYBR Green Supermix (Bio-Rad) and the experiment reaction was detected using the iQ5 Real-Time PCR detection system (Bio-Rad). Specifically, each cDNA sample contains an extra technical replicate, the total reaction volume for each sample is 10 μL and prokaryotic 16S rRNA was set as the endogenous control. We used previous reported primers (IDT) for both 16S rRNA and GFP amplification. The sequence of primers for 16S rRNA are 5'-GAATGCCACGGTGAATACGTT-3' (rrnB, forward, starting at the 1361st nucleotide), and 5'-CACAAAGTGGTAAGCGCCCT-3' (rrnB, reverse, starting at the 1475th nucleotide) and the sequence of GFP primers are 5'-CAGTGGAGAGGGTGAAGGTGA-3' (forward, starting at the 87th nucleotide); and 5'-CCTGTACATAACCTTCGGGCAT-3' (reverse,

starting at the 283th nucleotide). Bio-rad CFX Manager software version 3.1 was used to analyze the data. To investigate the fold change over mRNA levels, we averaged each Ct value of 16S rRNA and GFP with their biological replicates and calculated the delta Ct based on $Ct^{target} – Ct^{16S}$. Fold change for each sample was further calculated according to the biological control (circuit without dtRNA regulation) by 2-(ΔΔCt).

4.4.6 Hysteresis experiments

We used our previously reported protocol to perform the hysteresis experiments[135]. In detail, gene circuits of the synthetic positive feedback loop were constructed in a low-copy plasmid and transformed into *E. coli* K-12 MG1655 strain with lacI-/-. Single colonies for three replicates were picked for each sample and cultured at 37-degree, 220 rpm overnight in LB medium with 50 μg/mL kanamycin. For OFF-ON experiments, overnight cultured cells (initial OFF cells) were diluted into fresh LB medium at a 1:100 ratio and distributed into 5-mL polypropylene round-bottom tubes (Falcon) with various 3OC6HSL concentrations. Fluorescence of each sample was measured using an Accuri C6 flow cytometer (Becton Dickinson). In our experiments, GFP fluorescence became stable after ~12 hours of induction. For ON-OFF experiments, cells were first induced by 2 nM 3OC6HSL for 12 hours to ensure the fully induction as the initial ON state. These ON state cells were then collected through low speed centrifugation, washed once and further diluted to the fresh LB medium at 1:100 ratio. Various 3OC6HSL concentrations were then added to each sample for culture. Flow cytometry measurements were performed at 12 and 16 hours, respectively. We used 16-hour results as the ON-OFF dataset in Figure 4,7 and 4.8.

4.4.6 Hybrid dtRNA/toehold sensor plasmid construction

Synthetic DNAs encoding the redesigned norovirus-specific toehold sensors were synthesized by IDT. All cloning steps are following the general molecular biology technologies. Synthetic DNAs were amplified by PCR and inserted into the plasmid backbone using Gibson assembly[209]. Complete plasmids were further confirmed by Sanger sequencing (Biodesign Sequencing Core, ASU). Plasmids and primers were described previously[42].

4.4.7 Paper-based cell-free systems preparation

The protocols used for the paper-based cell-free reactions have been described previously[42]. Briefly, cell-free transcription-translation systems (PURExpress, NEB) were used to prepare the freeze-dried samples. The volume for each component of the reaction sample is 40% of cell-free solution A, 30% of cell-free solution B, 2% RNase inhibitor (Roche, 03335402001, distributed by MilliporeSigma) if needed, 2.5% chlorophenol red-b-D-galactopyranoside (Roche, 10884308001, distributed by MilliporeSigma, 24 mg/mL) and the remaining volume for toehold sensor DNA, lacZω and nuclease-free water. The final concentration for the synthetic DNA plasmid of each paper device is 30 ng/μL. The paper for the assays was first cut to a 2-mm diameter using a biopsy punch and transferred into PCR tubes. The prepared cell-free reaction mix (1.8 μL for each device) was then added into the PCR tubes with the paper disks and flash frozen in liquid nitrogen. Frozen devices were transferred to a lyophilizer to freeze-dry overnight. Completely dry paper devices were ready for use as viral diagnostics and can be stored at room temperature as previously described[42,83].

4.4.8 *In silico* design of synthetic dtRNA library based on NUPACK nucleic acid sequence design package

This section describes the method for in silico design of synthetic dtRNA library through NUPACK design package[126]. The same method is also used to design new dtRNAs for in vitro gene expression regulation and toehold sensor optimization for paper-based viral diagnostics.

4.4.8.1 Definition of dtRNA secondary structure domains

We first specify the secondary structure domains of dtRNA library. A single hairpin is set to be the basic structure frame for each dtRNA. As shown in Figure 4.3, factors such as the 5’ spacing, stem length and the number of GC pairs, and loop size are considered for structure optimization. Based on these features, we define the 5’ spacing region as domain “a”; the stem and loop of the hairpin frame as domains “b” and “c”, respectively; the 10 nt insulator sequence as domain “d”; and the rest of the downstream sequences are defined as domain “e”. Previous

research has demonstrated that gene expression is significantly correlated with the folding energy from the RBS region to +38 nt of the coding sequence[103,135]. Accordingly, we select 64 nt as the downstream sequence, which contains the RBS region (e.g., strong RBS BBa_B0034 in figure 4.3: AAAGAGGAGAA) and the first 38 nt of the GFP gene (Table 4.2, BBa_E0040). For T7 promoter induced gene expression (Figure 4.9 and 4.11), a GGG leader sequence is inserted at the beginning of 5' spacing (domain "a") for efficient transcription.

4.4.8.2 NUPACK scripts and dtRNA library sequence generation

After completing definition of the domains of dtRNA structure, NUPACK scripts need to be written to generate the sequence to fit the design principles. We first determine the basic settings for the design: the material is chosen to be RNA; the temperature is set at 37°C and the trial number is set as 10 which indicates the number of independent sequences to perform for one-time NUPACK design (Maximum 10).

We then define the base structure of each dtRNA in the library. In particular, we use DU+ notation to specify the single-stranded or base-paired nucleotides: U denotes the single-stranded nucleotides and D denotes the base-paired nucleotides. To define a hairpin structure with a 4 bp stem and 4 nt loop, for example, the algorithm format should be "D4 U4". Accordingly, the general format for the dtRNA structure with a 6 nt 5' spacing, 12 bp stem, 6 nt loop, 10 nt insulator sequence, and 64-nt downstream sequence is "U6 D12 U6 U10 U64". Specifically, for designs with an imperfect hairpin structure such as the introduction of a bugle within the stem region, we use brackets to specify the structural hierarchies. For example, "D3 (U3 D3 U6 U3)" denotes the structure with 9 bp stem interrupted by 3 nt symmetrical bugle. To ensure each domain will not interfere with the others, we maintain all sequences to be single stranded except the dtRNA hairpin structure during design process.

We next assign specific sequences to each domain. If the assigned sequence is not specified or needs the NUPACK design package to determine, we use the letter "N" to denote these nucleotides. Otherwise, using A, U, C and G to represent the four ribonucleotides. For example, a script with dtRNA = U6 D12 U6 U10, dtRNA.seq = a b c b* d (b* represents the

complementary sequence to b), domain a = UCUUCC, domain b = N3UCUUCCN3, domain c = UCUUCC and domain d = N10 represent a dtRNA with three RNase E cleavage sites UCUUCC inserted into 6 nt 5' spacing (domain "a"), the middle 6 bp of the stem (domain "b" and "b*"), and in the 6-nt loop (domain "c") while keeping the other nucleotides random.

For the final output of the synthetic dtRNA library, we choose Serra and Turner, 1995 as the basic RNA energy parameters and use 1.0 M Na+ and 0 M Mg2+ for the design algorithm[113]. To prevent runs of nucleotides or pairs of nucleotides, the following sequences were disallowed in the resulting designs: AAAAA, CCCCC, GGGGG, UUUUU, KKKKKK, MMMMMM, RRRRRR, SSSSSS, WWWWWW, YYYYYY.

4.4.8.3 Analysis and removal of unwanted designs

NUPACK design package calculates each design with a specific normalized ensemble defect which indicates the average percentage of incorrectly paired nucleotides at equilibrium relative to the design secondary structure which is evaluated by the Boltzmann-weighted ensemble of (unpseudoknotted) secondary structure. The best normalized ensemble defect is 0%, while 100% is the worst. We select the designs with the lowest normalized ensemble defect while removing the others to select the seed dtRNAs for each design criteria listed in Figure 4.3. These seed dtRNAs are further analyzed by NUPACK to make sure no interaction occurs between dtRNAs structure and the selected downstream sequences as shown in NUPACK structure prediction. Additionally, selected dtRNAs should keep a downstream sequence identical to their original structures. Seed dtRNAs that do not meet the specified criteria are removed from the designs. To prevent the introduction of transcriptional terminator sequences, insulator sequences with rU residues are fully removed from consideration. Based on this analysis, we chose AAAACCAAAA as the general insulator sequence for each dtRNA design unless otherwise specified.

The same method is used to denote the feature of dtRNAs to regulate gene expression in vitro and hybrid toehold sensors for viral diagnostics. In short, we select the desirable hairpin from the dtRNA library as the basal structure and define new 5' spacing and insulator sequence as the

design required (e.g., add GGG at the beginning of 5’ spacing for T7 promoter transcription preference). All designed dtRNAs are further analyzed and finalized as described above.

4.4.8.4 Examples of the scripts for dtRNA design

```
#
# Basic Settings of dtRNA structure design
material = rna
temperature = 37.0
trials = 10
sodium = 1.0
#


#
# Basic Sequence information
# Rnase E cleavage site = UCUUCC
# Common 3' end sequence (RBS to first 38 nt of GFP sequence, total 64 nt) =
TACTAGAGAAAGAGGAGAAATACTAGATGCGTAAAGGAGAAGAACTTTTCACTGGAGTTGTC
CC
# Common 3' end sequence structure = U13 D3 (U2 D3 (U1 D4 (U1 D2 (U3 D4 U8 U1)) U1) U1)
#


#
# dtRNA Structure Design
# example of dtRNA DU+ notation design of 6 nt 5' spacing, 12 bp stem 6 nt loop with 10 nt
insulator
structure dtRNA= U6 D12 U6 U10 U64
#
```

```
#
# Sequence denotation of each dtRNA domain
domain a = N6          # 5'spacing domain
domain b = N12         # dtRNA Stem region
domain c = N6          # dtRNA Loop region
domain d = N10         # Insulator sequence
domain e =
TACTAGAGAAAGAGGAGAAATACTAGATGCGTAAAGGAGAAGAACTTTTCACTGGAGTTGTC
CC     # 64 nt Common 3' end sequence
#


#
# Define each domain of dtRNA structure
dtRNA.seq = a b c b* d e
#


#
# Following sequence patterns are disallowed to prevent runs of nucleotides or pairs of
nucleotides
prevent = AAAAA, CCCCC, GGGGG, UUUUU, KKKKKK, MMMMMM, RRRRRR, SSSSSS,
WWWWWW, YYYYYY
#


#
# Output = qualified dtRNA sequence
#
```

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusions

This dissertation mainly focuses on synthetic biology research, specifically aiming to develop novel methods and tools that can benefit both research and industrial applications. In this section, I will summarize the major aims of each section and propose several possible works as the future direction.

The first chapter is mostly the introduction of the development of synthetic biology based on three divided time frames. In the early stage prior to synthetic biology emerging, scientists invented a variety of enabling techniques such as restriction enzymes, molecular cloning and PCR which serve as the foundation for using bottom-up approach to design and construct synthetic modules. As an increasing number of simple modules being developed, synthetic biologists were paying greater attention on engineering highly complex systems to achieve multi-layered and reprogrammable functions, which also indicates the transition from modules era to systems era. I also highlighted some examples of major contributions in synthetic biology in each period and discussed about the underlying challenges that could, to some extent, hamper synthetic biology development.

Chapter two describes about my first project that is to build and understand the major features that could affect gene expression in synthetic polycistronic circuits. A mathematical calculator was generated that aims to predict gene expression with designated circuit patterns without the need to perform experiments. This is actually very useful as one can select circuit design structures with desirable functions through calculation based on varying genetic features. This work could benefit gene circuit engineering without iterative design and construct process.

One of the drawbacks of this calculator is its limited prediction power caused by the lack of dataset and techniques to generate highly accurate prediction machines. Chapter three, therefore, acts as an excellent demonstration that by using machine learning (ML) based techniques, the prediction power can be significantly enhanced, from 63% to about 90% via

analyzing exact same dataset. This ability will be further advanced if more datasets were provided, indicating its potential to couple with high-throughput experiment analysis.

Chapter four is basically the extension from chapter two, where we engineered a 5' UTR module named dtRNA to control RNA stability. We also expended the usage of this tool and confirmed its ability to regulate gene expression in the other systems. The most interesting section in this project is to validate dtRNA's function to improve viral diagnostics while coupling with other RNA-based tools such as toehold switches or crRNAs. This holds the promise to redesign the RNA tools with dtRNAs to achieve novel functions.

5.2 Future works

Based on above projects, several possible future works can be proposed:

Aim 1: Characterize and model the role of 5' secondary structures in fine-tuning mRNA stability. In the previous study, we only characterize dtRNA modules in *E. coli*. Though certain principles have been drawn, some features are still not clear and thus require further investigation by engineering more dtRNAs with specific features. In order to draw a clear picture of dtRNA design principles, another direction might be to systematically analyze dtRNA integrated with fluorescence aptamers to directly visualize its functionality in cell-free system. This would help understand how dtRNA interacts with RNA molecules and could conclude detailed design principles to manipulate RNA dynamics in *in vitro* systems. To quantitatively analyze the system, models can be built for better characterization of dtRNA. Machine learning based techniques would also support model construction if enough data were provided.

Aim 2: Optimize sensing RNAs for detection of COVID-19. New viral diagnostic technologies, such as toehold switches and CRISPR-based SHERLOCK and DETECTR assays, are undergoing active development. Toehold switches rely on recognition of target viral nucleotides to unfold the switch RNA to initiate downstream reporter translation, while Cas12a/Cas13a based detection depend on collateral cleavage activated through crRNA-viral target binding. We have recently demonstrated the utility of dtRNA to improve toehold switches detection sensitivity by stabilizing them *in vitro* as well as dCas9's repression efficiency when

integrated with sgRNAs. The next step could be to investigate design strategies to quantify interactions between RNA structures formed by dtRNA and the stability of toehold switches or Cas12a/13a based crRNAs. This would help to further improve current SARS-Cov-2 diagnostics with these tools.

Aim 3: Use dtRNAs to enhance sample and amplified RNA stability for improved diagnostics. Recently we have shown that incorporating 5’ dtRNA can significantly enhance aptamer-based sensor activity as well as GFP protein expression *in vitro*. Here we propose to improve detection sensitivity by applying dtRNA strategies to enhance RNA preservation and by integrating them into isothermal amplification process, particularly NASBA to improve RNA stability while being produced. We believe the key is to quantitative understand the impacts of secondary structures on RNA’s stability and detectability.

REFERENCES

1. Endy D. Foundations for engineering biology. *Nature*. 2005;438(7067):449-453. doi:10.1038/nature04342

2. Kirschner MW. The meaning of systems biology. *Cell*. 2005;121(4):503-504. doi:10.1016/j.cell.2005.05.005

3. Sprinzak D, Elowitz MB. Reconstruction of genetic circuits. *Nature*. 2005;438(7067):443-448. doi:10.1038/nature04335

4. Hobom B. [Gene surgery: on the threshold of synthetic biology]. *Medizinische Klinik*. 1980;75(24):834-841.

5. Del Vecchio D, Dy AJ, Qian Y. Control theory meets synthetic biology. *J R Soc Interface*. 2016;13(120):20160380. doi:10.1098/rsif.2016.0380

6. Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*. 1961;3(3):318-356. doi:10.1016/S0022-2836(61)80072-7

7. Arber W, Linn S. DNA Modification and Restriction. *Annu Rev Biochem*. 1969;38(1):467-500. doi:10.1146/annurev.bi.38.070169.002343

8. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*. 1986;51 Pt 1:263-273. doi:10.1101/sqb.1986.051.01.032

9. Ideker T, Thorsson V, Ranish JA, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (New York, NY)*. 2001;292(5518):929-934. doi:10.1126/science.292.5518.929

10. Westerhoff HV, Palsson BO. The evolution of molecular biology into systems biology. *Nature Biotechnology*. 2004;22(10):1249-1252. doi:10.1038/nbt1020

11. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature*. 2000;407(6804):651-654. doi:10.1038/35036627

12. Cameron DE, Bashor CJ, Collins JJ. A brief history of synthetic biology. *Nature Reviews Microbiology*. 2014;12(5):381-390. doi:10.1038/nrmicro3239

13. McAdams HH, Arkin A. Towards a circuit engineering discipline. *Current biology: CB*. 2000;10(8):R318-320. doi:10.1016/s0960-9822(00)00440-1

14. McAdams HH, Shapiro L. Circuit simulation of genetic networks. *Science (New York, NY)*. 1995;269(5224):650-656. doi:10.1126/science.7624793

15. Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in Escherichia coli. *Nature*. 2000;403(6767):339-342. doi:10.1038/35002131

16. Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. *Nature*. 2000;403(6767):335-338. doi:10.1038/35002125

17. Chan CTY, Lee JW, Cameron DE, Bashor CJ, Collins JJ. “Deadman” and “Passcode” microbial kill switches for bacterial containment. *Nat Chem Biol*. 2016;12(2):82-86. doi:10.1038/nchembio.1979

18. Kotula JW, Kerns SJ, Shaket LA, et al. Programmable bacteria detect and record an environmental signal in the mammalian gut. *Proc Natl Acad Sci USA*. 2014;111(13):4838-4843. doi:10.1073/pnas.1321321111

19. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. Regulation of noise in the expression of a single gene. *Nat Genet*. 2002;31(1):69-73. doi:10.1038/ng869

20. Elowitz MB. Stochastic Gene Expression in a Single Cell. *Science*. 2002;297(5584):1183-1186. doi:10.1126/science.1070919

21. Blake WJ, KÆrn M, Cantor CR, Collins JJ. Noise in eukaryotic gene expression. *Nature*. 2003;422(6932):633-637. doi:10.1038/nature01546

22. Atkinson MR, Savageau MA, Myers JT, Ninfa AJ. Development of Genetic Circuitry Exhibiting Toggle Switch or Oscillatory Behavior in Escherichia coli. *Cell*. 2003;113(5):597-607. doi:10.1016/S0092-8674(03)00346-5

23. Stricker J, Cookson S, Bennett MR, Mather WH, Tsimring LS, Hasty J. A fast, robust and tunable synthetic gene oscillator. *Nature*. 2008;456(7221):516-519. doi:10.1038/nature07389

24. Tigges M, Marquez-Lago TT, Stelling J, Fussenegger M. A tunable synthetic mammalian oscillator. *Nature*. 2009;457(7227):309-312. doi:10.1038/nature07616

25. Cookson NA, Tsimring LS, Hasty J. The pedestrian watchmaker: genetic clocks from engineered oscillators. *FEBS letters*. 2009;583(24):3931-3937. doi:10.1016/j.febslet.2009.10.089

26. Weiss R, Knight TF. Engineered communications for microbial robotics. In: Condon A, Rozenberg G, eds. *DNA Computing*. Vol 2054. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2001:1-16. doi:10.1007/3-540-44992-2_1

27. Brenner K, Karig DK, Weiss R, Arnold FH. Engineered bidirectional communication mediates a consensus in a microbial biofilm consortium. *Proceedings of the National Academy of Sciences*. 2007;104(44):17300-17304. doi:10.1073/pnas.0704256104

28. Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R. A synthetic multicellular system for programmed pattern formation. *Nature*. 2005;434(7037):1130-1134. doi:10.1038/nature03461

29. Friedland AE, Lu TK, Wang X, Shi D, Church G, Collins JJ. Synthetic Gene Networks That Count. *Science*. 2009;324(5931):1199-1202. doi:10.1126/science.1172005

30. Ellis T, Wang X, Collins JJ. Diversity-based, model-guided construction of synthetic gene networks with predicted functions. *Nature Biotechnology*. 2009;27(5):465-471. doi:10.1038/nbt.1536

31. Sohka T, Heins RA, Phelan RM, Greisler JM, Townsend CA, Ostermeier M. An externally tunable bacterial band-pass filter. *Proceedings of the National Academy of Sciences*. 2009;106(25):10135-10140. doi:10.1073/pnas.0901246106

32. Basu S, Mehreja R, Thiberge S, Chen M-T, Weiss R. Spatiotemporal control of gene expression with pulse-generating networks. *Proceedings of the National Academy of Sciences*. 2004;101(17):6355-6360. doi:10.1073/pnas.0307571101

33. Levskaya A, Chevalier AA, Tabor JJ, et al. Engineering Escherichia coli to see light. *Nature*. 2005;438(7067):441-442. doi:10.1038/nature04405

34. Tabor JJ, Salis HM, Simpson ZB, et al. A synthetic genetic edge detection program. *Cell*. 2009;137(7):1272-1281. doi:10.1016/j.cell.2009.04.048

35. Wang B, Kitney RI, Joly N, Buck M. Engineering modular and orthogonal genetic logic gates for robust digital-like synthetic biology. *Nature Communications*. 2011;2:508. doi:10.1038/ncomms1516

36. Siuti P, Yazbek J, Lu TK. Synthetic circuits integrating logic and memory in living cells. *Nat Biotechnol*. 2013;31(5):448-452. doi:10.1038/nbt.2510

37. Bonnet J, Yin P, Ortiz ME, Subsoontorn P, Endy D. Amplifying Genetic Logic Gates. *Science*. 2013;340(6132):599-603. doi:10.1126/science.1232758

38. Purnick PEM, Weiss R. The second wave of synthetic biology: from modules to systems. *Nat Rev Mol Cell Biol*. 2009;10(6):410-422. doi:10.1038/nrm2698

39. Khalil AS, Collins JJ. Synthetic biology: applications come of age. *Nat Rev Genet*. 2010;11(5):367-379. doi:10.1038/nrg2775

40. Wang B, Barahona M, Buck M. A modular cell-based biosensor using engineered genetic logic circuits to detect and integrate multiple environmental signals. *Biosensors & Bioelectronics*. 2013;40(1):368-376. doi:10.1016/j.bios.2012.08.011

41. Pardee K, Green AA, Takahashi MK, et al. Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components. *Cell*. 2016;165(5):1255-1266. doi:10.1016/j.cell.2016.04.059

42. Ma D, Shen L, Wu K, Diehnelt CW, Green AA. Low-cost detection of norovirus using paper-based cell-free systems and synbody-based viral enrichment. *Synthetic Biology*. 2018;3(1):ysy018. doi:10.1093/synbio/ysy018

43. Ruder WC, Lu T, Collins JJ. Synthetic biology moving into the clinic. *Science (New York, NY)*. 2011;333(6047):1248-1252. doi:10.1126/science.1206843

44. Nissim L, Bar-Ziv RH. A tunable dual-promoter integrator for targeting of cancer cells. *Mol Syst Biol*. 2010;6(1):444. doi:10.1038/msb.2010.99

45. Xie Z, Wroblewska L, Prochazka L, Weiss R, Benenson Y. Multi-Input RNAi-Based Logic Circuit for Identification of Specific Cancer Cells. *Science*. 2011;333(6047):1307-1311. doi:10.1126/science.1205527

46. Lienert F, Lohmueller JJ, Garg A, Silver PA. Synthetic biology in mammalian cells: next generation research tools and therapeutics. *Nat Rev Mol Cell Biol*. 2014;15(2):95-107. doi:10.1038/nrm3738

47. Zhang R, Li J, Melendez-Alvarez J, et al. Topology-dependent interference of synthetic gene circuit function by growth feedback. *Nat Chem Biol*. 2020;16(6):695-701. doi:10.1038/s41589-020-0509-x

48. Ceroni F, Boo A, Furini S, et al. Burden-driven feedback control of gene expression. *Nature Methods*. 2018;15(5):387-393. doi:10.1038/nmeth.4635

49. Ceroni F, Algar R, Stan G-B, Ellis T. Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nature Methods*. 2015;12(5):415-418. doi:10.1038/nmeth.3339

50. Wu F, Menn DJ, Wang X. Quorum-Sensing Crosstalk-Driven Synthetic Circuits: From Unimodality to Trimodality. *Chemistry & Biology*. 2014;21(12):1629-1638. doi:10.1016/j.chembiol.2014.10.008

51. Callura JM, Cantor CR, Collins JJ. Genetic switchboard for synthetic biology applications. *Proceedings of the National Academy of Sciences*. 2012;109(15):5850-5855. doi:10.1073/pnas.1203808109

52. Callura JM, Dwyer DJ, Isaacs FJ, Cantor CR, Collins JJ. Tracking, tuning, and terminating microbial physiology using synthetic riboregulators. *Proceedings of the National Academy of Sciences*. 2010;107(36):15898-15903. doi:10.1073/pnas.1009747107

53. Lucks JB, Qi L, Mutalik VK, Wang D, Arkin AP. Versatile RNA-sensing transcriptional regulators for engineering genetic networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2011;108(21):8617-8622. doi:10.1073/pnas.1015741108

54. Mutalik VK, Qi L, Guimaraes JC, Lucks JB, Arkin AP. Rationally designed families of orthogonal RNA regulators of translation. *Nat Chem Biol*. 2012;8(5):447-454. doi:10.1038/nchembio.919

55. Mandal M. A Glycine-Dependent Riboswitch That Uses Cooperative Binding to Control Gene Expression. *Science*. 2004;306(5694):275-279. doi:10.1126/science.1100829

56. Ames TD, Breaker RR. Bacterial aptamers that selectively bind glutamine. *RNA Biology*. 2011;8(1):82-89. doi:10.4161/rna.8.1.13864

57. Sudarsan N, Wickiser JK, Nakamura S, Ebert MS, Breaker RR. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes & Development*. 2003;17(21):2688-2697. doi:10.1101/gad.1140003

58. Ren A, Rajashankar KR, Patel DJ. Fluoride ion encapsulation by Mg2+ ions and phosphates in a fluoride riboswitch. *Nature*. 2012;486(7401):85-89. doi:10.1038/nature11152

59. Montange RK, Batey RT. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. *Nature*. 2006;441(7097):1172-1175. doi:10.1038/nature04819

60. Lou C, Stanton B, Chen Y-J, Munsky B, Voigt CA. Ribozyme-based insulator parts buffer synthetic circuits from genetic context. *Nature Biotechnology*. 2012;30(11):1137-1142. doi:10.1038/nbt.2401

61. Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature*. 2004;428(6980):281-286. doi:10.1038/nature02362

62. Kruger K, Grabowski PJ, Zaug AJ, Sands J, Gottschling DE, Cech TR. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell*. 1982;31(1):147-157. doi:10.1016/0092-8674(82)90414-7

63. Hammann C, Luptak A, Perreault J, de la Peña M. The ubiquitous hammerhead ribozyme. *RNA (New York, NY)*. 2012;18(5):871-885. doi:10.1261/rna.031401.111

64. Chappell J, Takahashi MK, Lucks JB. Creating small transcription activating RNAs. *Nat Chem Biol*. 2015;11(3):214-220. doi:10.1038/nchembio.1737

65. Green AA, Silver PA, Collins JJ, Yin P. Toehold Switches: De-Novo-Designed Regulators of Gene Expression. *Cell*. 2014;159(4):925-939. doi:10.1016/j.cell.2014.10.002

66. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*. 2012;337(6096):816-821. doi:10.1126/science.1225829

67. Cong L, Ran FA, Cox D, et al. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*. 2013;339(6121):819-823. doi:10.1126/science.1231143

68. Cox DBT, Platt RJ, Zhang F. Therapeutic genome editing: prospects and challenges. *Nature Medicine*. 2015;21(2):121-131. doi:10.1038/nm.3793

69. Knott GJ, Doudna JA. CRISPR-Cas guides the future of genetic engineering. *Science (New York, NY)*. 2018;361(6405):866-869. doi:10.1126/science.aat5011

70. Wang H-X, Li M, Lee CM, et al. CRISPR/Cas9-Based Genome Editing for Disease Modeling and Therapy: Challenges and Opportunities for Nonviral Delivery. *Chem Rev*. 2017;117(15):9874-9906. doi:10.1021/acs.chemrev.6b00799

71. Qi LS, Larson MH, Gilbert LA, et al. Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell*. 2013;152(5):1173-1183. doi:10.1016/j.cell.2013.02.022

72. Perez-Pinera P, Kocak DD, Vockley CM, et al. RNA-guided gene activation by CRISPR-Cas9–based transcription factors. *Nat Methods*. 2013;10(10):973-976. doi:10.1038/nmeth.2600

73. Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH, Joung JK. CRISPR RNA–guided activation of endogenous human genes. *Nat Methods*. 2013;10(10):977-979. doi:10.1038/nmeth.2598

74. Chavez A, Tuttle M, Pruitt BW, et al. Comparison of Cas9 activators in multiple species. *Nat Methods*. 2016;13(7):563-567. doi:10.1038/nmeth.3871

75. Nihongaki Y, Yamamoto S, Kawano F, Suzuki H, Sato M. CRISPR-Cas9-based Photoactivatable Transcription System. *Chemistry & Biology*. 2015;22(2):169-174. doi:10.1016/j.chembiol.2014.12.011

76. Jusiak B, Cleto S, Perez-Piñera P, Lu TK. Engineering Synthetic Gene Circuits in Living Cells with CRISPR Technology. *Trends Biotechnol*. 2016;34(7):535-547. doi:10.1016/j.tibtech.2015.12.014

77. Liu Y, Zeng Y, Liu L, et al. Synthesizing AND gate genetic circuits based on CRISPR-Cas9 for identification of bladder cancer cells. *Nat Commun*. 2014;5(1):5393. doi:10.1038/ncomms6393

78. Nielsen AAK, Voigt CA. Multi-input CRISPR/Cas genetic circuits that interface host regulatory networks. *Mol Syst Biol*. 2014;10:763. doi:10.15252/msb.20145735

79. Cao Y, Ryser MD, Payne S, Li B, Rao CV, You L. Collective Space-Sensing Coordinates Pattern Scaling in Engineered Bacteria. *Cell*. 2016;165(3):620-630. doi:10.1016/j.cell.2016.03.006

80. Liu C, Fu X, Liu L, et al. Sequential establishment of stripe patterns in an expanding cell population. *Science*. 2011;334(6053):238-241. doi:10.1126/science.1209042

81. Smanski MJ, Zhou H, Claesen J, Shen B, Fischbach MA, Voigt CA. Synthetic biology to access and expand nature's chemical diversity. *Nat Rev Microbiol*. 2016;14(3):135-149. doi:10.1038/nrmicro.2015.24

82. Wright G. Perspective: Synthetic biology revives antibiotics. *Nature*. 2014;509(7498):S13. doi:10.1038/509S13a

83. Pardee K, Green AA, Ferrante T, et al. Paper-based synthetic gene networks. *Cell*. 2014;159(4):940-954. doi:10.1016/j.cell.2014.10.004

84. Din MO, Danino T, Prindle A, et al. Synchronized cycles of bacterial lysis for in vivo delivery. *Nature*. 2016;536(7614):81-85. doi:10.1038/nature18930

85. Mus F, Crook MB, Garcia K, et al. Symbiotic Nitrogen Fixation and the Challenges to Its Extension to Nonlegumes. *Appl Environ Microbiol*. 2016;82(13):3698-3710. doi:10.1128/AEM.01055-16

86. Wu F, Wang X. Applications of Synthetic Gene Networks. *Science Progress*. 2015;98(3):244-252. doi:10.3184/003685015X14368807556441

87. Zhang W, Nielsen DR. Synthetic biology applications in industrial microbiology. *Front Microbiol*. 2014;5:451. doi:10.3389/fmicb.2014.00451

88. Rocha EPC. The Organization of the Bacterial Genome. *Annu Rev Genet*. 2008;42(1):211-233. doi:10.1146/annurev.genet.42.110807.091653

89. Ma KC, Perli SD, Lu TK. Foundations and Emerging Paradigms for Computing in Living Cells. *Journal of Molecular Biology*. 2016;428(5):893-915. doi:10.1016/j.jmb.2016.02.018

90. Lee JW, Gyorgy A, Cameron DE, et al. Creating Single-Copy Genetic Circuits. *Molecular Cell*. 2016;63(2):329-336. doi:10.1016/j.molcel.2016.06.006

91. 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*. 2014;10(6):731. doi:10.15252/msb.20134955

92. Cameron DE, Collins JJ. Tunable protein degradation in bacteria. *Nat Biotechnol*. 2014;32(12):1276-1281. doi:10.1038/nbt.3053

93. Prindle A, Selimkhanov J, Li H, Razinkov I, Tsimring LS, Hasty J. Rapid and tunable post-translational coupling of genetic circuits. *Nature*. 2014;508(7496):387-391. doi:10.1038/nature13238

94. Yang L, Nielsen AAK, Fernandez-Rodriguez J, et al. Permanent genetic memory with >1-byte capacity. *Nat Methods*. 2014;11(12):1261-1266. doi:10.1038/nmeth.3147

95. Litcofsky KD, Afeyan RB, Krom RJ, Khalil AS, Collins JJ. Iterative plug-and-play methodology for constructing and modifying synthetic gene networks. *Nat Methods*. 2012;9(11):1077-1080. doi:10.1038/nmeth.2205

96. Wu F, Su R-Q, Lai Y-C, Wang X. Engineering of a synthetic quadrastable gene network to approach Waddington landscape and cell fate determination. *Elife*. 2017;6. doi:10.7554/eLife.23702

97. Chizzolini F, Forlin M, Cecchi D, Mansy SS. Gene Position More Strongly Influences Cell-Free Protein Expression from Operons than T7 Transcriptional Promoter Strength. *ACS Synth Biol*. 2014;3(6):363-371. doi:10.1021/sb4000977

98. Lim HN, Lee Y, Hussein R. Fundamental relationship between operon organization and gene expression. *Proceedings of the National Academy of Sciences*. 2011;108(26):10626-10631. doi:10.1073/pnas.1105692108

99. Brophy JAN, Voigt CA. Antisense transcription as a tool to tune gene expression. *Mol Syst Biol*. 2016;12(1):854. doi:10.15252/msb.20156540

100. Carr SB, Beal J, Densmore DM. Reducing DNA context dependence in bacterial promoters. *PLoS ONE*. 2017;12(4):e0176013. doi:10.1371/journal.pone.0176013

101. Yeung E, Dy AJ, Martin KB, et al. Biophysical Constraints Arising from Compositional Context in Synthetic Gene Networks. *Cell Syst*. 2017;5(1):11-24.e12. doi:10.1016/j.cels.2017.06.001

102. Taniguchi Y, Choi PJ, Li G-W, et al. Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science*. 2010;329(5991):533-538. doi:10.1126/science.1188308

103. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in Escherichia coli. *Science*. 2009;324(5924):255-258. doi:10.1126/science.1170160

104. Mao Y, Liu H, Liu Y, Tao S. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae. *Nucleic Acids Research*. 2014;42(8):4813-4822. doi:10.1093/nar/gku159

105. Tuller T, Waldman YY, Kupiec M, Ruppin E. Translation efficiency is determined by both codon bias and folding energy. *Proc Natl Acad Sci USA*. 2010;107(8):3645-3650. doi:10.1073/pnas.0909910107

106. Ahnert SE, Fink TMA, Zinovyev A. How much non-coding DNA do eukaryotes require? *Journal of Theoretical Biology*. 2008;252(4):587-592. doi:10.1016/j.jtbi.2008.02.005

107. Oliva G, Sahr T, Buchrieser C. Small RNAs, 5′ UTR elements and RNA-binding proteins in intracellular bacteria: impact on metabolism and virulence. *FEMS Microbiology Reviews*. 2015;39(3):331-349. doi:10.1093/femsre/fuv022

108. Chen H, Shiroguchi K, Ge H, Xie XS. Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Mol Syst Biol*. 2015;11(1):781. doi:10.15252/msb.20145794

109. Mackie GA. RNase E: at the interface of bacterial RNA processing and decay. *Nat Rev Microbiol*. 2013;11(1):45-57. doi:10.1038/nrmicro2930

110. Emory SA, Bouvet P, Belasco JG. A 5'-terminal stem-loop structure can stabilize mRNA in Escherichia coli. *Genes & Development*. 1992;6(1):135-148. doi:10.1101/gad.6.1.135

111. Selinger DW. Global RNA Half-Life Analysis in Escherichia coli Reveals Positional Patterns of Transcript Degradation. *Genome Research*. 2003;13(2):216-223. doi:10.1101/gr.912603

112. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. *Nat Biotechnol*. 2009;27(10):946-950. doi:10.1038/nbt.1568

113. Serra MJ, Turner DH. Predicting thermodynamic properties of RNA. *Meth Enzymol*. 1995;259:242-261. doi:10.1016/0076-6879(95)59047-1

114. de Smit MH, van Duin J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis. *Proceedings of the National Academy of Sciences*. 1990;87(19):7668-7672. doi:10.1073/pnas.87.19.7668

115. Xia T, SantaLucia J, Burkard ME, et al. Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson−Crick Base Pairs †. *Biochemistry*. 1998;37(42):14719-14735. doi:10.1021/bi9809425

116. Seffens W, Digby D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res*. 1999;27(7):1578-1584. doi:10.1093/nar/27.7.1578

117. Trotta E. On the normalization of the minimum free energy of RNAs by sequence length. *PLoS One*. 2014;9(11):e113380. doi:10.1371/journal.pone.0113380

118. Egbert RG, Klavins E. Fine-tuning gene networks using simple sequence repeats. *Proceedings of the National Academy of Sciences*. 2012;109(42):16817-16822. doi:10.1073/pnas.1205693109

119. Bennett MR, Hasty J. Overpowering the component problem. *Nat Biotechnol*. 2009;27(5):450-451. doi:10.1038/nbt0509-450

120. Brophy JAN, Voigt CA. Principles of genetic circuit design. *Nat Methods*. 2014;11(5):508-520. doi:10.1038/nmeth.2926

121. Nielsen AAK, Der BS, Shin J, et al. Genetic circuit design automation. *Science*. 2016;352(6281):aac7341. doi:10.1126/science.aac7341

122. Ferreira JP, Overton KW, Wang CL. Tuning gene expression with synthetic upstream open reading frames. *Proc Natl Acad Sci U S A*. 2013;110(28):11284-11289. doi:10.1073/pnas.1305590110

123. Li J, Liang Q, Song W, Marchisio MA. Nucleotides upstream of the Kozak sequence strongly influence gene expression in the yeast S. cerevisiae. *J Biol Eng*. 2017;11:25. doi:10.1186/s13036-017-0068-1

124. Mutalik VK, Guimaraes JC, Cambray G, et al. Precise and reliable gene expression via standard transcription and translation initiation elements. *Nat Methods*. 2013;10(4):354-360. doi:10.1038/nmeth.2404

125. Chen Y, Kim JK, Hirning AJ, Josić K, Bennett MR. SYNTHETIC BIOLOGY. Emergent genetic oscillations in a synthetic microbial consortium. *Science*. 2015;349(6251):986-989. doi:10.1126/science.aaa3794

126. Zadeh JN, Steenberg CD, Bois JS, et al. NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem*. 2011;32(1):170-173. doi:10.1002/jcc.21596

127. Espah Borujeni A, Salis HM. Translation Initiation is Controlled by RNA Folding Kinetics via a Ribosome Drafting Mechanism. *J Am Chem Soc*. 2016;138(22):7016-7023. doi:10.1021/jacs.6b01453

128. Goñi-Moreno A, Nikel PI. High-Performance Biocomputing in Synthetic Biology-Integrated Transcriptional and Metabolic Circuits. *Front Bioeng Biotechnol*. 2019;7:40. doi:10.3389/fbioe.2019.00040

129. Healy CP, Deans TL. Genetic circuits to engineer tissues with alternative functions. *J Biol Eng*. 2019;13(1):39. doi:10.1186/s13036-019-0170-7

130. Fortman JL, Chhabra S, Mukhopadhyay A, et al. Biofuel alternatives to ethanol: pumping the microbial well. *Trends Biotechnol*. 2008;26(7):375-381. doi:10.1016/j.tibtech.2008.03.008

131. Ro D-K, Paradise EM, Ouellet M, et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*. 2006;440(7086):940-943. doi:10.1038/nature04640

132. Trosset J-Y, Carbonell P. Synthetic biology for pharmaceutical drug discovery. *Drug Des Devel Ther*. 2015;9:6285-6302. doi:10.2147/DDDT.S58049

133. Levin-Karp A, Barenholz U, Bareia T, et al. Quantifying Translational Coupling in *E. coli* Synthetic Operons Using RBS Modulation and Fluorescent Reporters. *ACS Synth Biol*. 2013;2(6):327-336. doi:10.1021/sb400002n

134. Liao MJ, Din MO, Tsimring L, Hasty J. Rock-paper-scissors: Engineered population dynamics increase genetic stability. *Science*. 2019;365(6457):1045-1049. doi:10.1126/science.aaw0542

135. Wu F, Zhang Q, Wang X. Design of Adjacent Transcriptional Regions to Tune Gene Expression and Facilitate Circuit Construction. *Cell Syst*. 2018;6(2):206-215.e6. doi:10.1016/j.cels.2018.01.010

136. Dai Z, Liu Y, Guo J, Huang L, Zhang X. Yeast synthetic biology for high-value metabolites. *FEMS Yeast Res*. 2015;15(1):1-11. doi:10.1111/1567-1364.12187

137. Si T, Chao R, Min Y, Wu Y, Ren W, Zhao H. Automated multiplex genome-scale engineering in yeast. *Nat Commun*. 2017;8(1):15187. doi:10.1038/ncomms15187

138. Walker RSK, Pretorius IS. Applications of Yeast Synthetic Biology Geared towards the Production of Biopharmaceuticals. *Genes (Basel)*. 2018;9(7). doi:10.3390/genes9070340

139. Demain AL, Vaishnav P. Production of recombinant proteins by microbes and higher organisms. *Biotechnol Adv*. 2009;27(3):297-306. doi:10.1016/j.biotechadv.2009.01.008

140. Brown AJ, Gibson SJ, Hatton D, James DC. In silico design of context-responsive mammalian promoters with user-defined functionality. *Nucleic Acids Res*. 2017;45(18):10906-10919. doi:10.1093/nar/gkx768

141. Cuperus JT, Groves B, Kuchina A, et al. Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res*. 2017;27(12):2015-2024. doi:10.1101/gr.224964.117

142. Curran KA, Crook NC, Karim AS, Gupta A, Wagman AM, Alper HS. Design of synthetic yeast promoters via tuning of nucleosome architecture. *Nat Commun*. 2014;5(1):4002. doi:10.1038/ncomms5002

143. Decoene T, Peters G, De Maeseneire SL, De Mey M. Toward Predictable 5'UTRs in Saccharomyces cerevisiae: Development of a yUTR Calculator. *ACS Synth Biol*. 2018;7(2):622-634. doi:10.1021/acssynbio.7b00366

144. Ding W, Cheng J, Guo D, et al. Engineering the 5′ UTR-Mediated Regulation of Protein Abundance in Yeast Using Nucleotide Sequence Activity Relationships. *ACS Synth Biol*. 2018;7(12):2709-2714. doi:10.1021/acssynbio.8b00127

145. Jia L, Yarlagadda R, Reed CC. Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. *PLoS One*. 2015;10(9):e0138022. doi:10.1371/journal.pone.0138022

146. Li Y, Fang J. PROTS-RF: A Robust Model for Predicting Mutation-Induced Protein Stability Changes. Srinivasan N, ed. *PLoS ONE*. 2012;7(10):e47247. doi:10.1371/journal.pone.0047247

147. Tian J, Wu N, Chu X, Fan Y. Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinformatics*. 2010;11:370. doi:10.1186/1471-2105-11-370

148. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. *Nat Methods*. 2019;16(8):687-694. doi:10.1038/s41592-019-0496-6

149. Webb S. Deep learning for biology. *Nature*. 2018;554(7693):555-557. doi:10.1038/d41586-018-02174-z

150. Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. *BMC Genomics*. 2017;18(S9):845. doi:10.1186/s12864-017-4226-0

151. Geron A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Incorporated; 2019.

152. Telser LG, Aitchison J, Brown JAC. The Lognormal Distribution. *Journal of Farm Economics*. 1959;41(1):161. doi:10.2307/1235218

153. Limpert E, Stahel WA, Abbt M. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*. 2001;51(5):341. doi:10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

154. Koch AL. The logarithm in biology 1. Mechanisms generating the log-normal distribution exactly. *Journal of Theoretical Biology*. 1966;12(2):276-290. doi:10.1016/0022-5193(66)90119-6

155. Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-Normalizing Neural Networks. *arXiv:170602515 [cs, stat]*. Published online September 7, 2017. Accessed October 12, 2020. http://arxiv.org/abs/1706.02515

156. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *arXiv:14126980 [cs]*. Published online January 29, 2017. Accessed October 12, 2020. http://arxiv.org/abs/1412.6980

157. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*. 2014;15(56):1929-1958.

158. Cambray G, Guimaraes JC, Arkin AP. Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nat Biotechnol*. 2018;36(10):1005-1015. doi:10.1038/nbt.4238

159. Michaels YS, Barnkob MB, Barbosa H, et al. Precise tuning of gene expression levels in mammalian cells. *Nat Commun*. 2019;10(1):818. doi:10.1038/s41467-019-08777-y

160. Delivering the promise of RNA therapeutics. *Nat Med*. 2019;25(9):1321-1321. doi:10.1038/s41591-019-0580-6

161. Kole R, Krainer AR, Altman S. RNA therapeutics: beyond RNA interference and antisense oligonucleotides. *Nat Rev Drug Discov*. 2012;11(2):125-140. doi:10.1038/nrd3625

162. Slomovic S, Pardee K, Collins JJ. Synthetic biology devices for in vitro and in vivo diagnostics. *Proc Natl Acad Sci USA*. 2015;112(47):14429-14435. doi:10.1073/pnas.1508521112

163. Hammer K, Mijakovic I, Jensen PR. Synthetic promoter libraries--tuning of gene expression. *Trends Biotechnol*. 2006;24(2):53-55. doi:10.1016/j.tibtech.2005.12.003

164. Siegl T, Tokovenko B, Myronovskyi M, Luzhetskyy A. Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes. *Metab Eng*. 2013;19:98-106. doi:10.1016/j.ymben.2013.07.006

165. de Smit MH, van Duin J. Translational initiation on structured messengers. Another role for the Shine-Dalgarno interaction. *J Mol Biol*. 1994;235(1):173-184. doi:10.1016/s0022-2836(05)80024-5

166. de Smit MH, van Duin J. Control of translation by mRNA secondary structure in Escherichia coli. A quantitative analysis of literature data. *J Mol Biol*. 1994;244(2):144-150. doi:10.1006/jmbi.1994.1714

167. Oesterle S, Gerngross D, Schmitt S, Roberts TM, Panke S. Efficient engineering of chromosomal ribosome binding site libraries in mismatch repair proficient Escherichia coli. *Sci Rep*. 2017;7(1):12327. doi:10.1038/s41598-017-12395-3

168. Chen Y-J, Liu P, Nielsen AAK, et al. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nat Methods*. 2013;10(7):659-664. doi:10.1038/nmeth.2515

169. Gardner PP, Barquist L, Bateman A, Nawrocki EP, Weinberg Z. RNIE: genome-wide prediction of bacterial intrinsic terminators. *Nucleic Acids Res*. 2011;39(14):5845-5852. doi:10.1093/nar/gkr168

170. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol*. 2007;8(2):R22. doi:10.1186/gb-2007-8-2-r22

171. Green AA, Kim J, Ma D, Silver PA, Collins JJ, Yin P. Complex cellular logic computation using ribocomputing devices. *Nature*. 2017;548(7665):117-121. doi:10.1038/nature23271

172. Kim J, Zhou Y, Carlson PD, et al. De novo-designed translation-repressing riboregulators for multi-input cellular logic. *Nat Chem Biol*. Published online November 4, 2019. doi:10.1038/s41589-019-0388-1

173. Hoynes-O'Connor A, Hinman K, Kirchner L, Moon TS. De novo design of heat-repressible RNA thermosensors in E. coli. *Nucleic Acids Res*. 2015;43(12):6166-6179. doi:10.1093/nar/gkv499

174. Matharu N, Rattanasopha S, Tamura S, et al. CRISPR-mediated activation of a promoter or enhancer rescues obesity caused by haploinsufficiency. *Science*. 2019;363(6424). doi:10.1126/science.aau0629

175. Hawkins JS, Wong S, Peters JM, Almeida R, Qi LS. Targeted Transcriptional Repression in Bacteria Using CRISPR Interference (CRISPRi). *Methods Mol Biol*. 2015;1311:349-362. doi:10.1007/978-1-4939-2687-9_23

176. Cleto S, Jensen JV, Wendisch VF, Lu TK. Corynebacterium glutamicum Metabolic Engineering with CRISPR Interference (CRISPRi). *ACS Synth Biol*. 2016;5(5):375-385. doi:10.1021/acssynbio.5b00216

177. Siu K-H, Chen W. Riboregulated toehold-gated gRNA for programmable CRISPR–Cas9 function. *Nat Chem Biol*. 2019;15(3):217-220. doi:10.1038/s41589-018-0186-1

178. Oesinghaus L, Simmel FC. Switching the activity of Cas12a using guide RNA strand displacement circuits. *Nat Commun*. 2019;10(1):2092. doi:10.1038/s41467-019-09953-w

179. Hanewich-Hollatz MH, Chen Z, Hochrein LM, Huang J, Pierce NA. Conditional Guide RNAs: Programmable Conditional Regulation of CRISPR/Cas Function in Bacterial and Mammalian Cells via Dynamic RNA Nanotechnology. *ACS Cent Sci*. 2019;5(7):1241-1249. doi:10.1021/acscentsci.9b00340

180. Pfleger BF, Pitera DJ, Smolke CD, Keasling JD. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat Biotechnol*. 2006;24(8):1027-1032. doi:10.1038/nbt1226

181. Leppek K, Das R, Barna M. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat Rev Mol Cell Biol*. 2018;19(3):158-174. doi:10.1038/nrm.2017.103

182. Win MN, Smolke CD. Higher-Order Cellular Information Processing with Synthetic RNA Devices. *Science*. 2008;322(5900):456-460. doi:10.1126/science.1160311

183. Win MN, Smolke CD. A modular and extensible RNA-based gene-regulatory platform for engineering cellular function. *Proceedings of the National Academy of Sciences*. 2007;104(36):14283-14288. doi:10.1073/pnas.0703961104

184. Sastalla I, Chim K, Cheung GYC, Pomerantsev AP, Leppla SH. Codon-optimized fluorescent proteins designed for expression in low-GC gram-positive bacteria. *Appl Environ Microbiol*. 2009;75(7):2099-2110. doi:10.1128/AEM.02066-08

185. Yang TT, Cheng L, Kain SR. Optimized codon usage and chromophore mutations provide enhanced sensitivity with the green fluorescent protein. *Nucleic Acids Res*. 1996;24(22):4592-4593. doi:10.1093/nar/24.22.4592

186. Rauhut R, Klug G. mRNA degradation in bacteria. *FEMS Microbiol Rev*. 1999;23(3):353-370. doi:10.1111/j.1574-6976.1999.tb00404.x

187. Hui MP, Foley PL, Belasco JG. Messenger RNA degradation in bacterial cells. *Annu Rev Genet*. 2014;48:537-559. doi:10.1146/annurev-genet-120213-092340

188. Carpousis AJ. The RNA degradosome of Escherichia coli: an mRNA-degrading machine assembled on RNase E. *Annu Rev Microbiol*. 2007;61:71-87. doi:10.1146/annurev.micro.61.080706.093440

189. Arnold TE, Yu J, Belasco JG. mRNA stabilization by the ompA 5' untranslated region: two protective elements hinder distinct pathways for mRNA degradation. *RNA*. 1998;4(3):319-330.

190. Bouvet P, Belasco JG. Control of RNase E-mediated RNA degradation by 5'-terminal base pairing in E. coli. *Nature*. 1992;360(6403):488-491. doi:10.1038/360488a0

191. Carrier TA, Keasling JD. Library of synthetic 5' secondary structures to manipulate mRNA stability in Escherichia coli. *Biotechnol Prog*. 1999;15(1):58-64. doi:10.1021/bp9801143

192. 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*. 2014;42(4):2646-2659. doi:10.1093/nar/gkt1139

193. Gusarov I, Nudler E. The mechanism of intrinsic transcription termination. *Mol Cell*. 1999;3(4):495-504. doi:10.1016/s1097-2765(00)80477-3

194. d'Aubenton Carafa Y, Brody E, Thermes C. Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures. *J Mol Biol*. 1990;216(4):835-858. doi:10.1016/s0022-2836(99)80005-9

195. Varani G. Exceptionally stable nucleic acid hairpins. *Annu Rev Biophys Biomol Struct*. 1995;24:379-404. doi:10.1146/annurev.bb.24.060195.002115

196. Caron M-P, Bastet L, Lussier A, Simoneau-Roy M, Massé E, Lafontaine DA. Dual-acting riboswitch control of translation initiation and mRNA decay. *Proc Natl Acad Sci USA*. 2012;109(50):E3444-3453. doi:10.1073/pnas.1214024109

197. Ng W-L, Bassler BL. Bacterial quorum-sensing network architectures. *Annu Rev Genet*. 2009;43:197-222. doi:10.1146/annurev-genet-102108-134304

198. Wu M, Su R-Q, Li X, Ellis T, Lai Y-C, Wang X. Engineering of regulated stochastic cell fate determination. *Proc Natl Acad Sci USA*. 2013;110(26):10610-10615. doi:10.1073/pnas.1305423110

199. Karzbrun E, Tayar AM, Noireaux V, Bar-Ziv RH. Synthetic biology. Programmable on-chip DNA compartments as artificial cells. *Science*. 2014;345(6198):829-832. doi:10.1126/science.1255550

200. Dudley QM, Karim AS, Jewett MC. Cell-free metabolic engineering: biomanufacturing beyond the cell. *Biotechnol J*. 2015;10(1):69-82. doi:10.1002/biot.201400330

201. Jeong D, Klocke M, Agarwal S, et al. Cell-Free Synthetic Biology Platform for Engineering Synthetic Biological Circuits and Systems. *Methods Protoc*. 2019;2(2). doi:10.3390/mps2020039

202. Huang A, Nguyen PQ, Stark JC, et al. BioBits™ Explorer: A modular synthetic biology education kit. *Sci Adv*. 2018;4(8):eaat5105. doi:10.1126/sciadv.aat5105

203. Viegas SC, Apura P, Martínez-García E, de Lorenzo V, Arraiano CM. Modulating Heterologous Gene Expression with Portable mRNA-Stabilizing 5'-UTR Sequences. *ACS Synth Biol*. 2018;7(9):2177-2188. doi:10.1021/acssynbio.8b00191

204. Russell JB. The energy spilling reactions of bacteria and other organisms. *J Mol Microbiol Biotechnol*. 2007;13(1-3):1-11. doi:10.1159/000103591

205. Kwon Y-C, Jewett MC. High-throughput preparation methods of crude extract for robust cell-free protein synthesis. *Sci Rep*. 2015;5(1):8663. doi:10.1038/srep08663

206. Lavickova B, Maerkl SJ. A Simple, Robust, and Low-Cost Method To Produce the PURE Cell-Free System. *ACS Synth Biol*. 2019;8(2):455-462. doi:10.1021/acssynbio.8b00427

207. 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*. 2019;8(2):403-414. doi:10.1021/acssynbio.8b00430

208. Standage-Beier K, Zhang Q, Wang X. Targeted Large-Scale Deletion of Bacterial Genomes Using CRISPR-Nickases. *ACS Synth Biol*. 2015;4(11):1217-1225. doi:10.1021/acssynbio.5b00132

209. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods*. 2009;6(5):343-345. doi:10.1038/nmeth.1318