From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Master's Theses, 2022-2026 Chemical and Biological Engineering

2025

Towards a Functional Search for Novel Riboswitches in the Human Gut Microbiome

Kim, Christina Young

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01p8418r63n

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

TOWARDS A FUNCTIONAL SEARCH FOR NOVEL RIBOSWITCHES IN THE HUMAN GUT MICROBIOME

CHRISTINA YOUNG KIM

A MASTER'S THESIS

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF MASTER IN SCIENCE IN ENGINEERING

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF

CHEMICAL AND BIOLOGICAL ENGINEERING

Adviser: Mohamed S. Donia

May 2025

© Copyright by Christina Young Kim, 2025. All rights reserved.

Abstract

The human gut microbiome is a large, complex microbial system containing trillions of different organisms.[7] Gut microbes are constantly adapting to one another and the human host environment. As a result, they have evolved a rich variety of functions to compete, cooperate, and utilize the digestive byproducts and xenobiotics they encounter.[1,2] Variations in gut microbiome composition and function have been associated with diseases from gastrointestinal to respiratory to autoimmune and even neurological.[4,5,6] Characterizing the function of the gut microbiome has thus emerged as an area of high interest in human health.

The gut microbiome presents a dynamic, competitive environment. Understanding how gut bacteria regulate gene expression is vital to understanding how the overall community will react to natural changes or clinical interventions. To expand the current understanding of gene regulation in the microbiome, I focused on one ncRNA, the riboswitch. Riboswitches regulate the expression of downstream genes by sensing specific ligands. In prokaryotes, riboswitches can either prematurely terminate transcription or block ribosome binding to prevent mRNA translation.[13] Being entirely composed of RNA and acting in *cis* on adjacent coding sequences, riboswitches can be easily expressed heterologously or in cell-free systems,[40] making them useful components of biosensors and other genetic circuits.

This thesis establishes steps towards identifying and characterizing novel riboswitches in the human gut. Towards that end, I had two aims: construct and screen metagenomic libraries to identify riboswitches; experimentally characterize riboswitch candidates computationally identified from gut microbiome metatranscriptomic data. For the first, I developed a protocol for creating genomic libraries from single strains isolated from human fecal samples, which can be refined further for metagenomic libraries from microbiome cultures. For the second, I created a

testing platform for cloning riboswitch candidates into fluorescent reporter *E. coli* strains and selected computational candidates for proof of concept testing. With further development, the research in this thesis can be used for a function-first search for novel riboswitches in the human gut microbiome. Discovering and characterizing human gut riboswitches furthers understanding of the human gut microbiome regulatory landscape and advances efforts to engineer its function.

Table of Contents

Abstract **3**
Table of Contents **5**
Chapter 1: Introduction **6**
1.1: The human gut microbiome: A brief overview 6
1.2: Riboswitches: ncRNA cis-regulatory elements 7
Chapter 2: Background and Motivation **10**
Chapter 3: Methodology **12**
3.1: Genomic library preparation 12
3.2: Riboswitch testing in vitro 14
3.3: Filtering computationally identified candidate riboswitches 16
Chapter 4: Results **17**
4.1: Development of genomic library protocol 17
4.2: Genomic library creation 19
4.3: Validation of riboswitch testing construct 20
4.4: Computational candidate selection, initial testing 23
Chapter 5: Conclusion & Future Recommendations **25**
References **27**

Chapter 1: Introduction

1.1: The human gut microbiome: A brief overview

The human gut microbiome is a vast, complex community composed primarily of bacteria – on the order of 10^{13} individual organisms in the average adult[7] – along with fungi, viruses, and other microorganisms. Variations in gut microbiome function and composition have been correlated to disparate areas of human health from asthma to colorectal cancer to major depressive disorder.[4,5,6] However, much of the microbiome's metagenome remains unexplored, with underlying mechanisms yet to be discovered. For instance, of one widely cited database of >170 million human gut microbiome proteins, only 40% are functionally annotated.[8] With its far-reaching effects on overall health, filling gaps in our understanding of the metagenome of the gut microbiome - both coding and noncoding - is therefore of great interest.

With its high interpersonal variability,[45] there is also emerging interest in manipulating the gut microbiome as a means of personalized medicine. However, thus far only one treatment has been approved by the FDA: fecal microbiotal transplants (FMTs), replacing a patient's native gut microbiome with that of a healthy donor, for chronic *C. difficile* infection.[11] While this treatment has been successful, FMTs still have considerable inherent variability between donors, often inconsistent patient response, and can risk transferring viruses and other pathogens from donor to recipient.[12] Currently, no approved clinical interventions exist that can make smaller, more precise manipulations to a patient's existing microbiome.

Engineering the gut microbiome requires an understanding of how its members will respond to any changes made. Characterizing the regulatory landscape of gut bacteria expands our knowledge of the dynamics of the microbiome environment, as well as providing potential tools for the engineering itself. To begin characterizing gut microbiome regulation, I have

decided to focus on *cis*-regulatory elements, which derive from non-coding DNA and act on neighboring genes. Narrowing in on elements that act in *cis* allows for much easier identification of the gene being regulated by any given regulatory element. Examples of *cis*-regulatory elements in bacteria include riboswitches, promoters, enhancers, and transcription factor binding sites.[26] Of these, I'm particularly interested in riboswitches as, unlike other regulatory elements, riboswitches are composed only of non-coding RNA (ncRNA) and do not require additional proteins to function. This independence makes riboswitches much simpler to clone into model systems like *E. coli*, both for study and for application. As such, I chose to focus on the characterization and discovery of novel riboswitches.

1.2: Riboswitches: ncRNA *cis*-regulatory elements

Riboswitches are stretches of non-coding RNA (ncRNA) that form ligand-binding aptamers. Generally found in the 5' untranslated regions (UTRs) of mRNA, riboswitches allow for small-molecule-responsive gene regulation without additional protein factors.[13] In prokaryotes, riboswitches function by one of two mechanisms: transcriptional or translational control.[41] In transcriptional control, riboswitch conformation determines whether an intrinsic terminator stem can form and trigger premature disassociation of the RNA polymerase. In translational control, changes in riboswitch conformation in the presence or absence of a ligand determine whether a ribosome can bind to the mRNA transcript and allow for translation into protein. Both control mechanisms may be used in either "on" or "off" switches; i.e., a given riboswitch will either allow for gene expression only when ligand is present, or only when ligand is absent.[41]

Due to their involvement in many key processes in bacteria, such as nutrient transport, riboswitches present attractive druggable targets in ongoing efforts to develop new antibiotics. Pyrithiamine, a synthetic thiamine analogue with antibacterial and antifungal properties, was developed prior to the discovery of the TPP riboswitch but was later shown to cause bacterial cell death by binding the riboswitch and inhibiting its function.[15] Targeting common riboswitches is of interest to create broad-spectrum antibiotics, while targeting less prevalent riboswitches and specific riboswitch variants presents an avenue for narrow-spectrum antibiotics. For instance, a synthetic flavin analogue called 5FDQD has been developed to selectively target *C. difficile* while minimizing effects on other gut microbes.[17] As antibiotic resistance becomes an increasingly pressing issue in public health, riboswitch target analogues present attractive drug candidates, especially as ligand-binding domains are highly conserved[41] and therefore less able to evolve resistance.

In the field of genetic engineering, riboswitches are also useful for the construction of biosensors and other genetic circuits. For instance, Thavarajah, et al. incorporated a fluoride riboswitch taken from *B. cereus* into a cell-free system for the detection of unsafe levels of fluoride in drinking water.[40] Since transcriptional riboswitches only require a DNA template and RNAP to function, the researchers were able to create a paper strip test for fluoride testing in areas with limited resources. Once the relevant sequence and ligand are identified, riboswitches can be quickly and efficiently implemented in a range of applications.

Despite advances made in riboswitch discovery at large, the overall landscape of human gut microbiome riboswitches remains largely uncharacterized. Even so, bioinformatic searches have identified at least one class from gut-associated bacterial genomes: a conserved RNA motif was identified across multiple members of class *Clostridia*, and representative riboswitches from

C. scindens and *C. botulinum* were shown to respond to toxic concentrations of Ni^{II} and Co^{II}.[19] With the amount of uncharacterized sequences in the gut metagenome, further novel riboswitch classes, or novel usages of existing classes, likely remain to be discovered.

Chapter 2: Background and Motivation

In the years since riboswitches were first identified in 2002, as the most abundant riboswitch classes have been identified and characterized, the currently predominant strategy for identifying the rarer, still-undiscovered classes has been based in bioinformatics.[37] These computational approaches utilize comparative sequence analyses, searching genomic databases for highly conserved motifs.[41] Riboswitches contain specialized binding pockets for their target ligands and as such must be highly conserved in order to retain their function.[37] Candidate motifs identified using this approach are then generally tested using in-line probing,[33] an assay that takes advantage of RNA's inherent instability to look for a change in RNA sequence degradation patterns in the presence of ligand. Such a change is indicative of a structural change in the RNA induced by the ligand in the absence of any other components, such as protein factors. However, in-line probing is time intensive and can only test one candidate ligand at a time. While the bioinformatic approach has identified multiple experimentally verified novel classes,[46-48] many "orphan" riboswitch classes remain.[37] Orphans have sequence and structural properties similar enough to known classes such that they are likely to be riboswitches, but their binding ligands remain to be identified.

While computational approaches are powerful, searching for conserved sequences may miss riboswitches with few representatives in datasets, and searching for sequences similar to known classes is inherently limiting when attempting to discover novel classes. Meanwhile, experimental approaches have mainly focused on adapting existing riboswitches[29,31] and creating synthetic riboswitches for biosensing applications.[43] However, approaches for directed evolution of synthetic aptamers have also been adapted to the discovery of naturally occurring riboswitches. Genomic SELEX ("Systematic Evolution of Ligands with EXponential

enrichment") uses genomic DNA as the template for a library of RNA aptamers[49] and has led to the discovery of several ligand binding pocket domains.[36] Further research is needed to verify whether these domains represent novel riboswitch aptamers or instead exhibit serendipitous binding when removed from their original context.

The work in this thesis was motivated by a functional screening approach that seeks to avoid excluding potential riboswitches that are under- or unrepresented in current datasets. In previous work, the Donia lab constructed and screened a metagenomic library to identify a hydrocortisone reductase gene from a human donor microbiome that had been observed to metabolize the drug.[28] Adapting that protocol, metagenomic libraries could be constructed and screened drawing on existing techniques for substrate-induced gene expression (SIGEX) screening. SIGEX is a library functional screening method that was developed to isolate novel catabolic genes from environmental metagenomes.[36] In SIGEX, a metagenomic library is constructed using an environmental sample and a fluorescent reporter, and then FACS is used to isolate the low fluorescence population in the absence of the ligand of interest. This population is cultured with the ligand, and the strains which then begin to express high levels of the fluorescent reporter are sorted and sequenced.[36] Similar FACS approaches have also been used to screen synthetic riboswitches,[43] adapting the screening logic to search for changes in fluorescence expression in the presence of ligand. While this approach also takes inspiration from genomic SELEX, a library of *E. coli* strains carrying metagenomic DNA fragments in plasmids (as in SIGEX) is more stable than the pools of short RNA sequences required for SELEX and thus simpler to reuse and test across large pools of potential ligands.

Chapter 3: Methodology

3.1: Genomic library preparation

Fecal sample-derived single strain isolates were obtained in glycerol stock form from Catherine Day. Following 24-48 hour anaerobic culture at 37°C in BG media,[28] bacteria were spun down at 3900 rpm in a benchtop centrifuge. Resulting pellets were massed and genomic DNA (gDNA) was extracted using Zymo Quick-DNA Fungal/Bacterial Miniprep kit as instructed by the manufacturer.

The parent vector pGFPuv was miniprepped from Top10 *E. coli* stocks pre-existing in the lab. Isolated vector pGFPuv was linearized using PCR (PCR primers: forward, pGFP-IF-F = TAATGAATTCCAACTGA GCGCCGG and reverse, pGFP-IF-R = CATAGCTGTTTCCTGT GTGAAATTG. See Tables 1, 2.) The resulting 2.5 kbp product was verified using gel electrophoresis of 2 µL product diluted with 3 µL dH_2O. The remaining linearized product was brought up to 100 µL with dH_2O and cleaned using a Zymo DNA Clean & Concentrator kit as directed by the manufacturer. Final concentration of the cleaned product was verified using a Qubit fluorometer. To eliminate residual circular vector, the cleaned linearized vector was incubated with 1 µL Dpn1 restriction enzyme per µg DNA (plus CutSmart buffer) at 37°C for 1 hour, then incubated at 80°C for 20 minutes to inactivate the Dpn1. To prevent re-ligation, the vector was then incubated with 1 µL Antarctic phosphatase (AnP) per µg DNA (plus AnP reaction buffer) at 37°C for 15 minutes, then at 70°C for 5 minutes to inactivate the AnP. Finally, the 2.5 kbp product was isolated using gel electrophoresis: the appropriate band was excised, mashed, and then extracted using the QIAquick Gel Extraction kit as directed by the manufacturer.

Component	Volume (µl)
5X Phusion HF or GC Buffer	4
10 mM dNTPs	0.4
10 uM pGFP-IF-F	1
10 uM pGFP-IF-R	1
pGFPuv (10-15 ng/µl)	1
Phusion DNA polymerase	0.2
Milli-Q water	Make up to 20 µl total volume

Table 1. PCR reaction mixture for pGFPuv vector linearisation.

Step	Temperature (°C)	Time
Initial Denaturation	98	30 seconds
25-35 Cycles	98	10 seconds
	68	30 seconds
	72	30 seconds per 1 kbp product
Final Extension	72	5 minutes
Hold	4	Indefinite

Table 2. PCR thermocycler program for pGFPuv vector linearisation.

In parallel, extracted gDNA was sheared to a target size of 4-6 kbp using Covaris g-TUBEs. Four µg gDNA was brought up to 150 µL in dH_2O, loaded into a g-TUBE, and spun twice at 13,200 rpm for 30 seconds in an accuSpin Micro 17R Microcentrifuge with the "soft" setting disabled. The resulting fragments were then removed from the g-TUBE using a micropipette, brought back up to 150 µL in dH_2O if necessary, and the spin procedure repeated in a fresh g-TUBE. Fragment size was validated by running 10 µL of the sample through gel electrophoresis; if desired sample size was obtained, the sample volume was reduced to 30 µL or less on a SpeedVac. If desired sample size was not obtained, the g-TUBE spin procedure was repeated.

Once sample volume was reduced, gDNA fragments were isolated using gel electrophoresis, with bands excised on a blue light box to avoid potential degradation under UV light. The 4-6 kbp band was then mashed and extracted using a QIAquick Gel Extraction kit as

directed by the manufacturer. The extracted gDNA fragments were then end-repaired using a Thermo Scientific Fast DNA End-Repair kit and cleaned using a Zymo DNA Clean & Concentrator kit.

Linearized pGFPuv and end-repaired gDNA fragments were blunt-end ligated using T4 DNA ligase as directed by the manufacturer: 50 ng vector and 300 ng gDNA were incubated with ligase and buffer at 16°C overnight. The resulting ligated plasmid was precipitated using Pellet Paint NF Co-precipitant as directed by the manufacturer and brought up to 2 µL in dH_2O. Two µL ligated plasmid was transformed into MegaX DH10ß Electrocompetent Cells as directed by the manufacturer, then recovered in 1 mL LB medium.

The transformation product was brought up to 10.2 mL in LB with ampicillin, then divided into ten pools of 1 mL each. Each pool was incubated at 30°C overnight, then 0.75 mL of each was mixed with 0.75 mL of 50% glycerol and frozen at -80°C for long term storage. The remaining 0.2 mL transformation product was used to create serial dilutions, spread on LB/ampicillin plates, and incubated at 37°C overnight. Plates with countable amounts of colonies were counted to calculate library titer, as well as to pick random colonies for colony PCR to sample insert size.

3.2: Riboswitch testing *in vitro*

A plasmid was constructed from pGFPuv (see Fig. 1), sandwiching the riboswitch or riboswitch candidate between a *lac* promoter and the gene for GFPuv, a version of green fluorescent protein (GFP) that has been codon-optimized for *E. coli* and exhibits much higher emission than wild-type GFP under UV light.[39] While not used in the riboswitch screening, UV excitation produces emission strong enough to be seen by the naked eye on an agar plate or in a

cell pellet from liquid culture, allowing for convenient verification of GFPuv's presence or absence. The reporter construct allows for the measurement of fluorescence intensity as an indicator of riboswitch activity and the establishment of any dose-response relationship to the ligand(s) of interest.

The reporter construct was cloned into Stellar *E. coli* and stored in glycerol stocks at -80°C. For each test, the strain carrying the reporter - alongside a pGFPuv constitutive expression control and a non-fluorescent negative control (carrying carbenicillin resistance on a pUC19 plasmid, but without any fluorescent reporter) - was cultured overnight in LB media on a shaking incubator, then diluted to OD_{600} 0.05 and cultured overnight on a Tecan plate reader in M9 media supplemented with the appropriate dose of ligand (an escalating series of 0, 1, 10, and 100 µM, unless constrained by ligand solubility). All strains carried resistance to and were cultured with carbenicillin.

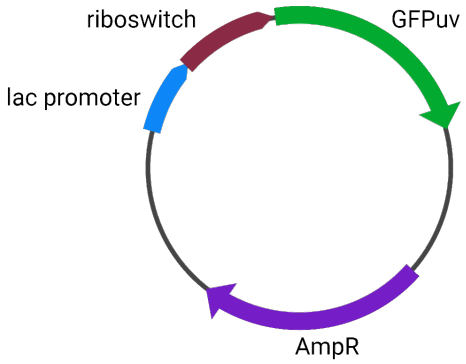


Figure 1. Riboswitch testing plasmid map. Due to its greater stability, strains were cultured with the related drug carbenicillin rather than ampicillin. (Created with BioRender.)

Plate reader results were then verified using flow cytometry. After being cultured on a shaking incubator as in the plate reader protocol, cultures were diluted to OD_{600} 0.05 in M9 supplemented with the appropriate dose of ligand and grown overnight in triplicate in a shaking incubator. The next day, the cultures were diluted to 10^6 cells/mL and run on a BD LSR II flow

cytometer. The negative control strain was used to set thresholds for expected cell size and baseline autofluorescence.

3.3: Filtering computationally identified candidate riboswitches

Ivy Chen, my computational collaborator, repurposed her existing computational pipeline for identifying small RNAs to search for riboswitches. The Donia lab has previously collected metatranscriptomic data from cultures of fecal samples from 20 healthy human donors. Chen started with that metatranscriptomic dataset, filtered out the coding DNA sequences (CDSes), isolated the non-coding intergenic regions, then searched them against a widely used non-coding RNA (ncRNA) family database, Rfam,[38] to determine which ncRNAs represented potential riboswitches.

Chen classified over 20,000 riboswitch candidate sequences falling into two categories: 1) 597 annotated as known riboswitch classes in Rfam, and 2) approximately 21,600 that do not match any known riboswitch class or any other known ncRNA family on Rfam. I then curated the candidate list further: first, to those riboswitch candidates with CDSes of known function (200 annotated, ~4700 unannotated); then, those 4700 unannotated candidates were limited to 818 which were less than 600 bp long. Typical riboswitches measure ~50-250 bp long;[42] the cutoff was selected to give some margin for error in case some identified sequences contained more than the riboswitch.

Chapter 4: Results

4.1: Development of genomic library protocol

The protocol was adapted from prior work by Javdan et al.[28] First, the protocol was assembled from the 2020 publication, accompanying supplementary materials, and lab notebooks, replacing components that had been discontinued. However, perhaps the most essential step, the mechanical shearing that generates random genomic DNA (gDNA) fragments, immediately required modification. The original protocol specified a single spin of 11,000 rpm for a target insert size of 2-4 kbp, which failed to be replicated over multiple trials in my testing. Increasing spin time would not have an effect; g-TUBEs shear DNA strands by forcing them through a narrow orifice. If the spin time is already sufficient to force through the full sample, increasing it will not increase the extent of shearing. Therefore, I tested the effects of shearing at 11,000 rpm vs increased speeds of 13,200 and 14,500 rpm. Even at 14,500 rpm, beyond the manufacturer's specifications for what the g-TUBEs can withstand, a target size of 2-4 kbp could not be achieved. The target size range was then relaxed to 4-6 kbp, which could be achieved using two sequential g-TUBE shears at 13,200 rpm. Small amounts of sample loss due to g-TUBE leaks were occasionally observed. Particularly precious samples were sheared at 11,000 rpm to reduce the chance of g-TUBE damage and subsequent sample loss, but this could necessitate a third sequential g-TUBE shear to achieve the same results.

It also became apparent that modifications were needed to preserve DNA yields throughout the many steps of the protocol, especially in the shearing and processing of gDNA. The shearing step was the largest source of sample loss: from a starting amount of 4000 ng of gDNA, early trials could result in yields as little as 350 ng of gDNA fragments after shearing and gel extraction. Cleaning steps would then reduce the yield further to less than 300 ng, insufficient

for a single ligation reaction. To minimize opportunities for sample loss, I replaced gel electrophoresis and subsequent gel extraction steps with DNA Clean & Concentrator where possible (bringing samples up to 100 µL before beginning the protocol also helped to minimize sample losses in the cleaning step). I also imaged gels containing gDNA fragments using blue light, rather than UV, to minimize degradation of target size fragments. To maximize yield during gel extraction, I tested Qiagen's QIAquick Gel Extraction kit against the protocol's original recommendation of the Zymoclean Large Fragment DNA Recovery kit and found that the QIAquick kit could provide over twice as much yield. Mashing agar gel prior to beginning the extraction protocol decreased melting time and further minimized risks of gDNA degradation. Other tested modifications, such as heating the extraction buffers, had minimal effect and were discontinued. With all shearing and cleaning step modifications taken together, final yields of gDNA fragments increased from around 250 ng to 500-700 ng from 4000 ng of starting gDNA.

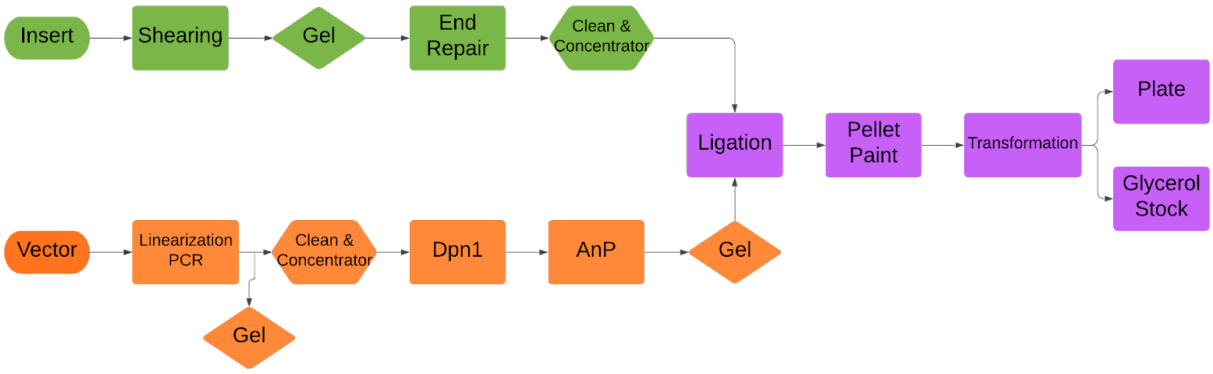


Figure 2. Final library protocol flowchart. Flowcharts were produced at intermediate phases of protocol development to summarize the full process and identify steps most responsible for DNA fragment loss.

Once all modifications were finalized, the protocol (see Fig. 2) reliably produced gDNA fragments and linearized vector in quantities sufficient for the ligation and transformation steps. Further refinement would be needed in order to improve ligation and transformation efficiencies

in order to reach the library size needed to transfer the protocol from single genomes to full microbiome metagenomes.

4.2: Genomic library creation

Catherine Day, another graduate student from the Donia Lab, provided me with single species stocks, one *V. atypica* and one *W. confusa*, that were isolated from a human donor fecal sample and whole-genome sequenced. Both strains had been observed by Day to inactivate the antibiotic colistin; genomic libraries can be used to test for the gene responsible for the inactivation phenotype. For my purpose, these donor isolates are relatively easy to culture at larger scales in an anaerobic chamber, making them easier to work with for the development of the library construction protocol than the donor samples needed for metagenomic libraries.

I produced full genomic libraries for both strains and estimated the size of the library and level of genomic coverage for each one. Following the established library protocol, I measured the number of unique clones per library by serially diluting and plating the transformation stock followed by CFU counting. I also used colony PCR (sampling 16 colonies from each library and amplifying their inserted genomic DNA fragments) and gel electrophoresis to estimate the average insert size. Multiplying the number of unique clones with the average insert size, scaled by a correction factor consisting of the proportion of inserts >0.5 kb on the colony PCR, gives the size of the library.[28] The times coverage is the ratio between the library size and the size of the genome. The *V. atypica* library had an average insert size of 0.9 kb and an estimated library size of 48 Mbp, for approximately 23 times coverage of a 2.1 Mbp genome. The *W. confusa* library had an average insert size of 0.6 kb and an estimated library size of 42 Mbp, for approximately 16 times coverage of a 2.6 Mbp genome. I also transferred the completed libraries back to Day;

she tested the libraries for the ability to inactivate colistin and observed that several strains were able to recapitulate the colistin inactivation phenotype (see Fig. 3).

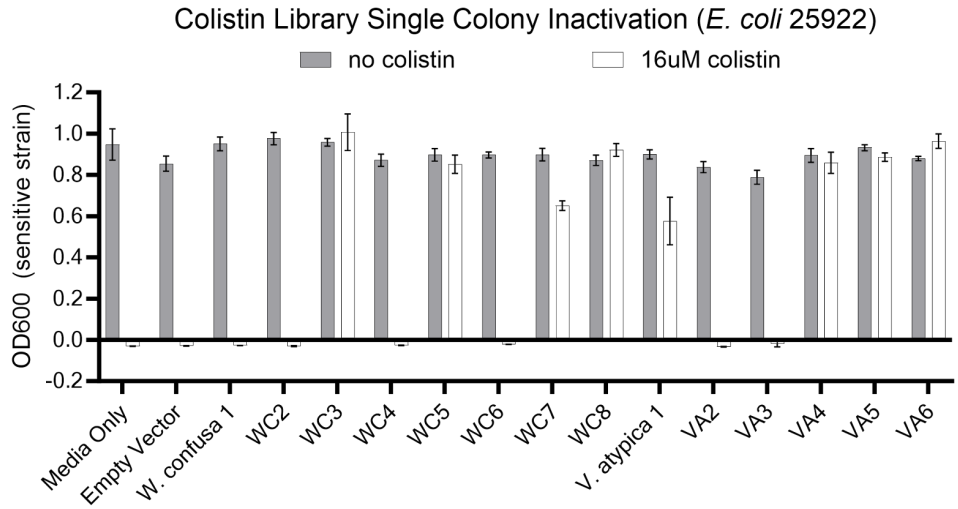


Figure 3. Figure courtesy of Catherine Day. Genomic libraries generated from *W. confusa* and *V. atypica* recapitulated the phenotype of interest: Colonies grown on plates with colistin were then cultured with 16 μM colistin. The spent media was then used to grow a susceptible strain of *E. coli*. Survival of the susceptible strain indicates successful inactivation.

4.3: Validation of riboswitch testing construct

I have created and tested a reporter construct (see section 3.2) with a synthetic thiamine pyrophosphate (TPP) riboswitch from the literature, ThiMn19, which has a known dose-response with a GFPuv reporter at varying concentrations of added thiamine.[31] After several iterations, the TPP reporter construct functions as expected: fluorescence activity is repressed below constitutive baseline expression in the absence of thiamine, before rapidly increasing in the presence of even 1 μM thiamine and further increasing with each increasing dose, reaching an apparent peak after 100 μM (see Fig. 4). The Stellar *E. coli* strain carrying the reporter - alongside a pGFPuv constitutive expression control and a non-fluorescent negative control - was cultured overnight in LB media, then diluted to OD_{600} 0.05 and cultured overnight in M9 media supplemented with the appropriate dose of thiamine. Following established protocols for

measuring riboswitch-mediated GFP activity,[29,31,43] I tested the TPP reporter twice on a plate reader, each time growing in triplicate and comparing normalized fluorescence between the riboswitch reporter strain and the constitutive and negative control strains. Across both tests, the reporter strain showed a low level of GFPuv fluorescence in the absence of thiamine and a significant increase in GFPuv fluorescence with the addition of thiamine, increasing further with increasing dose up to 100 µM.

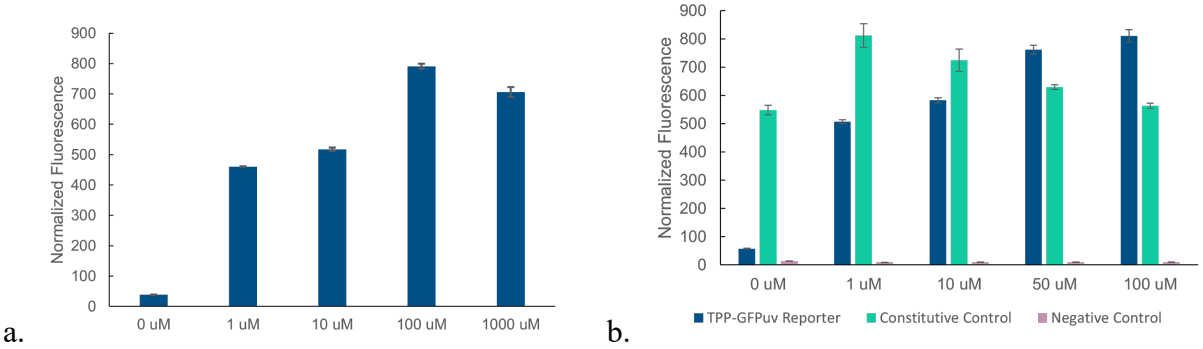


Figure 4. Endpoint fluorescence readings from plate reader. Strains were cultured at each dose in triplicate; error bars represent ±1 standard deviation. a) The TPP-GFPuv reporter cultured at 0, 1, 10, 100, and 1000 µM thiamine, with normalized fluorescence appearing to peak after 100 µM. b) Endpoint fluorescence measurements for the TPP-GFPuv reporter, constitutive expression control, and negative control at 0, 1, 10, 50, and 100 µM thiamine.

I further validated the reporter construct using flow cytometry. Based on the plate reader results, I selected an abbreviated dose series of 0, 10, and 100 µM thiamine. Instead of culturing on a plate reader, I diluted initial LB cultures into M9 with thiamine and grew them overnight in triplicate in a shaking incubator. The next day, the cultures were diluted to 10^6 cells/mL and run on a flow cytometer. The negative control strain was used to set thresholds for expected cell size and baseline autofluorescence. Representative histograms of individual samples are shown in Figure 5. Figure 5a shows the negative control, as well as the fluorescence gate that I used to threshold the data for Figure 6. Figure 5b shows a representative TPP-GFPuv sample cultured at 10 µM thiamine; as expected, the vast majority of the detected cells show strong fluorescence intensity. However, the constitutive expression control (representative example in Fig. 5c) did not behave as expected, instead showing a bimodal distribution. Suspecting potential contamination

of my starting glycerol stock, I conducted a follow-up fluorescence-activated cell sorting (FACS) experiment to separate the lowest and highest fluorescence intensity populations; sequencing of the full plasmids from both populations revealed whole, intact pGFPuv plasmid. For future flow cytometry experiments, a constitutive expression control with a strongly unimodal distribution of fluorescence intensity would be required, either by obtaining a new stock of constitutively-GFPuv-expressing *E. coli* or by modifying the protocol, depending on whether the observed bimodal distribution is due to a chromosome mutation or is caused by some factor in the protocol. While I was unable to obtain high quality flow cytometry data for constitutive levels of GFPuv expression, I was able to threshold the data for *E. coli* autofluorescence based on my nonfluorescent control and generate Figure 6: while the non-fluorescent control population remains effectively 0% fluorescent at all thiamine concentrations (see Fig. 6b), the TPP-GFPuv population shifts from <1% fluorescent in the absence of thiamine to >95% fluorescent at 10 and 100 µM thiamine.

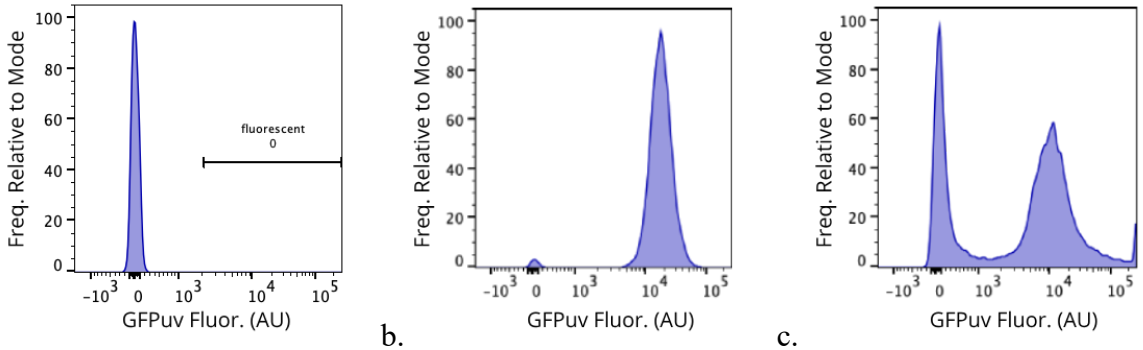
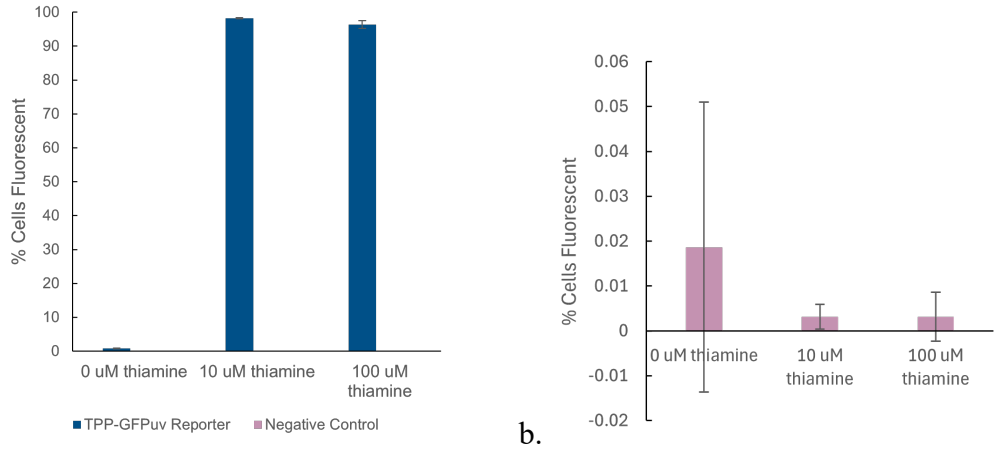


Figure 5. Representative histograms of single samples from flow cytometry. The y-axis represents the frequency of a detected cell relative to the frequency of the mode, while the x-axis represents GFPuv fluorescence intensity. a) The negative control strain at 0 µM thiamine, shown with the gate used to determine the frequency of fluorescent vs non-fluorescent cells. b) The TPP-GFPuv reporter strain at 10 µM thiamine. c) The pGFPuv constitutive control strain at 0 µM thiamine, displaying an unexpectedly bimodal distribution.



a. b.

Figure 6. Frequencies of fluorescent cells of negative control and TPP-GFPuv reporter strain cells on flow cytometry. Strains were cultured at each dose in triplicate and measured separately; error bars represent ±1 standard deviation. a) Data for the non-fluorescent negative control and TPP-GFPuv reporter strains. b) Isolated data for the non-fluorescent negative control, showing that the thresholding results in a value of effectively 0% fluorescence when variation between replicates is taken into account.

4.4: Computational candidate selection, initial testing

Initial filtering limited the original candidate list of over 22,000 to 818 candidates that could not be annotated to any family on Rfam, and 200 candidates that were annotated as members of known riboswitch classes (see section 3.3, Fig. 7) Of those 818, I selected 139 unannotated riboswitch candidates as having gene products of interest. Of the 200 annotated riboswitch candidates with known CDSes, the most common gene products were transporters (60/200 = 30%), followed by synthases (28/200 = 14%). Accordingly, I prioritized unannotated candidates accompanying transport and synthesis CDSes, along with other functions of general interest in microbiome research, such as antibiotic resistance. Finally, all of the candidates were filtered based on the number of base pairs between the end of the ncRNA transcript and the beginning of the CDS mRNA transcript; a threshold gap size of 100 bp or less was selected in order to minimize gap size while still capturing as many candidates of interest as possible. After curation, I was left with a shortlist of 173 annotated riboswitch candidates and 75 unannotated candidates.

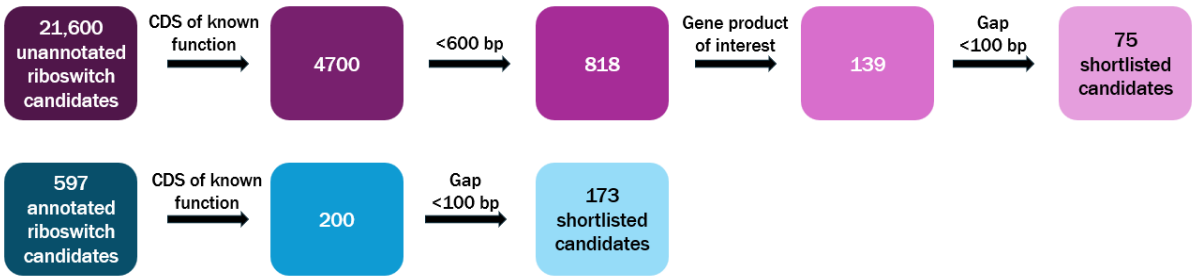


Figure 7. Candidate curation overview.

As proof of context, I selected four computationally identified riboswitches that were annotated as members of known classes for testing *in vitro.* These four consisted of three cobalamin (vitamin B12) riboswitches and one thiamine pyrophosphate (TPP) riboswitch. I ordered all four as IDT gBlocks and cloned them into the plasmid construct described in section 3.2, and tested all four using the protocol also described in 3.2. However, the experimental *E. coli* strains failed to grow sufficiently in initial experiments to determine whether a lack of observed response to ligand was due to insufficient growth or true lack of response. Further testing would be required to determine whether the testing protocol requires modifications for wild-type riboswitches rather than the synthetic TPP riboswitch that was used for its development.

Chapter 5: Conclusion & Future Recommendations

Greater understanding of gene regulation in the human gut microbiome both advances our overall understanding of a complex microbial community, and adds new tools to our arsenal in developing minimally invasive microbiome interventions. Novel riboswitches in this setting could provide gut microbiome-specific aptamers that can be applied in sensing applications: for instance, an engineered probiotic could carry a riboswitch ensuring that the regulated gene is only expressed in the colon instead of the small intestine or stomach.

The research in this thesis lays the groundwork for a function-first, metagenomic library-based search that would identify putative riboswitches together with their ligands from human gut microbiome cultures, as well as a platform for testing computationally identified riboswitch candidates *in vitro*. With further improvements to ligation and transformation efficiency, the genomic library protocol can be used to generate metagenomic libraries of riboswitch candidates within the riboswitch testing plasmid. Libraries of small molecules associated with the gut microbiome are commercially available[44]; library strains could be cultured alongside these small molecules and screened for ligand-mediated changes in fluorescence.

Identifying novel riboswitch classes in the gut microbiome would give us information about what gut-associated molecules are so prioritized for sensing that binding aptamers have evolved. Quantifying the frequencies of what molecules are detected is also valuable, for example if sensing certain nutrients is more common than sensing toxins that would give us insight into which factors have the most impact in regulation. Finding new riboswitches would also continue to inform and improve the computational identification approach. As computationally identified candidates are tested, information about the best-performing

candidates can be used to improve the computational pipeline. For example, if a previously unannotated candidate is confirmed to be a novel riboswitch, the datasets can be searched for similar sequences that may have previously been missed.

Furthermore, though the annotated riboswitch candidates are unlikely to yield truly novel riboswitch classes, the data may contain riboswitches of known classes that are being put to new or unexpected uses by gut bacteria. For example, if one species uses the concentration of a certain nutrient to regulate the expression of a seemingly unrelated toxin, that indicates that that nutrient may have a previously undescribed role in that species' defense systems. To further investigate that relationship and its effects on larger bacterial community dynamics, the donor microbiome community of origin could then be cultured *ex vivo.* 16S sequencing and transcriptomic sequencing could be used to observe how community makeup and gene expression changed in the presence and absence of that specific nutrient.

Identifying the gut microbiome's riboswitches would give us insight into regulatory activities potentially exclusive to the gut environment, as well as expanding tools for engineering gut microbiome function. Continuing to improve the accuracy of our computational approach would contribute to broader ongoing efforts in riboswitch discovery, especially in large metagenomic datasets. Overall, the work in this thesis provides tools to search for and characterize novel riboswitches, advancing efforts to understand the overall regulatory landscape of the human gut microbiome.

References

1. Heintz-Buschart, A., & Wilmes, P. (2018). Human gut microbiome: Function matters. *Trends in microbiology, 26*(7), 563-574.

2. Maurice, C. F., Haiser, H. J., & Turnbaugh, P. J. (2013). Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell, 152*(1), 39-50.

3. Pauer, H., Teixeira, F. L., Robinson, A. V., Parente, T. E., De Melo, M. A., Lobo, L. A., ... & Antunes, L. C. M. (2021). Bioactive small molecules produced by the human gut microbiome modulate *Vibrio cholerae* sessile and planktonic lifestyles. *Gut Microbes*, *13*(1), 1918993.

4. Depner, M., Taft, D. H., Kirjavainen, P. V., Kalanetra, K. M., Karvonen, A. M., Peschel, S., ... & Ege, M. J. (2020). Maturation of the gut microbiome during the first year of life contributes to the protective farm effect on childhood asthma. *Nature medicine, 26*(11), 1766-1775.

5. Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., ... & Yang, L. (2013). Human gut microbiome and risk for colorectal cancer. *Journal of the National Cancer Institute, 105*(24), 1907-1911.

6. Foster, J. A., Baker, G. B., & Dursun, S. M. (2021). The relationship between the gut microbiome-immune system-brain axis and major depressive disorder. *Frontiers in neurology*(12), 721126.

7. Sender, R., Fuchs, S., & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS biology, 14*(8), e1002533.

8. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., ... & Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature biotechnology, 39*(1), 105-114.

9. Puurunen, M. K., Vockley, J., Searle, S. L., Sacharow, S. J., Phillips III, J. A., Denney, W. S., ... & Brennan, A. M. (2021). Safety and pharmacodynamics of an engineered E. coli Nissle for the treatment of phenylketonuria: A first-in-human phase 1/2a study. *Nature Metabolism, 3*(8), 1125-1132.

10. Vazquez-Castellanos, J. F., Biclot, A., Vrancken, G., Huys, G. R., & Raes, J. (2019). Design of synthetic microbial consortia for gut microbiota modulation. *Current Opinion in Pharmacology, 49,* 52-59.

11. Lee, C., Louie, T., Bancke, L., Guthmueller, B., Harvey, A., Feuerstadt, P., ... & Dubberke, E. R. (2023). Safety of fecal microbiota, live-jslm (REBYOTA™) in individuals with recurrent *Clostridioides difficile* infection: Data from five prospective clinical trials. *Therapeutic Advances in Gastroenterology, 16*, 17562848231174277.

12. Panchal, P., Budree, S., Scheeler, A., Medina, G., Seng, M., Wong, W. F., ... & Osman, M. (2018). Scaling safe access to fecal microbiota transplantation: Past, present, and future. *Current gastroenterology reports, 20*, 1-11.

13. Breaker, R. R. (2009). Riboswitches: From ancient gene-control systems to modern drug targets. *Future microbiology, 4*(7), 771-773.

14. Li, S., & Breaker, R. R. (2013). Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucleic acids research, 41*(5), 3022-3031.

15. Sudarsan, N., Cohen-Chalamish, S., Nakamura, S., Emilsson, G. M., & Breaker, R. R. (2005). Thiamine pyrophosphate riboswitches are targets for the antimicrobial compound pyrithiamine. *Chemistry & biology, 12*(12), 1325-1335.

16. Panchal, V., & Brenk, R. (2021). Riboswitches as drug targets for antibiotics. *Antibiotics, 10*(1), 45.

17. Blount, K. F., Megyola, C., Plummer, M., Osterman, D., O'Connell, T., Aristoff, P., ... & Breaker, R. R. (2015). Novel riboswitch-binding flavin analog that protects mice against *Clostridium difficile* infection without inhibiting cecal flora. *Antimicrobial agents and chemotherapy, 59*(9), 5736-5746.

18. Genee, H. J., Bali, A. P., Petersen, S. D., Siedler, S., Bonde, M. T., Gronenberg, L. S., ... & Sommer, M. O. (2016). Functional mining of transporters using synthetic selections. *Nature chemical biology, 12*(12), 1015-1022.

19. Furukawa, K., Ramesh, A., Zhou, Z., Weinberg, Z., Vallery, T., Winkler, W. C., & Breaker, R. R. (2015). Bacterial riboswitches cooperatively bind Ni2+ or Co2+ ions and control expression of heavy metal transporters. *Molecular cell, 57*(6), 1088-1098.

20. Xu, J., & Cotruvo Jr, J. A. (2022). Iron-responsive riboswitches. *Current opinion in chemical biology*, 68, 102135.

21. Seshasayee, A. S. N., Sivaraman, K., & Luscombe, N. M. (2011). An overview of prokaryotic transcription factors: A summary of function and occurrence in bacterial genomes. *A handbook of transcription factors*, 7-23.

22. Kahramanoglou, C., Seshasayee, A. S., Prieto, A. I., Ibberson, D., Schmidt, S., Zimmermann, J., ... & Luscombe, N. M. (2011). Direct and indirect effects of H-NS and Fis on global gene expression control in *Escherichia coli*. *Nucleic acids research, 39*(6), 2073-2091.

23. Li, J. W., Zhang, X. Y., Wu, H., & Bai, Y. P. (2020). Transcription factor engineering for high-throughput strain evolution and organic acid bioproduction: A review. *Frontiers in Bioengineering and Biotechnology*, 8, 98.

24. Grau, M. F., Entwistle, R., Chiang, Y. M., Ahuja, M., Oakley, C. E., Akashi, T., ... & Oakley, B. R. (2018). Hybrid transcription factor engineering activates the silent secondary metabolite gene cluster for (+)-asperlin in Aspergillus nidulans. *ACS chemical biology, 13*(11), 3193-3205.

25. Nichols, R. G., & Davenport, E. R. (2021). The relationship between the gut microbiome and host gene expression: A review. *Human genetics, 140*(5), 747-760.

26. Wittkopp, P. J., & Kalay, G. (2012). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics, 13*(1), 59-69.

27. Copeland, M. F., Politz, M. C., & Pfleger, B. F. (2014). Application of TALEs, CRISPR/Cas and sRNAs as trans-acting regulators in prokaryotes. *Current opinion in biotechnology*, 29, 46-54.

28. Javdan, B., Lopez, J. G., Chankhamjon, P., Lee, Y. C. J., Hull, R., Wu, Q., ... & Donia, M. S. (2020). Personalized mapping of drug metabolism by the human gut microbiome. *Cell*, *181*(7), 1661-1679.

29. Dutta, D., Belashov, I. A., & Wedekind, J. E. (2018). Coupling green fluorescent protein expression with chemical modification to probe functionally relevant riboswitch conformations in live bacteria. *Biochemistry, 57*(31), 4620-4628.

30. Blanco, A., & Blanco, G. (2017). Chapter 23—Regulation of Gene Expression. In A. Blanco & G. Blanco (Eds.), *Medical Biochemistry* (pp. 525–533). Academic Press.

31. Muranaka, N., Sharma, V., Nomura, Y., & Yokobayashi, Y. (2009). An efficient platform for genetic selection and screening of gene switches in Escherichia coli. *Nucleic acids research*, *37*(5), e39-e39.

32. Gruber, A. R., Bernhart, S. H., & Lorenz, R. (2015). The ViennaRNA web services. *RNA bioinformatics*, 307-326.

33. Regulski, E. E., & Breaker, R. R. (2008). In-line probing analysis of riboswitches. *Post-transcriptional gene regulation*, 53-67.

34. Crameri, A., Whitehorn, E., Tate, E. *et al.* Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat Biotechnol* **14**, 315–319 (1996).

35. Takara Bio. (2015, November 12). *pGFPuv vector certificate of analysis*. Clontech Laboratories. https://www.takarabio.com/documents/Certificate%20of%20Analysis/632312/632312-121913.pdf

36. Uchiyama, T., Abe, T., Ikemura, T. *et al.* Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol* **23**, 88–93 (2005).

37. McCown, P. J., Corbino, K. A., Stav, S., Sherlock, M. E., & Breaker, R. R. (2017). Riboswitch diversity and distribution. *RNA (New York, N.Y.)*, *23*(7), 995–1011.

38. Ioanna Kalvari, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, Claire Toffano-Nioche, Daniel Gautheret, Zasha Weinberg, Elena Rivas, Sean R Eddy, Robert D Finn, Alex Bateman, Anton I

Petrov, Rfam 14: Expanded coverage of metagenomic, viral and microRNA families, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D192–D200.

39. Feilmeier BJ, Iseminger G, Schroeder D, Webber H, Phillips GJ. (2000). Green fluorescent protein functions as a reporter for protein localization in Escherichia coli. *J Bacteriol 182*(14):4068-76.

40. Thavarajah, W., Silverman, A. D., Verosloff, M. S., Kelley-Loughnane, N., Jewett, M. C., & Lucks, J. B. (2019). Point-of-use detection of environmental fluoride via a cell-free riboswitch-based biosensor. *ACS synthetic biology, 9*(1), 10-18.

41. Kavita, K., & Breaker, R. R. (2023). Discovering riboswitches: The past and the future. *Trends in biochemical sciences, 48*(2), 119-141.

42. Havill, J. T., Bhatiya, C., Johnson, S. M., Sheets, J. D., & Thompson, J. S. (2014). A new approach for detecting riboswitches in DNA sequences. *Bioinformatics (Oxford, England)*, *30*(21), 3012–3019.

43. Page, K., Shaffer, J., Lin, S., Zhang, M., & Liu, J. M. (2018). Engineering riboswitches in vivo using dual genetic selection and fluorescence-activated cell sorting. *ACS synthetic biology*, *7*(9), 2000-2006.

44. *Microbiome metabolite library of standards supplied by IROA Technologies*. Sigma Aldrich USA. (n.d.). https://www.sigmaaldrich.com/US/en/product/sigma/gutmls

45. Ursell, L. K., Clemente, J. C., Rideout, J. R., Gevers, D., Caporaso, J. G., & Knight, R. (2012). The interpersonal and intrapersonal diversity of human-associated microbiota in key body sites. *The Journal of allergy and clinical immunology*, *129*(5), 1204–1208.

46. Baker, J. L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R. B., & Breaker, R. R. (2012). Widespread genetic switches and toxicity resistance proteins for fluoride. *Science*, *335*(6065), 233-235.

47. White, N., Sadeeshkumar, H., Sun, A., Sudarsan, N., & Breaker, R. R. (2022). Lithium-sensing riboswitch classes regulate expression of bacterial cation transporter genes. *Scientific Reports*, *12*(1), 19145.

48. White, N., Sadeeshkumar, H., Sun, A., Sudarsan, N., & Breaker, R. R. (2022). Na+ riboswitches regulate genes for diverse physiological processes in bacteria. *Nature chemical biology*, *18*(8), 878-885.

49. Zimmermann, B., Bilusic, I., Lorenz, C., & Schroeder, R. (2010). Genomic SELEX: a discovery tool for genomic aptamers. *Methods*, *52*(2), 125-132.