

© 2022 Ryan M. Wyllie

STRATEGIES FOR THE GENETIC CHARACTERIZATION
OF THE ORAL STREPTOCOCCI

BY

RYAN M. WYLLIE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Bioengineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Assistant Professor Paul A. Jensen, Chair
Associate Professor Pablo Perez-Pinera
Assistant Professor Shannon Sirk
Professor Michael Miller

Abstract

The oral streptococci make up a diverse community of Gram-positive bacteria that inhabit the human oral cavity. Members of the oral streptococci include dental pathogens, like *Streptococcus mutans* and *Streptococcus sobrinus*, as well as commensal bacteria, like *Streptococcus salivarius*, which produce a broad range of antimicrobial compounds effective against human pathogens. However, the complex genetic circuits of these microbes, which govern metabolism, virulence, community-level communication, and stress responses, remain poorly understood. Central to this knowledge deficit is the lack of tools available for the genetic characterization of the oral streptococci.

In this dissertation, I develop synthetic biology tools and experimental methods that can be used to characterize genetic systems in the oral streptococci by modulating genetic inputs and measuring transcriptomic outputs. Using a novel, Golden Gate-compatible gene expression chassis designed specifically for the oral streptococci, I assembled thousands of combinatorial gene expression cassettes derived from an initial collection of 6,646 promoters, 22 ribosome binding sites (RBS), and 6 open reading frames (ORFs); and characterized their transcriptional and translational strengths. From these data, I provide a set of 12 characterized promoters for controlling transcription and 15 characterized promoter-RBS combinations for fine-tuned control of protein expression in *S. mutans*.

In the interest of addressing the high labor and monetary costs of current prokaryotic transcriptomic library preps, I describe Splintlock-seq, a novel library prep technology based on the activity of SplintR ligase on padlock probes targeting 95% of the entire annotated transcriptome for *S. mutans* UA159. I demonstrate how top-down, automated padlock probe design can be combined with next-generation sequencing assays to achieve a characterized probe pool capable of identifying differentially-expressed genes in *Streptococcus mutans*. Splintlock-seq is shown to produce comparable transcriptomic profiling data to RNA-seq, but for a fraction of the overall cost.

Acknowledgments

This work would not have been possible without the support of many mentors, colleagues, friends and family. I am grateful to:

- My advisor Paul A. Jensen for his mentorship, patience, support and friendship. I would not have pursued a PhD if it wasn't for his unique vision of what scientific research and mentorship should be. The lab culture that he has carefully cultivated is one which encourages curiosity, creativity, and above all learning. This was imperative to any scientific success which I have achieved.
- Members of the Jensen lab for their camaraderie, insights, and support. I am especially grateful to Kenan Jijakli, Walden Li, Adam Dama, Ben David, Noah Schmid, and Kevin Kim for their fellowship both inside and outside of lab.
- MANTIS for handling an enormous amount of liquid handling work, all without a single complaint.
- My dissertation committee—Pablo Perez-Pinera, Shannon Sirk, and Michael Miller—for their advice, time, flexibility, and patience. Did I mention patience?
- Ramdane Harouaka for his valuable insight during the development of Splintlock-seq
- Philip D. Laible and Deborah K. Hanson for their mentorship, patience, rigorous instruction in molecular biology techniques, and for introducing me to the world of scientific research.
- Glennys Mensing for helping a struggling, first-year graduate student completely out of his element. Cleanrooms can be scary places for the uninitiated.
- Alvaro Hernandez, Chris Wright, and Leslie Benson for their incredible insights and talents in all things pertaining to next-generation sequencing.
- Members of the facilities and administrative teams at the Department of Bioengineering, especially John Hardin, Mike Hansen, Krista Smith, Liezl Bowman, and Stacy Fournier.
- All past and present teachers and professors for their diligent and thoughtful instruction. I am especially grateful for Brian Estabrook, Cheryl Dispensa, Kevin Gannon, Jordan Zimmeroff, Paul Cheek, Princess Imoukhuede, Sua Myong, Marcia Pool, Brad Sutton, and Ralph Mathisen.

- All my friends in Champaign and Chicago for their encouragement, support, and fellowship.
- Alyssa Rose Collins for her love, support, and wonderful ability to restore perspective. She has been there for me through the rollercoaster of successes and failures that inevitably accompany scientific research. After failed experiments or a long night in the lab, she never failed to brighten my mood. I will always be thankful for that.
- My grandparents, Ray and LaVerne Stella; and Lawrence and Marilyn Wyllie for their never-ending support and status as role models for integrity, empathy, and humility.
- My parents, Sherri and Michael Wyllie for everything.

Table of contents

List of Figures	vii
List of Tables	ix
Chapter 1 Introduction	1
Part I Manipulating gene expression	3
Chapter 2 Curating genetic parts for oral streptococci	4
2.1 Overview	4
2.2 RACCET	4
2.3 pRW17: A modular gene expression platform in streptococci	6
2.4 Literature Mining for Gram-Positive Synthetic Biology Parts	9
Chapter 3 Transcriptional characterization of promoter libraries	16
3.1 Overview	16
3.2 Promoter Pool Design and Synthesis	17
3.3 Ribosome Binding Site Design and Synthesis	17
3.4 ORF Design and Synthesis	18
3.5 Chromosome Insertion Homology Arm Design and Synthesis	18
3.6 RBS-ORF Library Construction	20
3.7 Combinatorial Gene Expression Cassette Library Assembly	20
3.8 Transformation	21
3.9 Transcriptional Characterization Strain Library Construction	21
3.10 Genomic DNA and Total RNA Extraction	21
3.11 Preparation and Sequencing of Transcriptional Characterization NGS Libraries	22
3.12 Paired-end NGS Data Processing	23
3.13 Single-read NGS Data Processing	25
3.14 Transcriptional Strength Quantification	26
3.15 ANOVA for Transcriptional Characterization Data	26
3.16 Effect of Mutation Position on Transcriptional Strength	28
3.17 Selection of Promoters for qPCR Validation	28

3.18	qPCR Validation	30
3.19	Selection of Final Characterized Promoter Set	32
Chapter 4	Translational characterization of combinatorial gene expression cassettes	34
4.1	Overview	34
4.2	RMW6 Construction	35
4.3	High-throughput, Combinatorial Strain Construction	37
4.4	Measuring protein translation <i>in vivo</i>	40
4.5	Translation Strength Analysis	42
4.6	Selection of Final RACCET Promoter-RBS Pairs	44
4.7	Effect Contributions for Final RACCET Promoter/RBS Pairs	45
Part II	Measuring gene expression	48
Chapter 5	Splintlock-seq: Scalable and targeted gene expression profiling	49
5.1	Overview	49
5.2	Automated design of padlock probe pools	52
5.3	<i>S. mutans</i> UA159 Padlock Probe Pool v.1 Summary	55
5.4	Padlock Probe Pool Synthesis	56
5.5	Identification of highly non-specific probes in v.1 probe library	56
5.6	Splintlock-seq stoichiometry screen	59
5.7	Padlock probe target specificity assay	62
5.8	Padlock probe fold change quantification assay	65
5.9	Padlock probe library v.3 probe selection	69
5.10	Assessment of v.3 library rRNA orthogonality	70
5.11	Comparing Splintlock-seq and RNA-seq	71
Chapter 6	Conclusions	76
6.1	Summary	76
6.2	Future Directions	77
Bibliography		79
Appendix A	Probability calculations for padlock probe design	86

List of Figures

2.1	Overview of RACCET process	5
2.2	Plasmid diagram of pRW17	7
2.3	The plasmid pRW17 can be used to integrate gene expression cassettes into the chromosome of oral streptococci. A. Homology arms are ligated to the pRW17 gene expression cassette by Golden Gate Assembly. B. The homology arms allow double-crossover recombination <i>in vivo</i> to insert the expression cassette.	8
2.4	Sequence logo for <i>S. mutans</i> bacteriophage self-replication promoters	10
3.1	Overview of RACCET transcriptional characterization assay	17
3.2	Results of transcriptional characterization assay for <i>S. mutans</i>	27
3.3	ANOVA results for expression element contributions during transcription	27
3.4	Effect of promoter mutation position on transcriptional strength	29
3.5	qPCR validation of RACCET transcriptional characterization assay results	32
4.1	Overview of RACCET translational characterization assay	36
4.2	Results of RACCET translational characterization assay	43
4.3	ANOVA results for expression element contribution during translation	45
4.4	Expression element contributions in RACCET translational control toolkit	47
5.1	Padlock probe architecture	50
5.2	Overview of Splintlock-seq library prep	51
5.3	Costs of different bacterial transcriptomic library preps	52
5.4	Overview of automated padlock probe design pipeline	53
5.5	Initial non-specific probe identification results	59
5.6	Splintlock-seq library prep stoichiometry screen	63
5.7	Design and results for the probe target specificity assay	65
5.8	Experimental design for the probe fold change quantification assay	66
5.9	Representative results of padlock probe fold change validation assay	68
5.10	rRNA orthogonality of Splintlock-seq probes	71
5.11	Fold change estimates for transcripts profiled using RNA-seq and Splintlock-seq. According to both technologies, most transcripts have no fold change between growth on glucose and growth on fructose. The observed range of fold changes for Splintlock-seq is smaller than the range observed with RNAseq.	73

5.12	Volcano plots for Splintlock-seq and RNA-seq	75
A.1	Comparison of filtering results predicted using the derived probability equation to empirical results	88

List of Tables

2.1	Promoters mined from Gram-positive gene expression literature	10
2.2	Ribosome binding sites mined from Gram-positive gene expression literature	12
2.3	Reporter ORFs selected for use in protein expression assays	15
3.1	RACCET transcriptional control toolkit for <i>S. mutans</i>	33
4.1	RACCET toolkit for translational control	46

Chapter 1

Introduction

Gene expression occupies a central role in both the study and engineering of biological systems. Researchers measure gene expression to understand the physiological state of a cell, and the genomics revolution has made gene expression one of the most common proxies for studying biology. Scientists manipulate the expression level of individual genes to create the perturbations necessary to understand complex systems. Gene deletion and over-expression studies are fundamental tools in microbiology. Genetic perturbations can also modify organisms to achieve design goals, and synthetic biology offers engineers numerous options for realizing their designs *in vivo*.

We take for granted how easy it is to measure and manipulate gene expression. Next-generation sequencing technologies like RNA-seq have made genome-wide expression profiling routine. DNA registries allow engineers to pick and choose genetic parts to suit their goals. However, expectations rise as technologies improve. Every corner of microbiology uses RNA-seq, and its spread is limited by practical constraints (time & costs) rather than a lack of scientific interest. The cost of DNA sequencing has dropped exponentially, but the cost of RNA-seq has stagnated for more than a decade due to expensive and time-consuming library preparation protocols. Engineers demand finer control over gene circuits built for new, poorly-characterized species. Genetic toolboxes—while plentiful for *E. coli*—are crude and sparse for the vast majority of bacteria.

This thesis builds better tools for manipulating and measuring gene expression. Part I presents a pipeline for the **R**apid **A**ssembly and **C**haracterization of a **C**ombinatorial **E**xpression **T**oolbox (RACCET). Two advantages of RACCET are

1. RACCET focuses on characterizing poorly understood parts in non-model organisms. We use RACCET to build a complete gene expression toolbox for the Gram-positive human bacterium *S. mutans*—a species with an oversized impact on human health and an undersized set of synthetic biology tools. All of the techniques developed for RACCET can be expanded to other bacteria.
2. RACCET measures the interactions between individual genetic parts. We show that interactions can dominate when creating cassettes with very high or very low expression.

Part II of the thesis develops low-cost, targeted methods for measuring gene expression. We introduce Splintlock-seq, a probe-based method for profiling a subset of genes. We present high-throughput techniques for building and characterizing gene expression probes that combine bioinformatic screening with combinatorial sequencing assays. Although Splintlock-seq is designed to complement RNA-seq through low-cost profiling of a subset of genes, we show that our method scales to genome-wide screens for differentially expressed genes.

Taken together, the thesis provides efficient tools for manipulating the inputs to gene expression and measuring the outputs. We believe these tools will be a boon for synthetic biologists and microbiologists working with the oral streptococci or other understudied bacteria.

Part I

Manipulating gene expression

Chapter 2

Curating genetic parts for oral streptococci

2.1 Overview

The genetic toolbox for *S. mutans* is small compared to *E. coli*, *B. subtilis*, and other model species. My first goal was to expand the list of genetic parts for the streptococci and allow engineers to mix-and-match when creating gene expression cassettes. The parts described in this chapter will be characterized subsequently on the transcriptional (Chapter 3) and translational (Chapter 4) levels using RACCET, which we summarize in Section 2.2.

In Section 2.3 I present pRW17, a Gram-positive plasmid that was designed for modular assembly of gene expression cassettes using Golden Gate Assembly. The flexibility of pRW17 makes it the platform of choice for this thesis and all other genetic studies in our lab. All other parts in this thesis are designed for compatibility with pRW17 and Golden Gate Assembly.

Section 2.4 describes how I mined genetic parts from the literature or designed novel parts using computational tools. Gene expression studies pertaining to Gram-positive bacteria were mined in order to identify candidate gene expression elements. This resulted in a collection of 17 promoters (§2.4.1), 22 RBSs (§2.4.2), and seven reporter ORFs (§2.4.3) for use in the oral streptococci.

2.2 RACCET

Gene expression levels can be tuned and modulated through the use of well-characterized gene expression elements, specifically promoters and RBSs. However, the transcriptional or translational strength of a particular gene expression element is highly dependent on its biological context. Variables such as growth rate [1], sequence context [2], and the identity of the expression host are widely acknowledged to affect gene expression [3]. As such, the large sets of characterized promoters and RBSs available for use in model organisms, like *E. coli* and *B. subtilis*, are of little predictive use in less-studied bacteria. This highlights the need for species-level characterization,

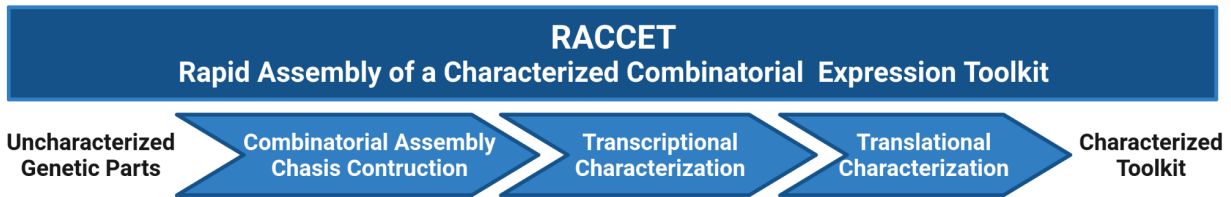


Figure 2.1: RACCET uses high-throughput assays to measure how promoters, RBSs, ORFs, and their interactions affect gene expression.

or re-characterization, of gene expression elements. However, the process of characterizing sets of promoters and/or RBSs for a given context using traditional methods is laborious, low-throughput, and usually convolves transcriptional and translational characterization. RACCET (Figure 2.1) integrates a modular gene expression platform into a high-throughput pipeline that will allow for the creation of a characterized set of promoter-RBS combinations offering control over a wide range of gene expression in *S. mutans*.

Traditional strategies for the characterization of gene expression elements rely on the quantification of bioluminescence, fluorescence, or colorimetric change that serves as a proxy variable for reporter protein abundance. As a result, the transcriptional and translational strengths of the promoter-RBS combination are convolved into a single readout, preventing element-specific characterization. Traditionally, these assays are low-throughput as they are limited by the need to individually clone each gene expression cassette prior to characterization. While recent studies have utilized fluorescence-activated cell sorting to improve throughput, they sacrifice characterization resolution, with gene expression bins being the fundamental characterized unit instead of individual cassettes. Finally, traditional reporter assays are typically subject to gene expression noise resulting from extrinsic factors within the cell, such as nutrient availability [4]. This further complicates the intrinsic characterization of promoters and RBSs. We decouple transcriptional characterization from translational characterization. In doing so, we increase the throughput of the entire process by first utilizing a highly multiplexed, molecular barcode counting assay for promoter characterization. This approach is compatible with pooled inputs which can be rapidly generated using RACCET’s modular expression system. After determining the transcriptional strengths of the combinatorial gene expression cassettes, we use a fluorescent reporter assay modified to control for noise derived from extrinsic factors affecting gene expression in order to characterize a collection of promoter-RBS pairs. Combining the information from these two processes allows us to build a characterized collection of gene expression elements for use in *S. mutans*. Although RACCET separately characterizes transcription (by NGS) and translation (by fluorescence reporters), both steps perturb all three elements that control gene expression: promoters, RBSs, and ORFs. These perturbations can be used to deconvolve the relative contribution of each component (§3.15 and §4.5.3) and identify interactions between the elements (§4.7).

2.3 pRW17: A modular gene expression platform in streptococci

pRW17 (Figure 2.2) was designed to be a streptococcal expression vector compatible with Golden Gate assembly of modular gene expression cassettes. The backbone of pRW17 was derived from pRW06, a derivative of pZX10 [5] with the Gram-negative origin of replication from pDL278 [6] inserted into the BbvCI site. The backbone of pRW06 — containing the *aad9* spectinomycin resistance cassette, the Gram-positive origin of replication, and the Gram-negative origin of replication — was amplified using primers that added BsaI restriction sites to the 5' and 3' ends of the amplicon. A synthetic insert molecule was designed to include insulating terminators and a polylinker region. The bidirectional terminator located at the end of the *hisB* gene within the histidine biosynthesis cluster of *L. lactis subsp. lactis* [7] was placed at the 5' end of the insert to prevent transcriptional read-through from genes within the upstream replicon. In the middle of the insert, I designed a polylinker region including recognition sites for many commonly utilized restriction enzymes. Two BsaI sites were placed at the flanks of the polylinker region to allow for polylinker removal during Golden Gate assembly. To prevent transcription of cloned gene cassettes from continuing outside the insert region, I placed a predicted, unidirectional terminator associated with ORF 20 of the *S. mutans* bacteriophage M102 [8] in series with a synthetic, bidirectional *rrnB*-T7 terminator on the 3' flank of the insert region. The entire insert sequence was synthesized as a gBlock Gene Fragment by IDT and amplified with primers to add BbsI sites, which upon digestion yield overhangs compatible with the overhangs produced by BsaI digestion of the pRW06-derived amplicon. Both amplicons were digested with their respective Type IIS enzyme and utilized in a standard T4 DNA Ligase (NEB) assembly reaction using 40 fmol of digested insert and 20 fmol of digested backbone in accordance with the enzyme's recommended ligation protocol. After 1 hour, 5 μ l of the 20 μ l ligation mix was used to transform competent *E. coli* DH5 α (Mix & Go, Zymo Research). After a 1-hour outgrowth, the cells were plated on LB plates supplemented with 100 μ g/mL of spectinomycin and incubated overnight at 37 °C with shaking (250 rpm). The following morning, colonies were screened with colony PCR using primers targeting the P0 and P4 sites. Clones yielding correct amplicon sizes as determined with DNA gel electrophoresis were used to start 5 mL cultures, which were incubated overnight as described above. Plasmids were then isolated from cultured cells using a GeneJet Plasmid Miniprep Kit (Thermo), and screened by restriction enzyme digest with PstI and BsaI. Successful pRW17 construction was then validated by Sanger sequencing candidate plasmids using primers targeting the P0 and P4 sites. For use in Golden Gate cloning efforts with pRW17, a standard set of five orthogonal four-nucleotide overhangs were selected by referencing similar designs [9] and T4 DNA ligase overhang fidelity and efficiency profiles [10] : ATAG (Overhang 1)¹, TCTA (Overhang 2), AATC (Overhang 3), GACT (Overhang 4), and CATA (Overhang 5).

Importantly, components designed for assembly with pRW17 can be easily repurposed to make double crossover homologous recombination cassettes for integration into a specific locus on the chromosome of *S. mutans*, or any other recombination-savvy microbe. To do this, the user simply

¹Unless otherwise stated, all DNA sequences are written 5' to 3'.

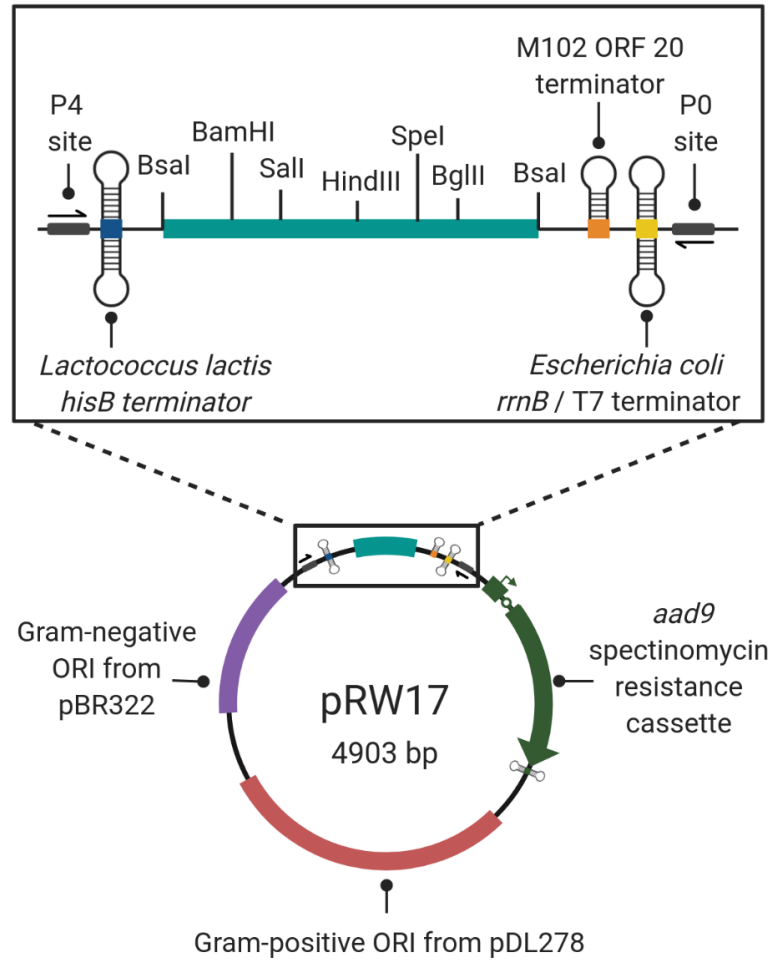


Figure 2.2: The plasmid pRW17 is an modular platform for cloning and gene expression in oral streptococci.

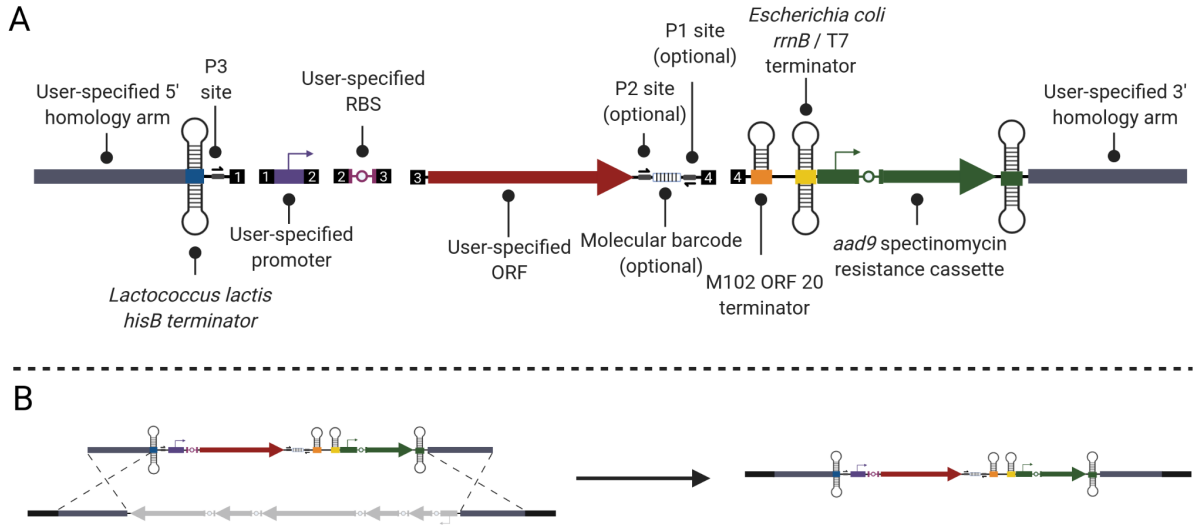


Figure 2.3: The plasmid pRW17 can be used to integrate gene expression cassettes into the chromosome of oral streptococci. **A.** Homology arms are ligated to the pRW17 gene expression cassette by Golden Gate Assembly. **B.** The homology arms allow double-crossover recombination *in vivo* to insert the expression cassette.

has to replace the pRW17 backbone with two, 1 kb amplicons containing nucleotide sequences with homology to the chromosome environment flanking the desired insertion locus and a selection marker cassette during Golden Gate Assembly. In this case, the 5' homology arm should contain a *BsaI* site capable of creating Overhang 1 at its 3' end. The selection marker cassette should contain a *BsaI* site capable of creating Overhang 4 at its 5' end and Overhang 5 at its 3' end. Similarly, the 3' homology arm should contain a *BsaI* site capable of creating Overhang 5 at its 5' end. Alternatively, if the user wants to reduce the number of parts involved in the assembly reaction while ensuring the presence of insulating terminators, the *hisB* terminator can be added to the 5' homology arm sequence and the ORF 20-*rrnB*/T7 terminator chain along with the *aad9* cassette from pRW17 can be added to the 3' homology arm using standard restriction cloning techniques prior to Golden Gate assembly of the desired gene cassette payload. An example of this approach is shown in Figure 2.3.

Used in conjunction with the standardized set of Golden Gate overhangs, over 180 pRW17-derived plasmid constructs have been created within the last 3 years. Furthermore, those plasmids have been used to successfully express gene cassettes in *S. mutans*, *S. sobrinus*, *S. gordonii*, *S. sanguinus*, *S. thermophilus*, *S. salivarius*, *S. downei*, *S. criceti*, and *S. ferus*. For members of the Jensen Lab, the pRW17 system has completely superseded pDL278 as the gene expression vector of choice for studies with the oral streptococci.

2.4 Literature Mining for Gram-Positive Synthetic Biology Parts

2.4.1 Promoters

Gene expression is the process by which a functional gene product is generated from the information stored within a gene. This can involve many sub-processes including transcription, translation, protein folding, and post-translational modification, along with all the regulatory circuits governing each step. For the purposes of modulating gene expression, researchers typically focus on transcription and translation due to availability of cis-acting DNA components which can be readily synthesized and cloned into gene expression cassettes. The transcriptional strength of a gene expression cassette is thought to be primarily governed by the promoter identity. The sequence of the promoter determines its ability to be recognized by and bind to transcriptional machinery. Promoter specificity is canonically determined by two hexameric boxes of DNA termed the -35 and -10 sites for their approximate location with regards to the transcription start site (TSS). UP elements consisting of T/A-rich tracts, and typically found 60–80 bp upstream of the TSS, also play a role in transcriptional machinery recruitment [11]. The DNA sequence from 0 to +20 of the TSS plays a significant role in promoter escape and subsequent transcriptional elongation [2]. Therefore, in this study we have designated the promoter module to encompass everything from -80 to +20 with respect to the putative TSS.

To ensure a diverse range of transcriptional strengths, 17 constitutive promoters were chosen from species with 1.) varying phylogenetic distance to the oral streptococci and 2.) varying transcriptional strength in their host organism. Three promoters were taken from *Bacillus subtilis*, one from the lactobacilli, one from the lactococci, three from the pneumococci, two from the salivarius group of streptococci, one from the mitis group of streptococci, and two from the mutans group of streptococci. I also included four synthetic promoters.

A crucial part of building a gene expression toolkit is achieving a wide range of transcriptional/translational strengths. To provide a baseline reference for transcriptional activity, a 100 nt randomizer free of any discernable promoter elements was selected to serve as a negative control promoter sequence. In an attempt to reach the upper limits of transcriptional strength, a promoter was designed using the consensus sequence of promoters identified within the self-replication module of three *S. mutans* serotype C lytic bacteriophages. The entire genomes of the three bacteriophages — M102, M102AD, and ϕ APCM01— were submitted to a deep learning neural network, CNNpromoter, trained to identify bacterial promoter sequences using experimentally validated data from *B. subtilis* [12]. The promoters returned from this tool were then mapped to genomic coordinates. All of the referenced *S. mutans* bacteriophages contain a genomic architecture similar to that of well-studied *S. thermophilus* lytic bacteriophages, whereby the genome is organized into four modules: packaging and phage head structure; phage tail structure; host cell lysis; and self-replication [8]. I hypothesized that the self-replication module of these lytic phages would be constitutively expressed at high levels and would be orthogonal to host efforts to regulate transcription via repressor binding. Importantly, homology searches for all ORFs within the genomes of M102, M102AD, and ϕ APCM01 revealed

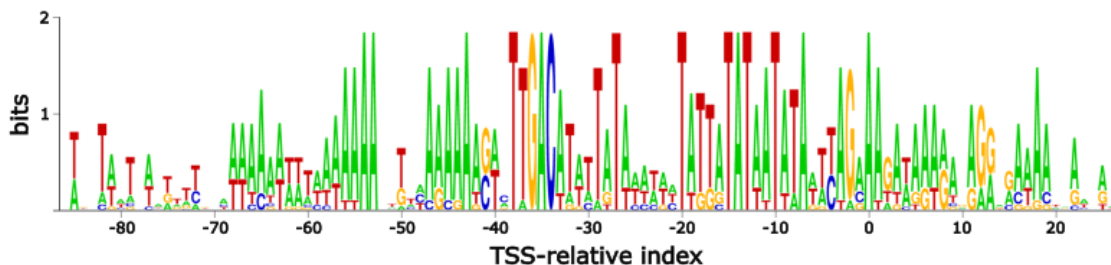


Figure 2.4: Sequence logo of 14 promoters from the self-replication modules of three *S. mutans* bacteriophages. The consensus sequence was used to design a promoter for subsequent transcriptional characterization with *S. mutans*. High levels of sequence agreement are noted at the transcriptional start site, the -10 site, the -35 site, and UP elements.

an absence of transcription-related proteins. This seems to indicate that the phages rely on host transcriptional machinery to recognize their promoter sequences and initiate transcription. To reduce the sequence information of the 14 CNNpromoter-identified promoters within the self-replication modules of the three bacteriophages, I calculated the consensus sequence using nucleotide information from -80 to $+20$ with respect to the CNNpromoter-predicted transcriptional start sites. The sequence logo for the multiple alignment of the 14 self-replication module promoters is shown in Figure 2.4.

Interestingly, a strong consensus sequence emerged with readily identifiable promoter motifs. An adenine was always observed at the predicted transcriptional start site. Particularly strong agreement was also observed from the -7 site through the -20 site, and from the -27 site through the -40 site. Three separate polyadenine tracts were also observed upstream of the canonical -35 box, hinting at the presence of UP elements within the promoter structure. Interestingly, the entire promoter consensus sequence had a GC content of only 13%. All of the promoters in the final toolbox are listed in Table 2.1.

Table 2.1: Promoter sequences were collected from the literature or developed for this study. Promoters were selected to span a wide range of transcriptional strengths in Gram positive bacteria.

ID	Name	Origin	Reference
PR1	P ₂₃	synthetic / <i>Lactococcus lactis</i> subsp. <i>lactis</i> Derivative of a synthetic promoter derived by combining the consensus -35 and -10 sequences of several <i>Lactococcus lactis</i> subsp. <i>lactis</i> promoters with random spacer sequences.	[13]
PR2	P _{veg}	<i>Bacillus subtilis</i> This promoter controls transcription of the <i>veg</i> gene, a small conserved protein that is transcribed at very high levels in growth phase and sporulation in <i>B. subtilis</i> .	[13]
PR3	P _{ldh}	<i>Streptococcus mutans</i> Ldh is a lactate dehydrogenase, an enzyme crucial to streptococcal metabolism.	[5]

Table 2.1 (continued)

PR4	P _{phage}	synthetic The consensus sequence of fourteen replication module promoters from <i>S. mutans</i> bacteriophages M102, M102AD, and Φ APCM01.	this study
PR5	P _{gtfB}	<i>Streptococcus mutans</i> GtfB is a β -glucosyltransferase that is involved in cell adhesion and biofilm formation.	[13]
PR6	P _{malM}	<i>Streptococcus pneumoniae</i> MalM is an enzyme involved in maltosaccharide utilization. The MalR recognition site was removed to prevent possible repression from MalR homologs in <i>S. mutans</i> .	[14]
PR7	P _{rrnO1}	<i>Bacillus subtilis</i> PrnO1 is upstream of the ribosomal RNA genes and demonstrates high levels of transcription during growth phase in <i>B. subtilis</i> .	[15]
PR8	P _{hlbA}	<i>Lactobacillus delbrueckii subsp. bulgaricus</i> HlbA is a histone-like DNA-binding protein. The T at position 96 in original sequence was changed to an A to prevent out of frame translation from potential RBS sequences in promoter.	[16]
PR9	P ₃	synthetic Synthetic promoter derived by combining the consensus -35 and -10 sequences of several <i>Lactococcus lactis subsp. lactis</i> promoters with random spacer sequences.	[17]
PR10	P _{ung}	<i>Streptococcus pneumoniae</i> Ung is a uracil-DNA glycosylase responsible for DNA repair.	[18]
PR11	P _{spaC}	<i>Bacillus subtilis</i> This sequence was derived from pIB167 with the <i>lac</i> repressor site removed.	[13]
PR12	P _{tRNAGlu}	<i>Streptococcus pneumoniae</i> This promoter controls transcription of a glutamic acid tRNA.	[19]
PR13	P ₁	synthetic A synthetic promoter derived by combining the consensus -35 and -10 sequences of several <i>Lactococcus lactis subsp. lactis</i> promoters with random spacer sequences.	[17]
PR14	P _{hppA}	<i>Streptococcus gordonii</i> HppA is an oligopeptide binding and transport protein associated with cell surface properties in <i>S. gordonii</i> .	[20]
PR15	P ₈	<i>Streptococcus thermophilus</i> P _{8w} as isolated from <i>S. thermophilus</i> chromosome fragments. It is found upstream of the adapter protein-encoding gene, <i>mecA</i> .	[21]

Table 2.1 (continued)

PR16	P ₂₅	<i>Streptococcus thermophilus</i> P _{25w} as isolated from <i>S. thermophilus</i> chromosome fragments. It is found upstream of STER_0676, a predicted Cof-type HAD-IIB family hydrolase.	[21]
PR17	P _{negcontrol}	synthetic A random 100-mer to serve as a negative control promoter.	this study

2.4.2 Ribosome Binding Sites

The translational strength of a gene cassette is commonly modulated through the choice of RBS. Translation occurs in two stages: ribosomal recruitment and translational initiation. The RBS influences both of these stages. Ribosomal recruitment efficacy is largely a function of the complementarity of the RBS sequence to the 3' end of the 16S rRNA. RBSs with a high similarity to the Shine-Delgarno sequence generally exhibit higher levels of ribosomal recruitment and translational initiation, although in a non-monotonic manner [22], [23]. The spacing distance between the RBS sequence and the start codon, as well as the nucleotide identity of the spacing sequence, affect rates of translation initiation, with stable secondary structures being particularly deleterious [24], [25]. Most ORF-independent, prokaryotic ribosome-RBS interactions are limited to the sequence 20 base pairs upstream of the start codon. For this reason, the RBS module is by default bounded by this limit.

I collected nine native RBSs and thirteen synthetic RBSs designed to offer a wide range of translational strength in *S. mutans*. To provide a baseline reference for translational activity and identify translation-initiating sequences within the promoter, a negative control RBS was constructed by taking a 22-nt random sequence and converting all purines to pyrimidines. This should greatly reduce the theoretical affinity of the RBS region within the mRNA to the 3' end of the ribosome-associated 16S rRNA. The twelve other synthetic RBS sequences were designed for *S. mutans* using the Salis RBS Calculator to provide a wide range of theoretical translation initiation rates. The native RBS sequences include eight sequences from promoters also included in the gene expression element collection and a single RBS derived from ORF 23 in the *S. mutans* lytic bacteriophage M102. The final set of RBSs are listed in Table 2.2.

Table 2.2: RBS's were collected from the literature or an online RBS design tool [22]. RBS's from literature were drawn from proteins the span a wide range of expression levels in Gram positive bacteria.

ID	Name	Origin	Reference
RBS1	SMU_SLRBS	synthetic Derived from expression vectors in the referenced study.	[13]

Table 2.2 (continued)

RBS2	<i>veg</i>	<i>Bacillus subtilis</i> This promoter controls transcription of Veg, a small conserved protein.	[26]
RBS3	<i>ldh</i>	<i>Streptococcus mutans</i> Ldh is a lactate dehydrogenase, an enzyme crucial to streptococcal metabolism.	[27]
RBS4	ORF_23	<i>Streptococcus mutans</i> bacteriophage M102 ORF 23 encodes a DnaC-like protein within the replication module of the bacteriophage M102.	[8]
RBS5	<i>gtfB</i>	<i>Streptococcus mutans</i> GtfB is a β -glucosyltransferase that is involved in cell adhesion and biofilm formation.	[27]
RBS6	<i>malM</i>	<i>Streptococcus pneumoniae</i> MalM is an enzyme involved in maltosaccharide utilization.	[28]
RBS7	<i>hlbA</i>	<i>Lactobacillus delbrueckii subsp. bulgaricus</i> HlbA is a histone-like DNA-binding protein.	[29]
RBS8	<i>ung</i>	<i>Streptococcus pneumoniae</i> Ung is a uracil-DNA glycosylase responsible for DNA repair.	[28]
RBS9	<i>hppA</i>	<i>Streptococcus gordonii</i> HppA is an oligopeptide binding and transport protein associated with cell surface properties in <i>S. gordonii</i> .	[30]
RBS10	NC_RBS	synthetic This sequence is a random 22-mer with no purines. It should not be able to base-pair efficiently with the ribosome-associated 16s rRNA. This sequence serves as a negative control.	this study
RBS11	DeNovo_Max	synthetic This sequence was predicted to achieve very high levels of translation.	[22]
RBS12	DeNovo_100000	synthetic This sequence was predicted to achieve very high levels of translation.	[22]
RBS13	DeNovo_10000	synthetic This sequence was predicted to achieve high levels of translation.	[22]
RBS14	DeNovo_1000	synthetic This sequence was predicted to achieve medium levels of translation.	[22]

Table 2.2 (continued)

RBS15	DeNovo_100	synthetic This sequence was predicted to achieve medium levels of translation.	[22]
RBS16	DeNovo_10	synthetic This sequence was predicted to achieve low levels of translation.	[22]
RBS17	DeNovo_1	synthetic This sequence was predicted to achieve very low levels of translation.	[22]
RBS18	DeNovo_50000	synthetic This sequence was predicted to achieve very high levels of translation.	[22]
RBS19	DeNovo_5000	synthetic This sequence was predicted to achieve high levels of translation.	[22]
RBS20	DeNovo_500	synthetic This sequence was predicted to achieve medium levels of translation.	[22]
RBS21	DeNovo_50	synthetic This sequence was predicted to achieve low-medium levels of translation.	[22]
RBS22	DeNovo_5	synthetic This sequence was predicted to achieve low levels of translation.	[22]

2.4.3 Fluorescent Reporter Proteins

For the purposes of translational reporter assays, six fluorescent protein ORFs and a chloramphenicol acetyltransferase gene (*cat*) were also collected. In order to ensure experimental compatibility, all fluorescent reporters were chosen to have a similar maturation time, oligomerization state, quantum-yield, brightness, and pK_a . To maximize spectral distinction and allow for potential multiplexing, reporters were chosen from along the entire range of the visible spectrum. All ORFs mined for use with the oral streptococci are listed in 2.3.

The fluorescent protein database [38] reports excitation and emission wavelengths for each ORF. However, a spectral scan in our plate reader found that the proteins in *S. mutans* responded best at slightly different wavelengths. The empirical wavelengths are listed in Table 2.3 and were used in the rest of our experiments.

Table 2.3: Five fluorescent proteins were selected to minimize the spectral overlap when used in pairs in streptococci. Empirical excitation and emission wavelengths maximize the signal-to-noise ratio during a wavelength scan in a monochromator-equipped plate reader. The non-fluorescent Cat protein was included as a negative control.

	mScarlet-I	mNeonGreen	mTFP1	sfgfp	mTurquoise2	mTagBFP2	Cat
Origin	synthetic	<i>Branchiostoma lanceolatum</i>	<i>Clavularia</i> sp.	<i>Aequorea victoria</i>	<i>Aequorea victoria</i>	<i>Entacmaea quadricolor</i>	<i>Streptococcus</i> mobile genetic element
Molecular Weight (kDa)	26.4	26.6	26.9	26.8	26.9	26.7	25
Max Excitation Wavelength (nm)	569	506	462	485	434	399	NA
Max Emission Wavelength (nm)	593	517	492	510	474	454	NA
Stokes Shift (nm)	24	11	30	25	40	55	NA
Empirical Max Excitation Wavelength (nm)	564	505	452	485	434	400	NA
Empirical Max Emission Wavelength (nm)	602	536	492	528	472	545	NA
Extinction Coefficient ($M^{-1} cm^{-1}$)	104000	116000	64000	83300	30000	50600	NA
Quantum Yield	0.54	0.8	0.85	0.65	0.93	0.64	NA
Brightness	56.16	92.8	54.4	54.15	27.9	32.38	NA
pKa	5.4	5.7	4.3	NA	3.1	2.7	NA
Maturation Time (minutes)	36	10	NA	13.6	33.5	12	NA
Oligomerization	Monomer	Monomer	Monomer	Weak dimer	Monomer	Monomer	Monomer
Reference	[31]	[32]	[33]	[34]	[35]	[36]	[37]

Chapter 3

Transcriptional characterization of promoter libraries

3.1 Overview

During the first decade of the 21st century, the advent of next generation sequencing (NGS) technologies enabled the accurate quantification of nucleic acid molecules with extremely high throughput. Sequencing-based counting assays have allowed for the determination of DNA and RNA abundances at the genomic and transcriptomic levels, respectively. As the NGS revolution progressed, molecular barcodes became increasingly popular as a tool to improve the multiplexing capacity of sequencing runs. Molecular barcodes are random or semi-random nucleic acids, typically DNA, which can be attached to other target nucleic acids through PCR or ligation in order to label them for experimental purposes. In the context of NGS, molecular barcodes are often added to label individual sequencing libraries prior to pooling and running in a single sequencing lane. During low-input transcriptomics experiments, such as scRNA-seq, molecular barcodes (often referred to as unique molecular identifiers in this context) are used to label individual cDNA molecules in order to preserve relative transcript abundance within the population prior to bias-prone amplification with PCR [39]. In 2010, researchers demonstrated that the relative abundance of 6,200 uniquely barcoded and pooled yeast mutants could be accurately quantified by amplifying and sequencing the barcodes from gDNA libraries [40]. The targeted sequencing of DNA barcodes is now widely applied to quantify population composition, being central to techniques such as Tn-seq which aim to assess fitness changes associated with transposon-mediated gene inactivation [41]. In this chapter, we use targeted barcode sequencing in order to quantify the transcriptional strength of combinatorial gene expression cassettes each labelled with a unique molecular barcode.

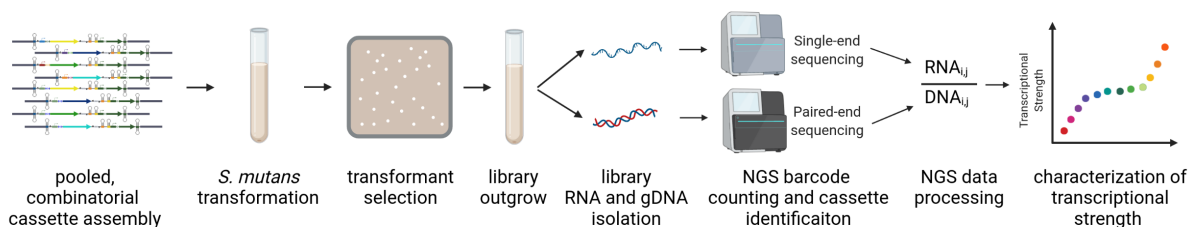


Figure 3.1: RACCET uses DNA barcodes to quantify transcription levels of expression cassettes. Pools of cassettes, each with a unique combination of promoter, RBS, and ORF, are transformed into *S. mutans*. Libraries of transformed cells are grown and DNA and RNA are harvested from the same culture. The abundance of cDNA barcodes made from the RNA is proportional to the number of transcripts in the pool. These counts are normalized by the DNA barcode counts to correct for the number of cells expressing the cassette. A separate paired-end NGS run reveals the promoter, RBS, and ORF corresponding to each barcode.

3.2 Promoter Pool Design and Synthesis

In order to increase the diversity of promoter sequences in the transcriptional characterization assay, and in doing so hopefully increase the range of transcriptional strengths observed, the 17 promoter sequences mined from relevant Gram-positive gene expression studies were mutated *in silico* to produce every single-position substitution and deletion mutant for each “base promoter” sequence. To do this, nucleotide sequences for the 17 base promoters were read into a custom R script. The script iterates through the input list and returns every possible single position substitution and deletion mutant for a given sequence. The resulting set of sequences are then screened to ensure that no mutation had resulted in the creation of a BsaI recognition site within the promoter sequence. A common amplification tag (ACTACAACAGACCTTGCTCC) as well as the Overhang 1 sequence for Golden Gate cloning (GGTCTCGATAG) was appended to the 5’ end of the promoter sequence. Likewise, the Overhang 2 sequence for Golden Gate cloning (TCTAGGGGAGACC) and a different common amplification tag (TTGTAGATCCAGGGTGGTGA) were added to the 3’ end of each promoter sequence. The resulting set of 6,646 mutant promoter sequences were synthesized on a DNA microarray (CustomArray, Inc.). The ssDNA promoter pool was then PCR amplified using primers targeting the common tags at the 5’ and 3’ ends of the promoter sequences. Amplicons were purified using a Thermo GeneJet PCR Cleanup column, validated by gel electrophoresis, and stored at -20°C .

3.3 Ribosome Binding Site Design and Synthesis

For each RBS cassette, two individual, partially complementary oligonucleotides were designed. The hybridization product of these two oligonucleotides results in a 5’ ssDNA overhang sequence a 3’ ssDNA overhang sequence for Golden Gate Assembly with promoter and ORF elements. Each oligonucleotide was synthesized separately and added to a hybridization and phosphorylation

reaction. Briefly, 1 μ l of 10 μ M stocks for each oligonucleotide in a given RBS cassette was added to 41 μ l nuclease free water, 5 μ l of 10x T4 DNA Ligase Buffer (NEB), and 2 μ l of T4 PNK (NEB, 10 U/ μ l). The reaction was incubated at 37 °C for 30 minutes, heated to 95 °C for 5 minutes, and then returned to 25 °C by ramping down at a rate of -5 °C/minute. RBS cassettes were then stored at -20 °C.

3.4 ORF Design and Synthesis

Primary amino acid sequences for mNeonGreen, mTFP1, mTurquoise2, mTagBFP2, and mScarlet-I were obtained from FPbase and converted to DNA sequences using IDT’s codon optimization tool with *S. mutans* UA159 codon bias data. After ensuring that no BsaI sites were added during codon optimization, a common amplification tag (ACTACAACAGACCTTGCTCC) as well as the Overhang 3 sequence for Golden Gate Assembly (GGTCTCGAATC) were added to the 5’ end of each ORF sequence. Similarly, the Overhang 4 sequence for Golden Gate Assembly (GACTGGAGACC) and a different common amplification tag (TTGTAGATCCAGGGTGGTGA) were added to the 3’ end. The resulting sequence for each ORF was then synthesized as a gBlock Gene Fragment, and used as template for PCR amplification with primers targeting the common amplification tags. A plasmid used in streptococcal gene expression studies, pComXgfp, provided by the Burne Lab at the University of Florida was used as template for the amplification of the sfgfp gene with ORF-specific primers with non-hybridizing tails containing the Overhang 3 and Overhang 4 sequences. The *cat* gene was amplified from pCAT, a plasmid developed in-lab by former lab technician Brent Wu, using analogous primers. Amplicons were purified using $0.5\times$ AMPure SPRI beads and subcloned into *E. coli* DH5 α using the pRW17 assembly platform. All plasmid constructs were created using P23 and RBS1. Clones were then screened with colony PCR using primers targeting the P0 and P4 sites, restriction digest with PstI-HF, and Sanger sequencing with primers targeting the P0 and P4 sites. Plasmids harboring the correct ORF sequence were used as template for PCR amplification with ORF-specific primers with non-hybridizing tails. The primer targeting the 5’ end of each ORF contained a non-hybridizing tail with baggage sequence and the Overhang 3 sequence for Golden Gate cloning (GTATACGGTCTCGAATC). The ORF-specific primer targeting the 3’ end of the ORF sequence contained a non-hybridizing tail that added the P2 sequence (GATGTCCACGAGGACTCT), a degenerate molecular barcode sequence designed with CutFree [42] to harbor no endogenous BsaI sequences (NNNNNANTNNMMNNMMNNNN), the P1 sequence (CGTACGCTGCAGGAGCTC), and the Overhang 4 sequence for Golden Gate Assembly (TTGTAGATCCAGGGTGGTGA). Amplicons were purified with $0.5\times$ AMPure SPRI beads, validated by gel electrophoresis, and stored at -20 °C.

3.5 Chromosome Insertion Homology Arm Design and Synthesis

The *S. mutans* mannitol utilization operon (1121766:1127060) starting with *mtlA1* and ending with *mtlD* was selected as the insertion site for gene expression cassettes. This locus had been used in

previous gene expression studies with *S. mutans*, and insertions removing the intervening four genes were shown to have no deleterious effects on strain growth or virulence when grown planktonically in THY. The 1 kb, 5' homology arm was selected to include native genomic sequence from index 1120766 to 1121765, while the 1 kb, 3' homology arm was selected to include native genomic sequence from index 1127061 to 1128060. These homology arms ensure that integrated gene cassette's coding strand is opposite that of the replaced genes. From this point forward, this insertion locus is referred to as the *phnA-mtlA* locus. To reduce the number of participating elements and boost assembly efficiency, it was decided that the *hisB* terminator and the P3 binding site should be appended to the 5' homology arm. To do this the bidirectional *hisB* terminator and the adjacent P3 site were amplified off pRW17. The 5' primer used to do this contained a BsaI site for subsequent Golden Gate Assembly (GTATACGGTCTCGATAG). The 1 kb 5' homology arm sequence was amplified off *S. mutans* UA159 genomic DNA to create an amplicon with a 3' BsaI site for subsequent Golden Gate Assembly (ATAGGGAGACCGGATCC). This amplicon and the pRW17-derived amplicon containing the *hisB* terminator and the P3 site were then joined in a Golden Gate Assembly reaction. The Golden Gate product was then diluted and used directly as template for a PCR amplification. The forward primer in this reaction hybridized with no baggage to the 5' end of the 5' homology arm sequence, while the reverse primer targeted the P3 sequence and had a non-hybridizing tail containing sequence to add the Overhang 1 sequence for Golden Gate Assembly. The resulting amplicon, referred to as *phn.HisB.P3.HA1*, was purified with 0.5× AMPure SPRI beads, validated with gel electrophoresis, and stored at −20 °C until needed.

Under the same rationale which led to the creation of *phn.HisB.P3.HA1*, it was determined that the dual terminator-*aad9* cassette should be joined with the 3' homology arm before combinatorial gene expression cassette assembly. To do this, the sequence containing the M102 ORF terminator, the *rrnB*/T7 terminator, and the *aad9* spectinomycin-resistance cassette was amplified using pRW17 as template using primers to add a BsaI site for subsequent Golden Gate Assembly (CATAGGAGACCAGCCGT) to the 3' end of the amplicon. The 1 kb 3' homology arm was amplified off *S. mutans* UA159 genomic DNA to create an amplicon with a 5' BsaI site for subsequent Golden Gate cloning (GTATACGGTCTCGCATA). This amplicon and the pRW17-derived amplicon containing the dual terminator-*aad9* cassette were then joined in a Golden Gate Assembly reaction. The Golden Gate product was then diluted and used directly as template for a PCR amplification. The forward primer in this reaction contained a non-hybridizing tail with sequence to add the Overhang 4 sequence for Golden Gate Assembly, while the reverse primer hybridized with no baggage to the 3' end of the 3' homology arm. The resulting amplicon, referred to as *SynTerm.SpecR.mtlA.HA2*, was purified with 0.5× AMPure SPRI beads, validated with gel electrophoresis, and stored at −20 °C until needed. To ensure sequence fidelity of these homology arm constructs, they were used in a Golden Gate Assembly reaction with P23, RBS1, and mNeonGreen. The resulting construct was then used to transform *S. mutans* UA159. Transformants were selected for by plating the entire transformation reaction on fresh THY agar plates supplemented with 100 µg/mL spectinomycin. Two colonies were picked and used to individually inoculate 5 mL of THY media supplemented

with spectinomycin. After overnight incubation in a 5% CO₂ incubator at 37 °C, genomic DNA was extracted from both cultures and used as template (50 ng) in PCR amplifications utilizing primers to regenerate the *phn.HisB.P3.HA1* and *SynTerm.SpecR.mtlA.HA2* amplicons. After 30 cycles of amplification, amplicons were purified using 0.5× AMPure SPRI beads, eluted into 20 µl of nuclease free water, and validated by gel electrophoresis and Sanger sequencing. A clone with correct *phn.HisB.P3.HA1* and *SynTerm.SpecR.mtlA.HA2* amplicons, termed RMW12, was then used to create a freezer stock. Genomic DNA from RMW12 was stored at −20 °C for use in any future PCRs geared towards the creation of homology arms for insertion into the *mtlD:mtlA1* locus.

3.6 RBS-ORF Library Construction

Preliminary efforts to construct combinatorial gene expression cassette libraries through Golden Gate assembly with promoter library amplicons, RBS cassettes, and ORF amplicons were plagued by low transformation efficiency. Troubleshooting experiments revealed this to be a function of the complexity of the RBS pool, likely indicating that spontaneous dehybridization of RBS cassettes within a complex mixture prevented faithful rehybridization, ultimately preventing ligation with the promoter and ORF amplicons. To remedy this issue, it was decided that each of the 22 RBS sequences would be directly added to each of the six ORF sequences by PCR. Primers targeting the 5' end of each ORF sequence were designed with non-hybridizing tails containing baggage sequence and the Overhang 2 sequence for Golden Gate Assembly (GTATACGGTCTCGTCTA), the RBS sequence, and the Overhang 3 “scar” sequence (AATC). These primers were paired with primers targeting the 3' end of each ORF sequence with non-hybridizing tails containing the P2 sequence, the degenerate molecular barcode sequence, the P1 sequence, and the Overhang 4 sequence for Golden Gate Assembly. For each PCR amplification, sequence-verified plasmid containing the respective ORF sequence was used as template. The resulting 132 amplicons were purified with 0.5× AMPure SPRI beads, validated with gel electrophoresis, pooled in an equimolar manner, and stored at −20 °C.

3.7 Combinatorial Gene Expression Cassette Library Assembly

Six individual gene expression cassette libraries were produced using pooled combinatorial Golden Gate Assembly. For each reaction, 40 fmol of promoter library amplicon (~3.0 ng), 40 fmol of RBS-ORF library amplicon (~21 ng), 40 fmol of *phn.HisB.P3.HA1* amplicon (~27 ng), 40 fmol of *SynTerm.SpecR.mtlA.HA2* amplicon (~54 ng) and 2 µl of 10× T4 DNA Ligase Buffer (NEB) were combined in a sterile 200 µl PCR tube (USA Scientific). The volume was adjusted to 18 µl using nuclease-free water followed by the addition of 1 µl of T4 DNA Ligase (NEB, 400 U/µl) and 1 µl of BsaI-HF v2 (NEB, 20 U/µl). The reaction was mixed by pipetting and then put into a preheated ProFlex PCR system (Applied Biosystems). The reactions were incubated for 30 cycles of 5 minutes at 37 °C followed by 5 minutes at 16 °C. Reactions were subjected to a final incubation at 37 °C

for 10 minutes before being stored at -20°C until needed.

3.8 Transformation

An overnight culture of the *S. mutans* strain RMW18 was diluted 1:40 into 5 mL of fresh THY. The dilution was grown in a 5% CO_2 incubator at 37°C until an OD_{600} of 0.12 was reached (~ 4 hours post-dilution). At this point, 5 μl of the *S. mutans* competence pheromone, CSP-18 (1 $\mu\text{g}/\mu\text{l}$) was added to the culture. The culture was vortexed briefly and returned to the incubator for 20 minutes. 6×500 μl aliquots of cells were created in sterile 1.5 mL Eppendorf tubes. 10 μl of respective Golden Gate Assembly reaction was added to the cells followed by gentle vortex mixing for 5 s. The transformation reactions were then returned to the incubator for 2 hours. For each library, the entire 500 μl transformation reaction was plated onto large, pre-warmed THY plates supplemented with 100 $\mu\text{g}/\text{mL}$ spectinomycin.

3.9 Transcriptional Characterization Strain Library Construction

After approximately 36 hours in the incubator, all visible colonies were scraped from selection plates and used to inoculate 6×35 mL cultures of fresh THY supplemented with 100 $\mu\text{g}/\text{mL}$ spectinomycin. These cultures were grown in a 5% CO_2 incubator at 37°C for 4 hours. Dilutions ($6 \times 1:35$) were then made into 35 mL of fresh selection media and grown overnight. After about 14 hours, the cultures were used to make $6 \times 1:35$ dilutions into 35 mL of fresh THY supplemented with 100 $\mu\text{g}/\text{mL}$ spectinomycin media. Once these cultures had reached an OD_{600} of 0.3–0.5, they were harvested by centrifugation at $3,250 \times g$ for 10 minutes. The supernatant was discarded and pellets were resuspended in 1 mL of fresh THY and 335 μl of 60% (vol/vol) glycerol. The resulting cell solution was then mixed by pipetting and used to make 100 μl aliquots in sterile 1.5 mL Eppendorf tubes. Strain libraries were then frozen at -80°C until needed.

3.10 Genomic DNA and Total RNA Extraction

Aliquots for each of the six *S. mutans* strain libraries for transcriptional characterization were thawed on ice. A 5 μl aliquot of each cryostock was used to inoculate 10 mL of fresh THY supplemented with 100 $\mu\text{g}/\text{mL}$ spectinomycin. These cultures were grown until mid-log phase ($\text{OD}_{600} = 0.3\text{--}0.5$). At this point each culture was divided in two, with 5 mL of culture set aside for genomic DNA extraction and 5 mL for total RNA isolation. All cells were harvested by centrifugation at $3,250 \times g$ for 10 minutes. The supernatants were discarded and pellets destined for genomic DNA isolation were immediately frozen at -20°C . The remaining cell pellets were immediately used in a total RNA isolation protocol. Briefly, cell pellets were resuspended in 750 μl of Trizol reagent and transferred into a 1.5 mL O-ring microcentrifuge tube filled with roughly 500 μl of 0.1 mm glass beads. Cells were then lysed using a bead mill homogenizer (Ohaus) at 1,600 RPM for 1.5 minutes. The O-ring

tubes are then heated at 60 °C for 5 minutes and spun down at 15,000×g for 5 minutes. The supernatant is then collected and transferred into 1 mL of fresh Trizol reagent, mixed by pipetting, and incubated at room temperature for 5 minutes. At this point samples may enter the standard total RNA extraction protocol for Zymo Direct-zol RNA minipreps according to the manufacturer’s instructions. Total RNA was eluted from silica binding columns using 60 μ l of nuclease free water and immediately stored at –80°C.

For genomic DNA extraction from *S. mutans* libraries, cell pellets were thawed on ice and resuspended in 400 μ l of TES buffer. To ensure maximal lysis, 32 μ l of diluted Ready-Lyse lysozyme (Lucigen, 1:24 into TES) was added to cell suspensions followed by gentle vortex mixing. Solutions were then incubated at room temperature for 15 minutes. Tissue Lysis Buffer (400 μ l) from a Monarch Genomic DNA Purification Kit (NEB) was added and the entire solution was transferred to a 1.5 mL O-ring microcentrifuge tube filled with roughly 500 μ l of 0.1 mm glass beads. Cells were lysed using a bead mill homogenizer (Ohaus) at 1,600 RPM for 1.5 minutes. The O-ring tubes are then heated at 65 °C for 10 minutes. At this point, samples may enter the standard genomic DNA extraction protocol using a Monarch Genomic DNA Purification kit according to the manufacturer’s instructions. Genomic DNA was eluted from silica binding columns using 100 μ l of nuclease free water.

3.11 Preparation and Sequencing of Transcriptional Characterization NGS Libraries

Total RNA samples isolated from *S. mutans* transcriptional characterization strain libraries were reverse transcribed using a primer targeting the P1 sequence and a NEB Protoscript II First Strand cDNA Synthesis Kit in accordance with the manufacturer’s instructions. The resulting cDNA was purified with a Thermo GeneJet PCR Purification kit. A 50 μ l NEBNext Ultra II Q5 polymerase-mediated PCR reaction was assembled for each library using 5 ng of cDNA and primers (2.5 μ l of 10 μ M stocks for each) targeting the P1 and P2 sites. These primers contain non-hybridizing tails that enable the addition of Illumina TruSeq adapter sequences to amplicons, along with library-specific, custom phase-shifting dual indices developed for digital sequencing applications. After 25 cycles of amplification, the RNA barcode amplicon libraries were purified with 1.25× volumes of AMPure beads and eluted in 20 μ l of nuclease free water. For single-read DNA barcode libraries, 50 μ l NEBNext Ultra II Q5 polymerase-mediated PCR reactions were assembled for each library using 50 ng of genomic DNA and primers (2.5 μ l of 10 μ M stocks for each) targeting the P1 and P2 sites. Paired-end gene expression cassette libraries were prepared using genomic DNA templates with primers targeting the P1 and P3 sites and amplified independently of single-read libraries. After 23 cycles of amplification, both single-read DNA barcode libraries, and paired-end gene expression cassette libraries were purified with AMPure beads in the manner described for RNA barcode libraries. Single-read RNA and DNA barcode libraries were pooled in an equimolar manner and sequenced on a single SP 1×100nt flowcell using an Illumina NovaSeq 6000.

This process generated over 250 million reads. Paired-end gene expression cassette libraries were pooled in an equimolar manner and sequenced on a single SP 2×250nt flowcell using an Illumina NovaSeq 6000. This process generated over 900 million reads (R1 and R2).

3.12 Paired-end NGS Data Processing

The raw FASTQ files (R1, R2, and I1) of paired-end data were demultiplexed according to each amplicon’s i7 index using the `illumina-utils` package [43] developed by the Meren Lab at the University of Chicago allowing for no barcode mismatches. For the second round of demultiplexing using the custom, phase-shifting barcodes, I created a “pseudo-index” FASTQ file from each i7-demultiplexed R1 FASTQ file. This was done by trimming off all the R1 sequence except for the first eight nucleotides. This should contain only the custom phase-shifting barcode associated with R1. I then repeated the demultiplexing process with no allowed mismatches using this “pseudo-index” FASTQ in place of the I1 FASTQ file in order to fully demultiplex the paired-end libraries. I then used Trimmomatic v.0.39 [44] to trim and quality filter reads using a sliding window of size 5 and a quality PHRED score threshold of 20. Only reads with a length of 100 nucleotides or more were retained.

Paired end data for each library was then streamed into R in batches of 10 million read-pairs using the ShortRead package and processed simultaneously to extract molecular barcode sequence, base promoter identity, and ORF identity from each read-pair. The term base promoter refers to one of the 17 promoter sequences which served as template in the single position mutagenesis scheme. The processing pipeline, developed for this study, collapses the data from read-pairs to molecular barcodes by calculating barcode count, consensus base promoter and ORF ID, and consensus R2 sequence from all read-pairs with a given barcode sequence in the same streaming batch.

Barcodes were extracted from the R1 sequence using two methods. The first method takes the R1 sequence (acquired by reading from the P1 site) and tries to identify the P2 hybridization sequence, which should exist immediately after the molecular barcode. If the P2 hybridization sequence (GATGTCCACGAGGACTCT) is found, only the read information prior to it is retained. This information is then taken as the candidate barcode sequence. Because this strategy does not allow for sequence ambiguity, a second strategy was also employed. The second strategy uses the theoretical index range (1–20) of the molecular barcode downstream of the P1 site to select the candidate molecular barcode sequence. While these two strategies demonstrated a high degree of agreement in barcode sequence identity (>95%), a final quality control step was implemented to determine the true barcode sequence for each read-pair. Because the molecular barcodes were developed using the CutFree algorithm [42] to ensure that all barcodes were free of endogenous BsaI sites, all true barcode sequences should adhere to a known degenerate pattern, NNNNNANTNNMMNNMMNNNN. The quality control algorithm checks for this pattern in all candidate molecular barcodes identified using both barcode extraction strategies. If the candidate barcode sequence is full length (20 bp) and contains this degenerate pattern, it is accepted as the final barcode sequence for the read-pair.

Using this approach high confidence barcode identifications were achieved for >99% of the reads. Reads without a barcode identification were discarded. ORF identifications were made for each read by calculating the Levenshtein distance between the 50 nucleotides of each R1 sequence immediately after the P2 binding site (determined by theoretical indices) and the 50 nucleotides at the 3' end of the six ORF sequences used in combinatorial gene expression cassette assembly. The ORF with the minimum edit distance to the R1 50-mer was taken to be the ORF associated with that read-pair. An analogous approach using the first 50 nucleotides downstream of the P3 site in the R2 sequence and the first 50 nucleotides of each base promoter sequence was used to determine the base promoter identification for each read-pair.

Finally, all read-pair data was subjected to a reduction process. This process takes all read-pairs with a particular barcode sequence and checks to ensure they have the same base promoter and ORF IDs. The most abundant base promoter and ORF ID combination for a given barcode was taken to be the true ID pair and all non-matching read-pairs were discarded. All remaining read-pairs for a given barcode were then used to generate a consensus R2 sequence that could be used for final promoter and RBS identifications. The base promoter ID, ORF ID, consensus R2 sequence, and barcode count were then added to a new data structure where each observation corresponded to a single molecular barcode. This data structure was then appended to the data structure created from previous batches of streamed paired-read data. Once the streaming and processing pipeline had finished analyzing every read-pair within a particular library, the resulting data structure was subjected to a final round of data reduction using the same process outlined above to ensure that each observation was unique. In this case, consensus R2 sequences for the same barcode observed in different batches were created by scaling the existing consensus sequence by the number of barcode counts for each batch prior to consensus generation.

Upon inspection of the resulting data, it became evident that minor data cleaning was still required. By looking at barcodes with the same base promoter-ORF combination, it was observed that some such barcodes had extremely similar sequences. The likelihood of encountering two barcodes a short edit distance from one another, within the same library, and having the same base promoter-ORF combination is extremely low. In fact, by applying the statistical derivation in Appendix A to a situation with 100,000 random 20-mers (roughly the number of barcodes observed for a single library after processing the FASTQ streams), the probability of observing a random barcode within a Hamming distance of two from a given random barcode sequence, and those two barcodes being associated with the same base promoter-ORF ID (assuming cassette assembly is unbiased and random) would approximately be 1.57×10^{-6} . Simulations accounting for the degenerate nature of the barcode sequence lead to essentially the same result (data not shown). Thus, for the number of barcodes in a given library, one would not expect to observe a case where the barcodes in the above scenario are truly distinct. Therefore, the most likely scenario is that sequencing errors combined with the batched processing of read-pairs discussed above led to artificial inflation of barcode diversity. To remedy this, an error correction algorithm was developed and employed on all barcode sequences for each library. For each barcode, the algorithm calculates the

Levenshtein distance from said barcode to all other barcodes within the library. If other barcodes are observed within an edit distance of two, the algorithm compares the base promoter-ORF IDs and barcode counts of the relevant barcodes. If the two barcodes have the same base promoter-ORF ID, the barcode sequence corresponding to the barcode with a higher count is assumed to be correct and the sequence of the less-abundant barcode is changed to reflect this.

Additionally, some low abundance barcodes from the paired-end data produced noisy consensus sequences which did not appear similar to any base promoter. To address this issue, the entire promoter sequence was extracted from the R2 consensus sequence using theoretical indexing values based on the length of the base promoter and the known sequence of the Golden Gate scar (AATC) between the promoter and the RBS. The Levenshtein distance of this sequence to the known base promoter sequence was then calculated. Any barcode whose promoter sequence was greater than 10 edits from the base promoter sequence was discarded. Furthermore, only barcode sequences which were observed at least 20 times were kept in order to ensure a high degree of confidence in the generated consensus sequence.

With these data processing steps finished, it became possible to identify the exact promoter variant associated with each barcode. This was accomplished by extracting the entire promoter sequence as described above and calculating the Levenshtein distance from that sequence to the approximately 400 possible promoter sequences for that cassette's identified base promoter. The mutation-level resolution promoter was then identified by the sequence with the minimum edit distance to the observed promoter consensus sequence. The edit distance for that match was stored for future reference. For roughly 42% of barcodes across all six libraries, the exact promoter sequence could not be identified because multiple candidate sequences were observed to have the same minimized edit distance to the consensus sequence.

In order to determine which RBS was present for each cassette, the R2 consensus sequence was trimmed to remove the promoter sequence, leaving only the ribosome binding site and a portion of the ORF sequence. A fuzzy matching algorithm was then utilized to look for each of the 22 known ribosome binding site sequences provided a maximum Hamming distance of two between the query and the trimmed R2 consensus sequence. Exactly one RBS match was observed in every R2 consensus sequence, enabling confident identification of the RBS in each cassette.

3.13 Single-read NGS Data Processing

FASTQ files produced from gDNA and cDNA-templated single read libraries were demultiplexed using `fastx.barcode.splitter` (no mismatches allowed) and trimmed to only include the 20 nucleotides immediately downstream of the P1 site using `fastx.trimmer`. This sequence was taken to be the candidate molecular barcode for each read. The resulting FASTQ files were input into `fastq.collapser` to generate read counts for each barcode. The collapsed, single read DNA barcode file and collapsed, single read RNA barcode file were then joined by barcode sequence and converted to CSV files. These CSV files were then read into R and subjected to quality control. Any barcode

with less than counts in either the respective DNA or RNA libraries was discarded. An error correction algorithm was then employed to ensure the correct barcode sequence had been identified. For each barcode, the algorithm calculates the Levenshtein distance from said barcode to all other barcodes within the library. If other barcodes are observed within an edit distance of 1, the algorithm compares the relative counts of the relevant barcodes. The barcode sequence corresponding to the barcode with a higher count is assumed to be correct and the sequence of the less-abundant barcode is corrected to reflect this. After error correction, the data were re-collapsed and counts adjusted to ensure that each observation was unique. At this point, all barcodes counted less than 20 times in either the DNA or RNA single end libraries were discarded. The remaining single read barcode data was then joined with the paired-end data by barcode sequence to produce the final data structure for transcriptional characterization analysis.

3.14 Transcriptional Strength Quantification

Only barcodes observed in all three types of libraries were retained for transcriptional characterization. In addition, sequences which had a frequency greater than 0.001 in any of the three types of libraries were labeled as overrepresented and discarded. For every barcode, transcriptional strength was quantified as the base-10 logarithm of the ratio of said barcode’s frequency in its cDNA-templated library to the barcode’s frequency in either its paired-end, gDNA-templated library or its single-read, gDNA-templated library. This gave two different metrics for transcriptional strength. In order to minimize noise due to amplification bias or jackpotting, only barcodes where the two metrics were in agreement were retained for future analysis. A barcode’s transcriptional strength metrics were said to be in agreement if the difference between the two values was less than or equal to 0.3. From this point forward, the transcriptional strength of a given cassette refers to the metric calculated using DNA frequencies derived from paired-end, gDNA-templated libraries.

Figure 3.2 illustrates the transcriptional strengths for the 30,228 gene expression cassettes which passed quality filtering. These transcriptional strengths span over four orders of magnitude. The 17,645 gene expression cassettes had single-based mutation resolution. This revealed 3,735 different promoter varieties. For cassettes with a definitive promoter identity, 13,827 unique promoter-RBS combinations and 16,148 unique promoter-RBS-ORF combinations were observed.

3.15 ANOVA for Transcriptional Characterization Data

In order to understand the sources of variation within the transcriptional characterization assay, a multivariate linear regression model was constructed within a custom R script using data derived from the 16,409 gene expression cassettes with high-confidence, mutation-level promoter identifications. The model tried to predict the transcriptional strength of a given gene expression cassette provided its mutation-level promoter ID, RBS ID, ORF ID, and all pairwise interactions between base promoter ID, RBS ID, and ORF ID. Because 3,735 different promoter sequences were identified

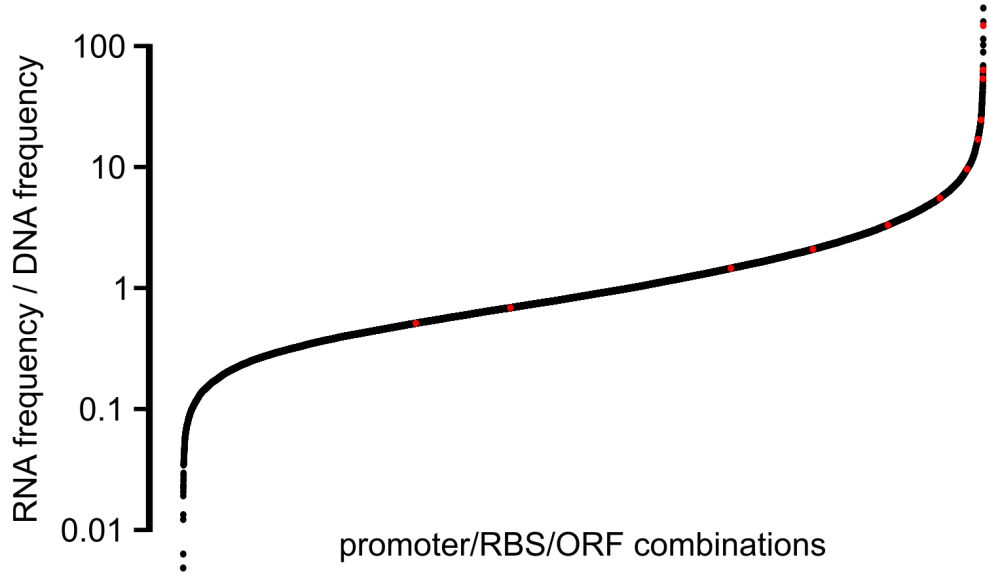


Figure 3.2: Transcriptional strengths was measured as the ratio of RNA barcodes to DNA barcodes in a pool of cells with different expression cassettes. The promoter variants selected for translational characterization appear in red and in Table 3.1.

within this dataset, a model using mutation-level promoter IDs to calculate pairwise interactions would require an enormous number of observations in order to produce a statistically powered model. To capture information on potential interactions between the promoter sequence and the other gene expression elements, base promoter IDs were substituted in interaction calculations for this model. For each term within the linear model, the sequential sum of squares was calculated. The percentage of variance corresponding to each term was then calculated by dividing the sum of squares associated with that term by the total sum of squares for the model and multiplied by 100.

The results for this variance decomposition analysis are shown in Figure 3.3. Promoter effects dominate the explained variance of transcriptional strength as determined using the NGS barcode

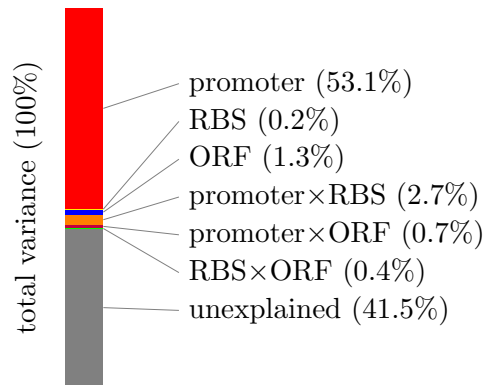


Figure 3.3: Promoter effects dominate the explained variance of transcriptional strength in the NGS assay. Total variance was decomposed by ANOVA on a linear model that predicts transcriptional strength as a function of promoter, RBS, ORF, and all two-way interactions.

counting assay. This is expected, as the promoter sequence is known to be the cassette component required for initiating transcription. Somewhat surprisingly, the effects for RBS and ORF as well as all pairwise interactions did not explain large portions of observed variance. This would seem to suggest, at least in the context of these combinatorial cassettes as expressed in *S. mutans*, that the RBS and ORF sequences do not effect transcription in a significant manner. A major caveat to this is that a large portion of the total variance in transcriptional strength as determined using this assay remains unexplained. Higher order interactions involving the RBS or ORF may exist and affect transcription. In addition, the use of RBSs and ORFs not in this toolkit may affect transcription rates while using a promoter from this toolkit. That said, for *S. mutans* synthetic biology experiments using the gene expression cassettes characterized herein, fine-tuned control of transcription can be accomplished without concern for non-promoter parts.

3.16 Effect of Mutation Position on Transcriptional Strength

While mutation position was not able to be used for modeling purposes due to imperfect information regarding transcription start sites for many of the promoters in our initial parts collection, previous studies have shown that mutations within the -35 and -10 sites of the promoter can have large effects on resulting gene expression [9]. To see if these effects could be observed within the transcriptional characterization data, the transcriptional strength of cassettes with mutation-level promoter IDs and the same base promoter sequence were visualized across the range of observed mutation locations. For well-sampled base promoters with a reasonably high level of median transcriptional strength, these effects were indeed observed, as shown in Figure 3.4. On average, mutations occurring within the -10 and -35 sites of PR13, PR12, PR9, and PR6 all result in promoters that are roughly 2–10 fold weaker than the respective base promoter (taken as the median value for all cassettes with a given base promoter). This information can be used to creating derivative constitutive promoter variants to further tune transcript levels.

3.17 Selection of Promoters for qPCR Validation

The goal of the transcriptional characterization assay was to identify a robust set of promoters whose transcriptional strengths span the observed range in a log-linear manner, while minimizing potential interaction effects with RBS and ORF sequences. This was accomplished in the following manner. To ensure only cassettes with high element identification confidences were considered for selection, gene expression cassettes without a single probabilistic, mutation-level promoter sequence identification and a Levenshtein distance greater than one between the calculated promoter consensus sequence and its theoretical mutation-level sequence were discarded. Any cassette whose transcriptional strength was calculated to be lower than the median transcriptional strength of cassettes harboring negative control-derived promoters were also eliminated. All remaining cassettes were then assigned to one of 24, log-linear spaced bins spanning the range of observed transcriptional strengths.

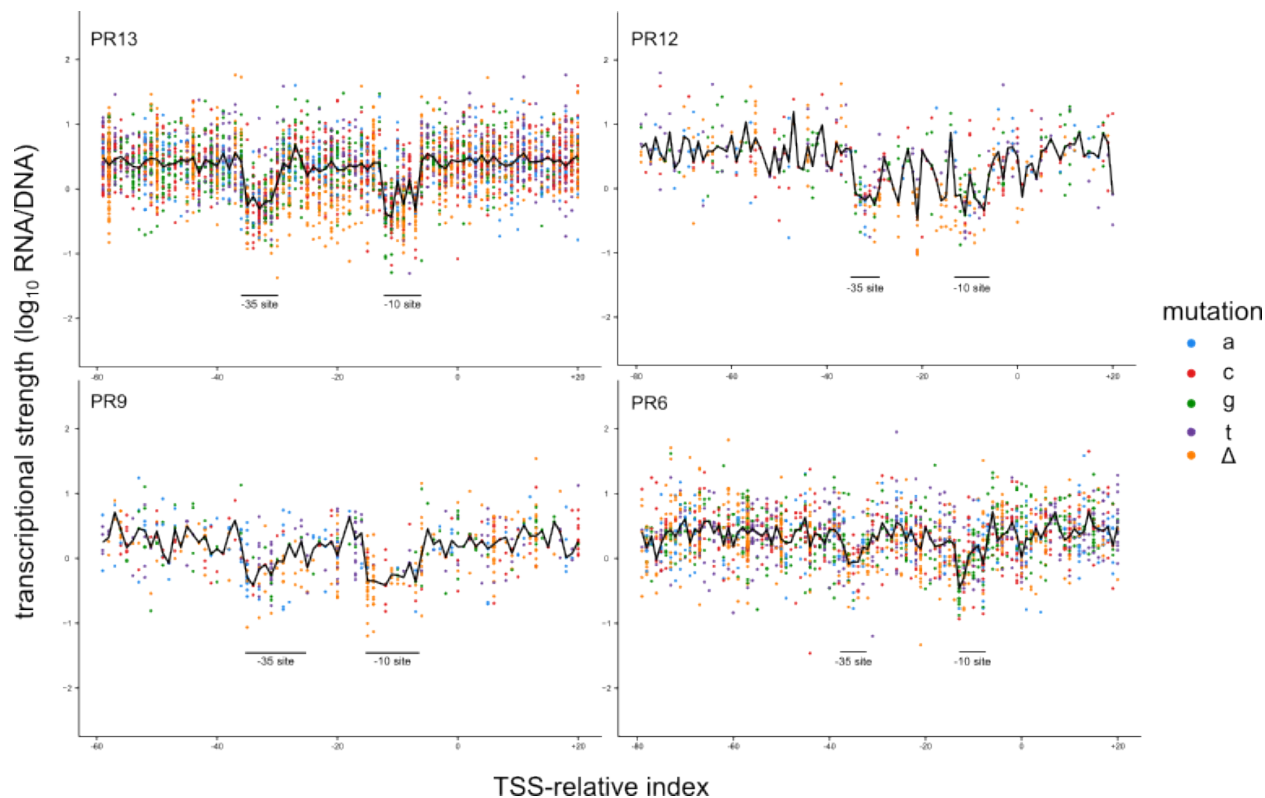


Figure 3.4: The effects of mutation position on base promoter transcriptional strength are shown for PR13, PR12, PR9 and PR6. While the ability to observe these effects depended heavily on the sampling rate of a particular base promoter sequence and its median transcriptional strength, these examples demonstrate the deleterious effects that mutations within the -35 and -10 site have on transcriptional strength.

To minimize the effects of assay noise on promoter set selection, a global confidence metric was calculated for each gene expression cassette. This metric is itself a function of “base-bin distance” and “local confidence”. Base-bin distance is calculated as the difference between the median transcriptional strength for a given cassette’s base promoter and the midpoint of its assigned bin. Assuming that for cassettes with the same base promoter, transcriptional strengths calculated using higher abundance barcodes are more reliable than those calculated with low abundance barcodes, the local confidence for a given gene expression cassette was calculated as the maximum of either the observed DNA barcode frequency or the observed RNA barcode frequency divided by the maximum barcode frequency observed for all gene expression cassettes with the same base promoter. In every transcriptional strength bin, the base-bin distances and local confidence scores were ranked, with low base-bin distances and high local confidence scores being preferred. Global confidence scores were then calculated as the unweighted mean of a gene expression cassettes’ local confidence ranking and its base-bin distance ranking. To select the “best” set of promoter sequences, the cassette with the highest global confidence score within each bin was chosen to be included in the final set for qPCR-validation.

To ensure that the selected gene expression cassettes transcriptional strengths were not dependent on the identity of the RBS or ORF sequence, I built a multivariate linear regression model using data from the 16,409 gene expression cassettes with high-confidence, mutation-level promoter IDs. The model predicts a gene expression cassette’s transcriptional strength using its mutation-level promoter ID, RBS ID, ORF ID, and all pairwise interactions between base promoter ID, RBS ID, and ORF ID. Unfortunately, a model using mutation-level promoter IDs to calculate pairwise interactions would require far more observations than I had available. Thus, I had to assume that interactions which arise between the promoter sequence and the RBS sequence or the ORF sequence are solely determined by the identity of the base promoter sequence. All statistically significant ($p < 0.05$) interaction terms were extracted from the model and cross-referenced with the promoters in the selected gene expression cassette set. In the context of predicting transcriptional strength, none of the 24 selected cassettes harbored promoters whose base promoter sequence was involved in any predicted significant interactions with assayed RBS sequences or ORF sequences.

3.18 qPCR Validation

The transcript abundance predictions from the NGS barcode assay were validated by qPCR. Signal corresponding to total barcode abundance was measured using qPCR primers targeting the P1 and P2 sites. Signal corresponding to the cassette-specific abundance was measured using primers targeting the P2 site and the molecular barcode sequence of each of the 24 selected gene expression cassettes. All primers were designed according to NEB’s recommendations for use with Luna qPCR kits. DNA barcode frequencies were determined using gDNA template, while RNA barcode frequencies were determined using cDNA template. Preliminary tests using end-point PCR with and without template revealed the need for preamplification of barcodes from genomic DNA and cDNA

libraries. This was accomplished with 20 cycles of amplification with NEBNext Ultra II Q5 DNA polymerase (NEB) and primers targeting the P1 and P2 sites. The resulting amplicons were purified using $1.25\times$ AMPure SPRI beads and eluted in nuclease-free water. qPCR reactions were assembled in the following manner: on ice, $0.5\ \mu\text{l}$ of a $10\ \mu\text{M}$ solution for each participating primer was added to $5\ \text{ng}$ of preamplified template inside a sterile, optically clear $200\ \mu\text{l}$ PCR tube (Thermo-Fisher Scientific). Nuclease free water was then added to a volume of $10\ \mu\text{l}$ before adding an additional $10\ \mu\text{l}$ of $2\times$ Luna Universal qPCR Master Mix (NEB). All qPCR reactions were assembled in triplicate. For each primer combination, three no-template qPCR reactions were also set up to check for non-specific amplification due to primer-primer interactions. Reactions were mixed by gently pipetting up and down. Tubes were then briefly spun down in a microcentrifuge at $10,000\times g$ and placed into a preheated QuantStudio 3 qPCR machine (Applied Biosciences) preprogrammed with NEB’s recommended thermocycling protocol for use with Luna Universal qPCR kits. After 42 cycles, threshold cycle numbers were determined using the QuantStudio 3 software package with default settings for SYBR-based analyses. Relative DNA and RNA barcode frequencies were calculated using the base-2 exponent of the difference between the cycle threshold for total barcode signal (primers targeting P1 and P2), and specific barcode signal (primers targeting P2 and the molecular barcode in question) for amplification templated with gDNA-derived amplicons and cDNA-derived amplicons, respectively. The transcriptional strength as determined by qPCR was then taken as the base-10 logarithm of the ratio between relative RNA barcode frequency and relative DNA barcode frequency. Because amplicons generated by qPCR were more similar to those generated during single-read NGS library preparation, comparisons made between the transcriptional strengths determined by qPCR and the molecular barcode counting assay used DNA barcode frequencies calculated with the single-read libraries. During analysis of the qPCR data, mean threshold cycle values for reactions using barcode specific primers were compared with mean threshold cycle values for those primers in the absence of template. If the templated reactions produced a mean threshold cycle number within two cycles of the template-independent reactions with the same primers, the respective transcriptional strength comparison was flagged as potentially noisy. Comparisons of qPCR-determined transcriptional strength and NGS-determined transcriptional strength are shown in Figure 3.5.

Overall, NGS and qPCR-determined transcriptional strengths agree well. The grey, univariate regression line fit to the data demonstrates a consistent relationship between the two measurement techniques ($R^2 = 0.65$). The reactions at the extremes of transcriptional strength (colored orange and yellow) where the data deviate most from the regression line are cases where the cycle threshold values from barcode-specific quantifications were similar to appropriate no template controls. This is to be expected as cassettes with very low or very high transcriptional strengths are likely to have relatively low abundances of RNA or DNA barcodes respectively. If these low confidence points are removed from consideration, the regression fit improves significantly ($R^2 = 0.79$). The results of the qPCR validation experiment confirm the reliability of the transcriptional strengths determined by the barcode counting assay.

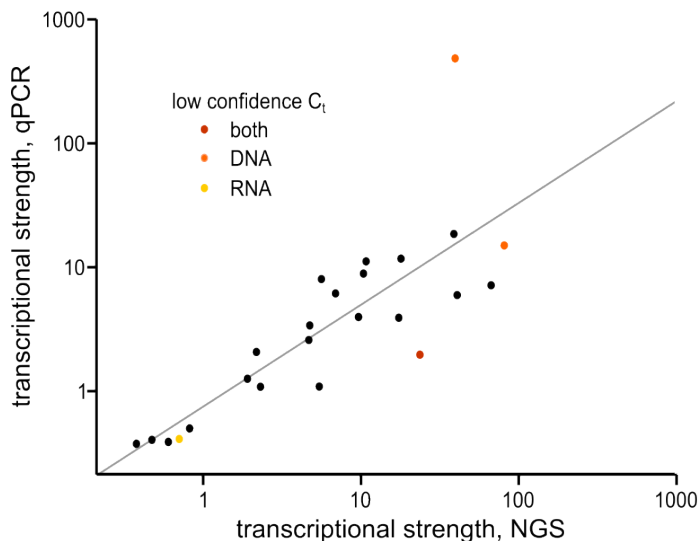


Figure 3.5: Comparison of transcriptional strengths measured with qPCR and with a NGS barcode counting assay. Colored data points indicate cases where the threshold cycle for barcode specific reactions was similar to no template controls.

3.19 Selection of Final Characterized Promoter Set

The final set of promoters to be carried into use with translational characterization experiments (Chapter 4) were selected in the following manner. The transcriptional strengths measured by qPCR and both varieties of NGS barcode counts were compared for all 24 gene expression cassettes measured with qPCR. The ANOVA run on the above model to predict transcriptional strength revealed the main effect of the mutation level promoter to be responsible for over 93% of the explained variance. Thus, it can be reasonably stated that in the context of comparing transcriptional strengths, only promoter identity matters. For the purposes of creating a set of characterized promoters for transcriptional control within *S. mutans*, we can therefore ignore the effects of all other gene expression cassette elements in these 24 cassettes and focus solely on their promoters. We decided that 24 promoters was too many to span the ~ 2.5 orders of magnitude of transcriptional strength. Instead, a final set of 12 promoters was selected from the 24 qPCR-validated gene expression cassettes by comparing the transcriptional strengths estimated using qPCR to both values calculated using the NGS-barcode counting assay. Every two bins considered in the qPCR experiments were merged to create a so called “superbin” starting from the lowest transcriptionally active bins and moving upwards. For each bin, I manually inspected all candidate promoter sequences, and selected the promoter whose transcriptional strength metrics seemed to be in the greatest agreement. The final set of characterized promoters are shown in Table 3.1.

Promoter	Strength	Sequence (5'–3')
PR8.22.Δ	0.42	tgtgacattctcttgaaaaaggctaagccgaggctaaaagccttgtagg gacaacgttcttgatatttttagagagcagggaattaattgaaagcag
PR10.77.Δ	0.51	atatagtcgatttattaaatctttagataaaactgtaaaaagtggttcca tagccactttttgctataatagaggcgtaaaattgaaggagtcggctat
PR9.59.A	1.5	cactagtcattctacagtttattcttgacattgcactgtccccctgggat aataactaaacatgcaagatctaaatagga
PR6.65.T	1.7	gttagtctccttaaaaaataatcaaaaaagtttttttaaaaacgcttgcaat tatgcttgaaaagtggtataacttataagtactgcaaaatacgcaacgttc
PR12.99.T	2.5	aactgaagctgttgtaaaattaataaaaaaagtttaaaaaagtagttgac aaagtttgaaaagactgtataatagtaagagttgaaaataacaactcatg
PR13.31.A	6.2	gaaaaaatatttttcaaaaccacttgacaatttagctatatatgtttat aataactatgaggatatctgctgctaggatcta
PR12.92.A	7.6	aactgaagctgttgtaaaattaataaaaaaagtttaaaaaagtagttgac aaagtttgaaaagactgtataatagtaagagttgaaaataaaaactcagg
PR12.33.C	12	aactgaagctgttgtaaaattaataaaaaaagcttaaaaaagtagttgac aaagtttgaaaagactgtataatagtaagagttgaaaataacaactcagg
PR4.28.G	13	aataaattttgaaaaaaatttataaaaaagatttcaaaaaagacttgacata ttataaattatattatataatataattagaaagataaagagaggagaaaa
PR13.71.T	34	gaaaaaatatttttcaaaaccacttgacacttttagctatatatgtttat aataactatgaggatatctgttgctaggatcta
PR4.52.G	50	aataaattttgaaaaaaatttataaaaaaatttcaaaaaagacttgacata tgataaattatattatataatataattagaaagataaagagaggagaaaa
PR4.12.Δ	110	aataaattttgaaaaaaatttataaaaaaatttcaaaaaagacttgacatat tataaattatattatataatataattagaaagataaagagaggagaaaa

Table 3.1: Twelve promoter variants were selected for the fluorescent reporter assays. The promoter variants span a range of transcriptional strengths, calculated as the ratio of RNA:DNA barcode sequences in the NGS reporter assay. The promoter PR8.22.Δ is a variant of the PR8 promoter with a single base deletion at position 22. PR9.59.A is a variant of PR9 where position 59 is mutated to an A.

Chapter 4

Translational characterization of combinatorial gene expression cassettes

4.1 Overview

Fluorescent proteins have revolutionized the fields of molecular and synthetic biology by allowing proteins to be detected and quantified *in vivo*. While bioluminescent reporters such as firefly luciferase offer a wider dynamic range, fluorescent proteins have become the reporter of choice for large-scale gene expression screens due to their ease of use [9]. Fluorescent proteins are also easier to multiplex, as spectrally distinct reporter ORFs can be expressed within the same cell and quantified independently of one another without significant cross talk. The first generation of fluorescent proteins, including avGFP22 [9] and DsRed23 [45], have given rise to rational mutagenesis and directed evolution efforts in order to expand the arsenal of fluorescent reporters [46]. These efforts were motivated by the fact that first generation fluorescent proteins often had inefficient quantum yields, slow maturation times, poor signal generation properties, or required oligomerization in order to form a functional fluorescent unit. The vast majority of the fluorescent proteins previously used in the oral streptococci have been first generation fluorescent proteins. This limits the temporal resolution and sensitivity of fluorescent reporter assays.

The effect of RBSs on translation is commonly quantified using fluorescent reporter ORFs. These proteins enable sensitive and highly specific *in vivo* detection of heterologous gene expression cassettes. The signal generated from these cassettes can be modulated by changing the RBS; however, RBS identity is only one factor in determining the efficiency of translation. Many extrinsic factors such as ribosomal protein abundance, rRNA abundance, tRNA abundance, and metabolic load influence the levels of gene expression within a given cell. Rudge *et al.* demonstrated that using an expression host that produces a spectrally distinct fluorescent protein from a reference cassette allows for signal normalization in order to account for these extrinsic factors [47]. By

utilizing the ratio of experimental cassette fluorescence to reference cassette fluorescence at multiple time points, these authors were able to robustly characterize gene expression elements. We apply a similar approach when constructing reporter strains (§4.2).

4.2 RMW6 Construction

In the context of measuring expression levels from synthetic cassettes, the use of ratiometric response variables has been shown to partially eliminate noise derived from external factors like cellular stresses and variations in the levels of transcriptional and translational machinery [47]. Bright, rapidly maturing fluorescent proteins like those in Table 2.3 are particularly well suited for these roles. For the purposes of RACCET, an expression strain of *S. mutans* was engineered to express mScarlet-I from the *gtfA* locus, while simultaneously harboring the IFDC3 positive and negative selection marker cassette [48] in the *phnA-mtlA* locus. Red fluorescent proteins produced from the former site serve as a reference level of protein expression, while the latter cassette acts as a “landing zone” for insertions into the *phnA-mtlA* locus and provides a convenient means of transformant selection.

It is important to note that the promoter/RBS pair driving expression of mScarlet-I is PR1 and RBS1. These represent the P₂₃ promoter—currently held to be the most reliable high-expression constitutive promoter for the oral streptococci—and RBS1, the RBS used by Biswas *et al.* in their early work on gene expression within the oral streptococci [13]. As the translational strength determined in RACCET is the ratio of two fluorescent signals, a translational strength of one would indicate, barring significant ORF effects, a promoter/RBS combination with equal strength to the field’s standard for high expression levels.

S. mutans RMW6 was created in a stepwise manner. First, homology arms targeting the *gtfA* gene were amplified using PCR. The resulting 5’ homology amplicon consisted of the native *S. mutans* UA159 chromosome sequence from 833110–833679 with an Overhang 1 sequence at the 3’ end. The 3’ homology amplicon consisted of an Overhang 4 sequence at the 5’ end of the construct followed by the native *S. mutans* UA159 chromosome sequence from 834008–835126. Insulating terminator arms were added to these amplicons in the manner described in Section 3.5. The IFDC3 cassette, containing a mutant *pheS* gene allowing for the charging of the toxic phenylalanine analog 4-chlorophenylalanine and the *aph3* kanamycin resistance marker was synthesized as a gBlock (IDT) and amplified using primers that add Overhang 1 and Overhang 4 sequences at the 5’ and 3’ end of the amplicon, respectively. The IFDC3 and homology arm amplicons were purified using 1× AMPure SPRI beads and eluted in nuclease free water. The insertion cassette was created with Golden Gate assembly and immediately used to transform *S. mutans* UA159. Transformants were selected for on THY agar plates supplemented with 200 µg/mL of kanamycin. A single colony was picked and verified with colony PCR using primers targeting a region upstream of the insertion site and the 3’ end of the *aph3* gene. The verified colony was then used to start an overnight culture in fresh THY supplemented with 200 µg/mL of kanamycin. Meanwhile, the *mScarlet-I*

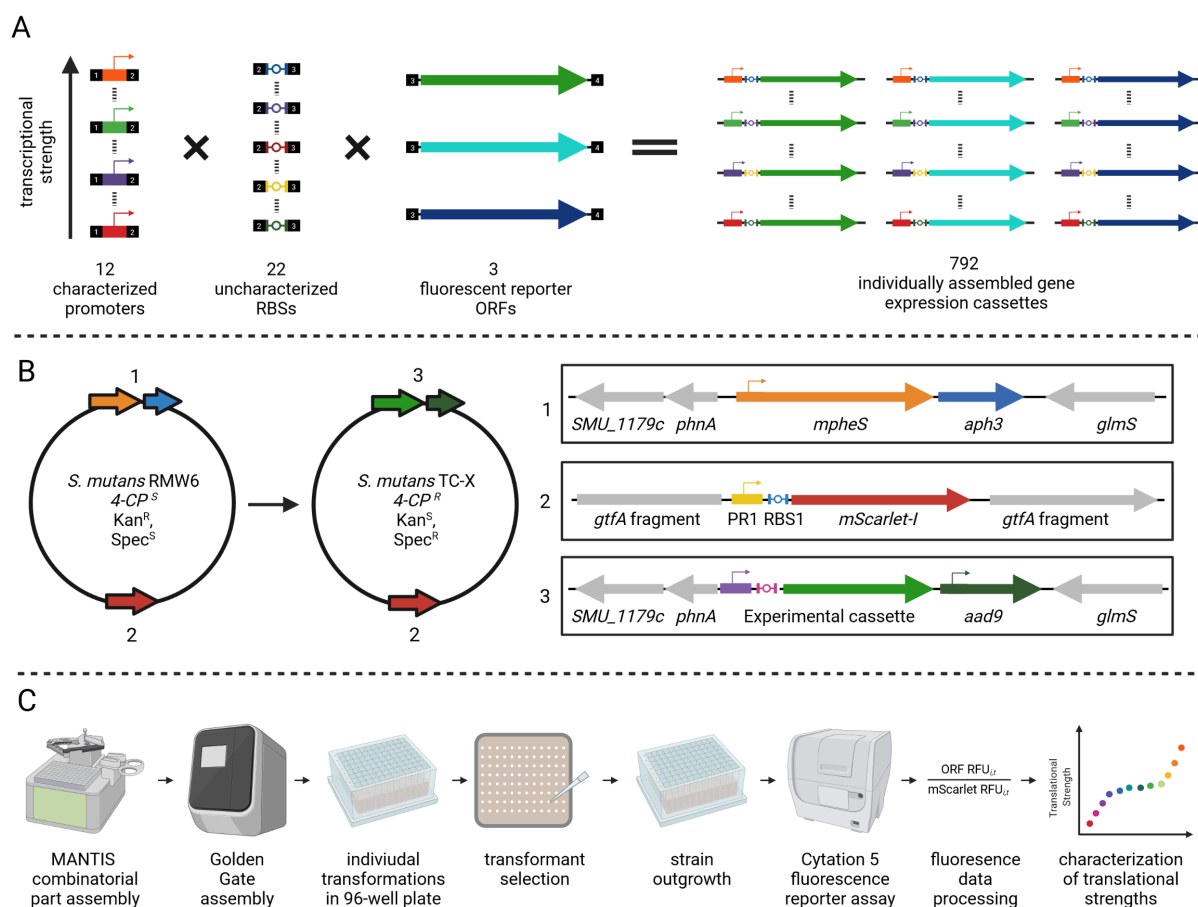


Figure 4.1: Overview of the translational characterization stage of RACCET. 792 strains are individually constructed by combining the set of 12 transcriptionally characterized promoters with all 22 RBSs and three bright, monomeric, fluorescent reporter ORFs: mNeonGreen, mTurquoise2, and mTagBFP2 (**A.**). These resulting gene expression cassettes are used to transform *S. mutans* RMW6, a strain preloaded with a counterselection marker in the insertion site, and a reference expression cassette expressing mScarlet-I in a distal site on the chromosome (**B.**). **C.** shows the workflow starting from cassette assembly and ending in quantification of translational strength.

gene was amplified off PRW49, a sequence verified pRW17-derivative, using primers targeting the 5' and 3' ends of the ORF. The resulting amplicon contained Overhang 1 and Overhang 2 sites at the 5' and 3' sites respectively. The purified amplicon was then used in a Golden Gate Assembly reaction with the same *gtfA*-targeting homology arms mentioned above—PR1 and RBS1. The next day, the clone harboring the IFDC3 cassette in the *gtfA* locus was transformed with the *mScarlet-I* insertion cassette and transformants were selected for on THY plates supplemented with 4 mg/mL of 4-chlorophenylalanine. A single colony was picked and verified with colony PCR using primers targeting a region upstream of the insertion site and the 3' end of the *mScarlet-I* gene. The verified colony was then used to start an overnight culture in fresh THY supplemented with 4 mg/mL of 4-chlorophenylalanine. Meanwhile, the homology arm constructs used to target the *phnA-mtlA* locus and described in Section 3.5 were used in a Golden Gate Assembly reaction with the IFDC3 amplicon mentioned above. The next day, the clone harboring the *mScarlet-I* gene in the *gtfA* locus was transformed with the IFDC3 cassette targeted to the *phnA-mtlA* locus. Transformants were selected for on THY agar plates supplemented with 200 μ g/mL of kanamycin. Four colonies were picked and verified with two separate colony PCRs. the first PCR was done using primers targeting a region upstream of the *gtfA* insertion site and the 3' end of the *mScarlet-I* gene. The second PCR was done using primers targeting a region upstream of the *phnA-mtlA* locus and the 3' end of the *aph3* gene. Clones which passed colony PCR validation were functionally screened for sensitivity to 4-chlorophenylalanine by overnight growth in THY supplemented with or without 4 mg/mL 4-chlorophenylalanine. The *mScarlet-I* cassette for all clones sensitive to 4-chlorophenylalanine were amplified and validated with Sanger sequencing. A clone who passed all validation steps was renamed *S. mutans* RMW6.

4.3 High-throughput, Combinatorial Strain Construction

4.3.1 Promoter Assembly

The sequences of the final set of 12 characterized promoters were used to design two pairs of primers per promoter sequence that would be used for assembly of the promoter cassette. The first pair, deemed the “core pair” of primers were designed to have approximately 20–30 base pairs of complementarity to each other at their 3' ends, with approximately 30–40 nucleotides making up non-hybridizing tails. The sequence overlap between the core primer pair was specifically designed such that they had an annealing temperature between 60–65 °C according to the NEB T_m Calculator for amplifications with Q5 DNA Polymerase. Core primer pairs were also screened with the Oligo Calc [49] to ensure each pair contained no significant unwanted secondary structures which may interfere with faithful amplification. The sequence information contained by this first pair of primers accounted for the most interior 80 base pairs of the promoter sequence. For all of the final 12 promoters except those derived from PR4 (P_{phage}), a second primer pair deemed the “extension pair” was also designed. These pairs of primers were designed to target sequences in the non-hybridizing portions of the core pair, with roughly 20–30 base pairs of complementarity. The

extension pair also contained non-hybridizing tail sequences which contained the remainder of the promoter sequence information as well as sequences to create Overhang 1 sites at the 5' end of the final amplicon, as well as Overhang 2 sites at the 3' end of the final amplicon. Importantly, the regions of complementarity between the extension pair and their core pair counterparts were designed to have an annealing temperature within 5 °C of the annealing temperature for the core pair of the same promoter cassette, as determined using the NEB T_m Calculator for amplifications with Q5 DNA Polymerase. All primer pairs were individually synthesized as oligonucleotides at the 25 nmol scale using IDT's standard DNA oligo synthesis service. Because the promoter sequence of PR4-derived promoters is so GC-poor (~13%), the two-pair assembly strategy outlined above was not viewed as practical. Instead, the entire promoter sequence including an Overhang 1 sequence at the 5' end and an Overhang 2 sequence was synthesized as an Ultramer at the 4 nmol scale (IDT). A common pair of extension primers was designed to target the Overhang 1 and Overhang 2 sequences at the end of all PR4-derived promoter cassettes. Promoter cassette assembly was accomplished in the following manner: 1 μ l of 20 pM stocks for each core primer (or 1 μ l of 20 pM Ultramer for PR4-derived promoters) was added to a sterile 200 μ l PCR tube (USA Scientific) with 18 μ l of nuclease free water. Stocks (2.5 μ l at 10 μ M) for each respective pair of extension primers were added, followed by 25 μ l of 2 $\times$ Hot Start High Fidelity Q5 DNA Polymerase Master Mix (NEB). After 25 cycles of amplification according to the standard protocol recommended by the polymerase's manufacturer, promoter amplicons were purified using 1.25 $\times$ AMPure SPRI beads, eluted in 20 μ l of nuclease free water, and validated by gel electrophoresis. All promoter amplicon concentrations were measured using a NanoDrop One C and normalized to 40 fmol/ μ l (~3 ng/ μ l) by the addition of appropriate amounts of nuclease free water. Promoter amplicon stocks were then stored at -20 °C.

4.3.2 RBS Assembly

For each of the 22 RBS cassettes in the initial gene expression toolkit, two individual, partially complementary oligonucleotides were designed. The hybridization product of these two oligonucleotides would result in a 5' ssDNA overhang sequence and a 3' ssDNA overhang sequence mimicking the digestion products of the Overhang 2 and Overhang 3 sequences thus enabling participation in Golden Gate Assembly with promoter and ORF elements. Each oligonucleotide was synthesized (IDT) separately and individually added to a hybridization and phosphorylation reaction with its designated partner. Briefly, 1 μ l of 10 μ M stocks for each oligonucleotide in a given RBS cassette was added to 41 μ l nuclease free water, 5 μ l of 10 $\times$ T4 DNA Ligase Buffer (NEB), and 2 μ l of T4 PNK (NEB, 10 U/ μ l). The reaction was incubated at 37 °C for 30 minutes, heated to 95 °C for 5 minutes, and then returned to 25 °C by ramping down at a rate of -5 °C/minute. RBS assembly reactions were then diluted 1:10 into nuclease free water to create stocks with an approximate concentration of 40 fmol/ μ l (0.75 ng/ μ l) and stored at -20 °C.

4.3.3 ORF Assembly

Plasmids harboring the correct ORF sequence for mNeonGreen, mTurquoise2, and mTagBFP2 were used as template for PCR amplification with ORF-specific primers with non-hybridizing tails. The primer targeting the 5' end of each ORF contained a non-hybridizing tail with baggage sequence and the Overhang 3 sequence for Golden Gate Assembly. The ORF-specific primer targeting the 3' end of the ORF sequence contained a non-hybridizing tail that added the Overhang 4 sequence for Golden Gate Assembly. Amplicons were purified with 0.5× AMPure SPRI beads and validated by gel electrophoresis. All ORF amplicon concentrations were measured using a NanoDrop One C and normalized to 200 fmol/ μ l ($\sim$ 90 ng/ μ l). ORF amplicon stocks were stored at -20°C .

4.3.4 Homology Arm Assembly

Archived genomic DNA from RMW12 was used as template in PCRs to generate phn.HisB.P3.HA1 and SynTerm.SpecR.mtlA.HA2 amplicons. These amplification products were purified using 0.5× AMPure SPRI beads and validated with gel electrophoresis. For both amplicons, concentrations were measured using a NanoDrop One C and normalized to 200 fmol/ μ l ($\sim$ 150 ng/ μ l for phn.HisB.P3.HA1, and $\sim$ 300 ng/ μ l for SynTerm.SpecR.mtlA.HA2). Homology arm construct amplicons were then stored at -20°C .

4.3.5 Golden Gate Reactions

The construction of 792 completely unique Golden Gate Assembly reactions by hand is impractical and likely to result in pipetting errors by the experimenter. To remedy this, a Golden Gate reaction “order form” was created using a custom R script. The R script randomly assigns a 96-well plate position to each reaction and calculates the required volume for every participating reagent in that reaction. For the purposes of future fluorescence reporter assays, reactions were divided by ORF, with all 264 reactions for a given ORF spread across three 96-well plates. The output from this script then serves as input into a custom Python script. The Python script converts the order form into a dispense list text file readable by Formulatrix liquid handling robots.

A Formulatrix MANTIS was used to assemble all Golden Gate Assembly reactions for translational characterization. The prepared promoter, RBS, ORF, and homology arm constructs were thawed on ice. All MANTIS chips were washed with 70% ethanol and nuclease-free water prior to use. Due to the MANTIS' limit of 28 reagents at one time, reaction assembly was completed in two rounds. In the first round, 100 μ l of each of the 12 promoter stocks (40 fmol/ μ l) was loaded into a corresponding low-volume MANTIS chip. The MANTIS then dispensed the appropriate promoter into its assigned well within a sterile TriFlex 96-well PCR plate (Applied Biosystems). The chips were then washed with 70% ethanol and nuclease-free water. In the second round, 100 μ l of each of the 22 RBS stocks (40 fmol/ μ l) was loaded into a corresponding low-volume MANTIS chip. Master mix solutions containing water, ORF stock, and homology arm stocks were created and loaded into a corresponding high-volume MANTIS chip. The MANTIS then dispensed the appropriate amount

of each reagent into its respective well. The result was 792 unique 10 μ l solutions containing 40 fmol of the appropriate promoter, RBS, ORF, and homology arm components.

To prepare Golden Gate reactions from these solutions, 6 μ l nuclease-free water, 2 μ l of 10 $\times$ T4 DNA Ligase Buffer (NEB), 1 μ l T4 DNA Ligase (NEB), and 1 μ l BsaI-HF v2 (NEB) were added. Reactions were mixed by pipetting and placed into a preheated ProFlex PCR system (Applied Biosystems). The reactions were incubated for 30 cycles of 5 minutes at 37 °C followed by 5 minutes at 16 °C. Reactions were subjected to a final incubation at 37 °C for 10 minutes before being stored at -20°C until needed.

4.3.6 Golden Gate Reaction Quality Control

In order to ensure the faithful assembly of all expression cassettes, the following quality control scheme was adopted. Eight reactions were randomly selected from each plate of Golden Gate Assemblies. These reactions were sampled and used as template in eight PCR amplifications with OneTaq DNA Polymerase (NEB) and primers targeting the 5' end of the 5' homology arm and the 3' end of the 3' homology arm. Reactions were run according to the manufacturer suggested protocol for OneTaq DNA Polymerase. After 34 cycles of amplification, 15 μ l of each 25 μ l reaction was used for fragment size validation with DNA gel electrophoresis. A reaction passed quality control if the correct ~ 4 kb amplicon representing the full-length assembly product was observed. An acceptable fail limit was selected using ANSI standards for general batch inspection to ensure an overall success rate of 90%. This resulted in a batch being rejected if three or more quality control PCRs failed. Plates which failed quality control were discarded and all associated Golden Gate reactions were re-assembled in the same manner discussed above. If a plate passed quality control, it was allowed to supply DNA for transformation reactions.

4.4 Measuring protein translation *in vivo*

4.4.1 Translational Characterization Strain Construction and Validation

The following operations were performed with one plate of Golden Gate reactions each day. An overnight culture of *S. mutans* RMW6 in fresh media supplemented with 200 $\mu\text{g}/\text{mL}$ of kanamycin was diluted 1:40 into 70 mL of fresh THY. The dilution was grown in a 5% CO_2 incubator at 37 °C until an OD_{600} of approximately 0.15 was reached (~ 4 hours post-dilution). At this point, 70 μ l of the *S. mutans* competence pheromone, CSP-18 (1 $\mu\text{g}/\mu\text{l}$) was added to the culture. The culture was vortexed briefly and returned to the incubator for 20 minutes. During this time, 10 μ l of each Golden Gate reaction was transferred into its corresponding well in a 1.2 mL, 96-deep well plate (Thermo Fisher Scientific). *S. mutans* RMW6 culture (500 μ l) was added to each well. Transformation reactions were mixed by pipetting. Plates were sealed using sterile, breathable, rayon sealing tape (Thermo Fisher Scientific) and returned to the incubator for 2 hours. After this time, 5 μ l of transformation reaction was spotted onto pre-warmed, THY plates supplemented

with 100 $\mu\text{g}/\text{mL}$ spectinomycin and allowed to dry. After 10 minutes, plates were returned to the incubator.

After 36 hours, colony spots (~ 10 – 100 CFUs) each corresponding to a single expression strain were individually scraped into 1 mL of fresh THY supplemented with 100 $\mu\text{g}/\text{mL}$ spectinomycin within a 2.2 mL, 96-deep well plate (Thermo Fisher Scientific) using sterile toothpicks. Plates were sealed using sterile, breathable, rayon sealing tape (Thermo Fisher Scientific) and incubated at 37 °C with 5% CO_2 overnight. The next morning, 2 μl of the overnight culture was used as template for quality control PCR amplifications with OneTaq DNA Polymerase (NEB) and primers targeting the genomic sequence 50 bp upstream of the insertion locus and the 3' end of cassette's ORF. Reactions were run according to the manufacturer suggested protocol for OneTaq DNA Polymerase. After 34 cycles of amplification, 15 μl of each 25 μl reaction was used for fragment size validation with DNA gel electrophoresis. A reaction passed quality control if the correct ~ 2.1 kb amplicon representing the full-length gene expression cassette correctly inserted into the *phnA-mtlA* locus was observed. Of the 792 expression strains attempted, 770 expression strains (97.2%) were created and passed quality control. Any strain that did not pass quality control was excluded from future analyses. Freezer stocks for each strain were created by combining 375 μl of overnight culture with 375 μl fresh THY and 250 μl of 60% (vol/vol) glycerol in a sterile 96-well microplate (Costar). Cell solutions were mixed by pipetting. Freezer stock plates were then sealed using aluminum cryogenic seal tape (USA Scientific) and placed in a -80 °C freezer. Overnight cultures of expression strains were started by inoculating 1 mL of fresh THY supplemented with 100 $\mu\text{g}/\text{mL}$ spectinomycin within a 2.2 mL 96-deep well plate using the appropriate freezer stock plate.

4.4.2 Fluorescence Reporter Measurements

Reporter assay plates were laid out in following manner. In each row of the first column of a black-wall 96-well microplate (Costar), 245 μl of prewarmed CDM was inoculated using 5 μl of a 5 mL overnight culture of *S. mutans* UA159 in THY. The other 11 columns of the plate were filled with 245 μl of prewarmed CDM and inoculated using 5 μl of an overnight culture of an experimental expression strain grown overnight in 1 mL of THY supplemented with 100 $\mu\text{g}/\text{mL}$ spectinomycin inside a 2.2 mL 96-deep well plate. The plate was then grown inside a Cytation 5 multi-mode plate reader at 37 °C with 5% CO_2 and periodic shaking to prevent cell clumping. Every 10 minutes over the next 24 hours, OD_{600} , mScarlet-I fluorescence, and experimental ORF fluorescence for each well were measured. The empirical max excitation and emission wavelengths for mScarlet-I, mNeonGreen, mTurquoise2, and mTagBFP2 found in Table 2.3 were used in all fluorescence reporter assays. After 24 hours, assay data were exported to Microsoft Excel for future analysis.

4.5 Translation Strength Analysis

4.5.1 Fluorescence Data Processing

For each plate, the single Excel file containing OD₆₀₀, raw mScarlet-I fluorescence, and raw experimental ORF fluorescence values was broken into three CSV files corresponding to the three measured variables. A fourth CSV file containing the data to map from well position to expression cassette ID for each plate was also created. These four files were read into a custom R analysis pipeline. First, a new data structure combining information from these four files was created where each observation corresponds to a particular well within a certain plate. Wells where no growth was observed were removed by calculating the difference between the initial and final OD₆₀₀ measurements. Wells with a difference less than 0.3 were discarded from consideration. To eliminate noisy non-monotonic signal from each observation, the first five measurements in the time series for OD₆₀₀, raw mScarlet-I fluorescence, and raw experimental ORF fluorescence were discarded. To account for background fluorescence, two nonlinear local polynomial LOESS regression models were built from fluorescence data for either the reference and experimental channels using data from non-fluorescent wild-type control wells. To supplement or replace wild-type controls, background subtraction models can also be built from wells harboring PR8.22.Δ-containing cassettes, as these were found to behave extremely similarly on average to the wild-type strain. These models were trained to predict background fluorescence in their respective channel provided only the OD₆₀₀ of the culture. Background fluorescence was then removed from the fluorescence values for each well by subtracting the model-predicted value from the observed value. Translational strength at each timepoint was calculated by taking the ratio of the experimental ORF fluorescence without background to the reference ORF fluorescence without background. The timeseries of the ratiometric response for the entire plate were then visualized to determine the bounds of the steady state component of the ratio signal. A noisy transient phase was observed from the beginning of measurement until roughly an OD₆₀₀ of 0.3–0.4. Remarkably, the ratio signal’s steady state component lasted from an OD₆₀₀ of roughly 0.4 until an OD₆₀₀ of roughly 1.25 for all fluorescent ORFs used in the reporter assay. After this, the signal became relatively unstable, likely due to stationary phase effects such as cell lysis and death. The translational strength of a cassette was taken to be the mean value of the ratio measurements within steady state.

Of the 770 expression strains participating in the fluorescent reporter assays, 654 produced signal above background. Their translational strengths are shown in Figure 4.2. The translational strengths span four orders of magnitude. Of the possible 264 combinations of promoter and RBS, 228 were observed above background for at least two of the three ORFs and 143 were observed above background for all three ORFs.

4.5.2 Modeling Translational Strength

To determine what expression cassette components affect translational strength in *S. mutans*, two linear multivariate regression models were built using expression data from the 654 assayed strains.

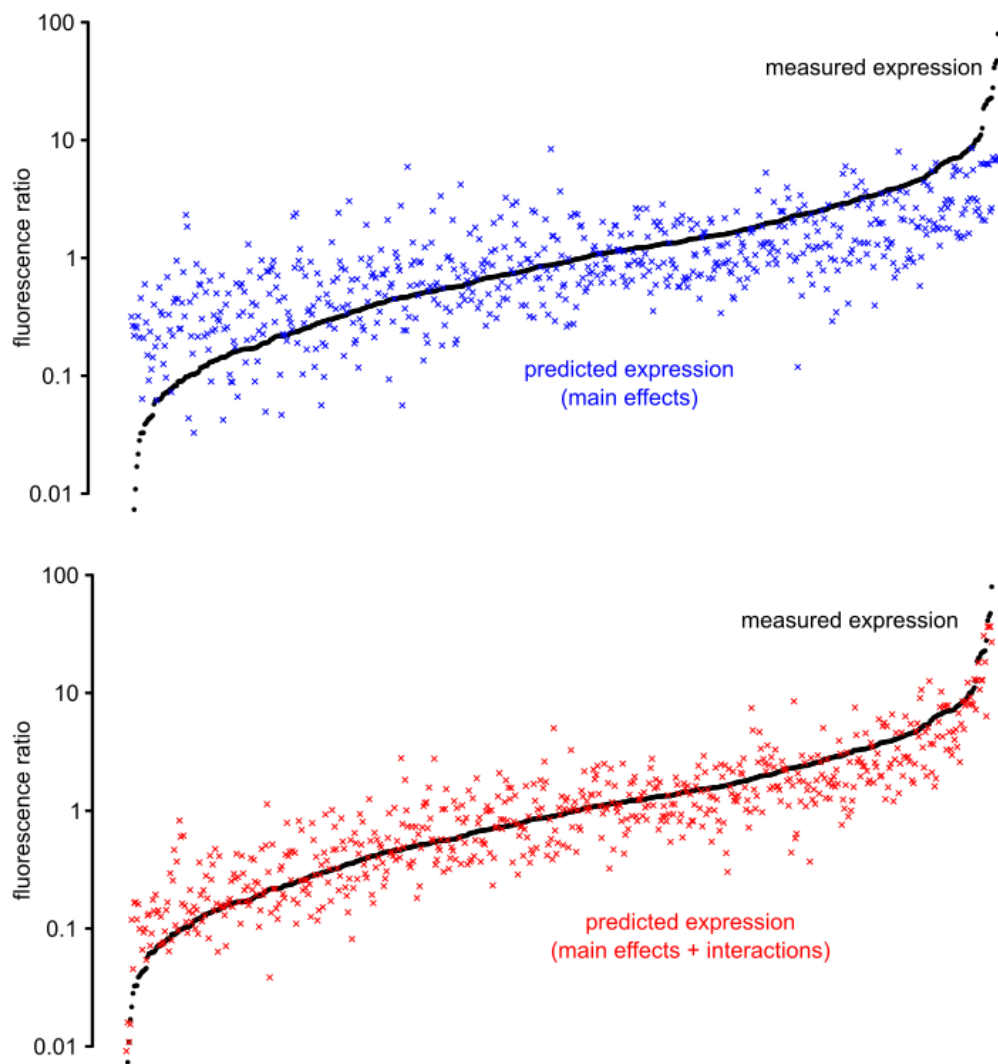


Figure 4.2: Translational strength of the combinatorial expression cassettes was measured as the ratio of the expressed fluorescent protein to a mScarlet-I control protein. A linear model considering only main effects (promoter, RBS, and ORF) does not predict high- or low-expression cassettes. A second model that includes main effects and interactions (promoter $\times$ RBS, promoter $\times$ ORF, and RBS $\times$ ORF) predicts better at the extremes of the data, suggesting that high and low expression is dominated by interactions among genetic parts.

The first model predicts the log-transformed translational strength of an expression cassette using only the main effects provided by the promoter, RBS, and ORF IDs. This model’s predictions are overlayed in blue on the translational strengths in Figure 4.2. The second model predicts the log-transformed translational strength of an expression cassette using the main effects provided by the promoter, RBS, and ORF IDs *as well as all pairwise interactions*. This model’s predictions are overlayed in red on the translational strengths in Figure 4.2.

The model using only main effects has a multiple R^2 value of 0.47 and a residual standard error of 0.48 while the model with interactions has a multiple R^2 value of 0.81 and a residual standard error of 0.39. This indicates that the model using interactions can explain a much greater fraction of the observed variance while also improving prediction accuracy. Interestingly, the greatest areas of improved performance by adding interactions appear to be the extremes of translational strength, suggesting that interaction terms dominate in low expression and high expression cassettes.

4.5.3 ANOVA for Translational Characterization Data

In order to understand the sources of variation within the translational characterization assay, ANOVA was used to decompose the variance and calculate the sequential sum of squares for every predictor in the pairwise interaction model. Each predictor’s sum of squares was then divided by the total sum of squares and multiplied by 100 to achieve that predictor’s contribution to the observed variance in translational strength. The results of this analysis are shown in Figure 4.3. The statistical model for translational strength explains 80.6% of the observed variance, while the analogous model for transcriptional strength (Figure 3.3) only explained 58.5% of the observed variance. Promoter, RBS, and ORF main effects explain nearly half the variance in translational strength, but interestingly the single predictor that explained the greatest fraction of the observed variance was the interaction between the promoter and the RBS. This term—along with the interaction between the RBS and the ORF—explain nearly a third of the observed variance, indicating that interactions significantly affect protein levels.

4.6 Selection of Final RACCET Promoter-RBS Pairs

The goal of translational characterization was to identify a set of promoter/RBS combinations which would span the observed range of translational strengths while remaining as robust as possible with respect to ORF-specific effects. This was accomplished by manual curation according to the following procedure. To prevent the selection of promoter/RBS pairs with low-confidence measurements of translational strength, cassettes who appeared above background for only one ORF were discarded from consideration. For all remaining cassettes, the ORF-specific percentile of translational strength was calculated. Promoter/RBS combinations which had significantly different percentile rankings between all three ORFs were discarded from consideration. Next promoter/RBS pairs were placed into approximately log-linear bins spanning the observed translational strength range according to their mean observed translational strength. In each bin, the promoter/RBS pairs which produced

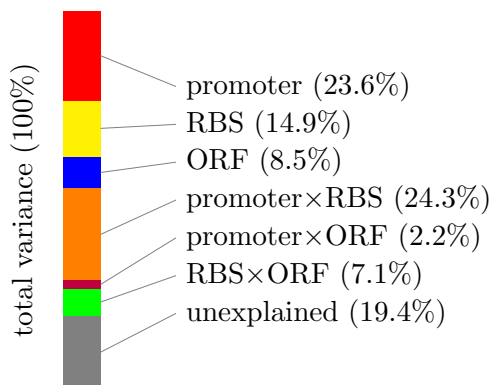


Figure 4.3: Promoter, RBS, and ORF effects explain nearly half of the variance in protein expression. Interactions—especially between the promoter/RBS and RBS/ORF—significantly effect protein levels. Overall, a linear statistical model of translation explains more variance than a similar model of transcription (Figure 3.3).

significantly different levels of translational strength across assayed ORFs were discarded. The final set of characterized promoter/RBS pairs was then selected from the remaining candidates by choosing the most robust (in terms of compared ORF percentile rankings and translational strengths) 15 pairs spanning the observed range of translational strengths. The promoter/RBS pairs representing the final RACCET set are shown in Table 4.1.

4.7 Effect Contributions for Final RACCET Promoter/RBS Pairs

To determine the contributions of promoter, RBS, and their two-way interaction on translational strength for the final RACCET pairs, a linear multivariate regression model was built using the translational characterization data for all 654 assayed strains. This model predicts the log-transformed translational strength of a cassette using the main effects of promoter and RBS as well as their interaction. The observed log transformed mean translational strength of the cassettes, their predicted log-transformed translational strength using this model, and the effect sizes of the each predictor are shown in Figure 4.4.

Importantly, the model predictions of translational strength with no information on ORF match the observed mean translational strengths very well for all chosen promoter/RBS pairs. This indicates that the selected pairs are relatively robust to ORF effects for the ORFs screened in this approach. Furthermore, the importance of promoter/RBS interactions in explaining the translational strength of assayed expression cassettes is made clear by their relatively large effect sizes. A case which particularly highlights this point is the PR12.92.A/RBS7 pair. The main effects for promoter and RBS alone would predict a much weaker expression level than what is actually observed. Instead a synergy between these parts results in a much higher level of translational strength.

Our results contradict a common assumption in synthetic biology: high expression is achieved when a strong promoter is combined with a strong RBS. Instead, the interaction between the promoter and RBS dominate in our data. Predicting the strongest or weakest gene expression

Combination		Overall		mNeonGreen		mTurquoise2		mTagBFP2	
Promoter	RBS	Strength	SD	Strength	Percentile	Strength	Percentile	Strength	Percentile
PR10.77.Δ	RBS6	0.082	0.05	0.12	2.7	NA	NA	0.046	3.3
PR13.71.T	RBS20	0.26	0.010	0.25	9.7	NA	NA	0.27	20
PR13.31.A	RBS2	0.32	0.18	0.35	14	0.13	17	0.49	37
PR10.77.Δ	RBS7	0.49	0.089	0.43	17	NA	NA	0.56	42
PR6.65.T	RBS2	0.54	0.24	0.36	14	0.46	47	0.81	55
PR4.52.G	RBS13	0.64	0.19	0.64	25	0.84	57	0.45	33
PR12.33.C	RBS1	0.73	0.12	0.86	33	0.61	52	0.72	51
PR9.59.A	RBS11	1.0	0.19	1.2	46	0.89	59	0.90	58
PR13.31.A	RBS7	1.3	0.30	1.7	58	1.3	67	1.1	63
PR6.65.T	RBS7	1.6	0.46	2.1	64	1.6	73	1.1	65
PR12.99.T	RBS3	2.3	0.89	3.3	78	1.7	76	1.8	81
PR12.92.A	RBS7	2.6	1.0	2.6	71	3.6	89	1.6	78
PR12.33.C	RBS7	3.5	0.89	4.2	83	3.8	90	2.5	87
PR4.12.Δ	RBS9	9.6	2.3	11	96	11	99	6.9	98
PR4.28.G	RBS4	35	40	80	100	23	100	2.3	86

Table 4.1: Fifteen promoter/RBS pairs were selected for the final RACCET set. Expression strength is the ratio of fluorescence relative to the mScarlet-I control. The combinations were selected to span a wide range of expression strengths while showing little dependence on the ORF.

cassettes requires a model with interactions, and such a model requires combinatorial experiments as training data.

Figure 4.4 highlights the difficulty in predicting the “middle” of the expression continuum. Because promoter/RBS interactions are so strong, the interaction can easily overwhelm the individual strengths of the promoter or RBS. By sampling all possible promoter/RBS/ORF combinations, RACCET offers experimentally-validated options for engineers who need moderate levels of expression.

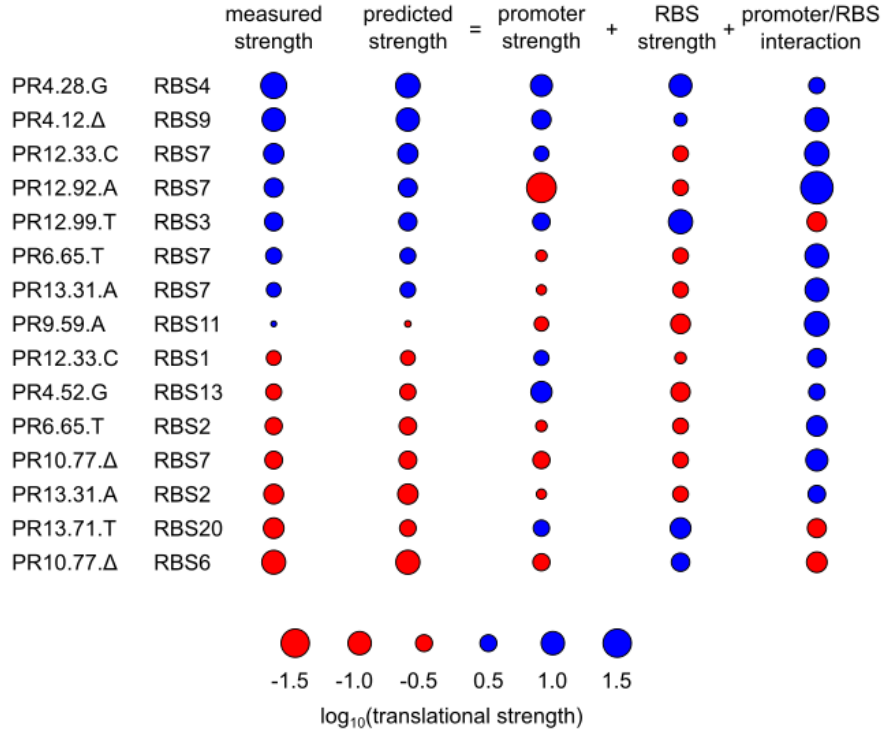


Figure 4.4: Fifteen characterized promoter/RBS combinations were selected as the final RACCET set in *S. mutans*. The combinations span a 426-fold range of gene expression strength as measured by fluorescent reporter assays. The promoter, RBS, and interaction strengths are the effect sizes in a linear model. The reference level for the promoters is the weakest of the final characterized promoter (Table 3.1), and the reference level for the RBS effect is the negative control RBS10.

Part II

Measuring gene expression

Chapter 5

Splintlock-seq: Scalable and targeted gene expression profiling

5.1 Overview

The considerable time and cost associated with prokaryotic, transcriptomic library preps have limited the scope of gene expression studies in bacteria. The absence of polyadenylated mRNA prevents transcript enrichment and requires rRNA depletion prior to cDNA library generation. Experimental workflows developed to address these challenges are labor and time intensive. The experimental workflow accompanying one major, commercially available RNA-seq library prep kit requires no less than 89 steps spread across 12 sub-protocols for a single library. In general, the quality of transcriptomic data is more dependent on the number of replicate libraries than the depth of sequencing [50]. For this reason, 4-6 biological replicates of each condition are recommended when designing transcriptomic experiments. The most basic RNA-seq experiment with one control group and one experimental group, would require the prep of 8-12 libraries. Most non-profit, institutional DNA sequencing cores offer prokaryotic RNA-seq library prep services for roughly \$250 per library [51] depending on scale. For the minimal RNA-seq experiment outlined above, this would mean a total cost of \$3,000 to \$5,000 dollars for library prep alone. Advances in sequencing technology, DNA barcoding, and computational demultiplexing have resulted in the cost of per library sequencing decreasing dramatically [52]. Thus, sequencing each of these libraries at the recommended depth can cost \$20 or less, representing roughly 8% of the library prep cost. That said, many fundamental biological processes cannot be addressed by a simple two group experimental framework. Metabolic gene expression profiling on different carbon or nitrogen sources, for example, would require multiple experimental groups. Furthermore, studies attempting to profile stress responses usually need high-resolution time series data. Without a low cost, prokaryotic, transcriptomic library prep these types of studies are labor and cost prohibitive.

While RNA-seq remains the most commonly performed transcriptomic analysis, several studies have sought to address the labor and cost issues associated with RNA-seq library preps [53].

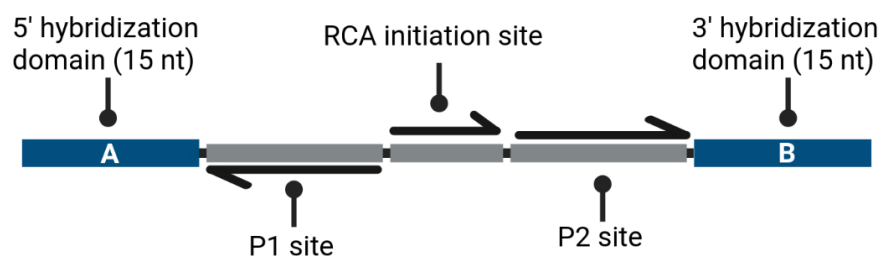


Figure 5.1: Typical architecture of a padlock probe. Variable 15-nt sequences capable of hybridizing to a target 30-nt stretch of mRNA are found at the 5' and 3' ends of the ssDNA molecule. The common backbone sequence often contains engineered sites for future amplification with PCR and/or isothermal rolling circle amplification with Φ 29 DNA polymerase.

RNA tag-seq reduces the library prep procedure to 16 steps and saves on total cost by enabling the pooling of libraries prior to rRNA depletion. However, the experimental workflow for this method remains complex and utilizes multiple, inefficient single stranded ligations and purification steps [54]. RASL-seq eliminated the need for adapter ligations by utilizing SplintR ligase and pairs of DNA probes to quantitatively profile gene expression. In this approach, pairs of DNA probes hybridize adjacent to one another on the target mRNA and are ligated together by SplintR ligase [55]. Subsequent amplification and tagging can be accomplished by making use of known sequences in the probe tails. However, RASL-seq suffers from high levels of noise at low target RNA concentrations and requires manual probe design. The latter requirement has limited the scope of RASL-seq studies to a small subset of the transcriptome. We aim to improve upon RASL-seq library prep by building an automated, computational pipeline capable of designing thousands of probes to target the entire transcriptome of a particular organism. In contrast to paired probed, we will use padlock probes, whose architecture improves assay sensitivity as well as specificity and allows for unbiased signal amplification. In addition, the proposed approach, referred to henceforth as Splintock-seq, offers a simple, three-step, two-pot transcriptomic library prep with a per library cost of less than \$10.

Chorella virus DNA ligase, also known as SplintR ligase, catalyzes the formation of phosphodiester linkages between adjacent residues of single stranded DNA molecules hybridized on RNA [56]. SplintR ligase exhibits minimal bias for nucleotide identity at the ligation junction while demonstrating a high degree of sensitivity to ligation junction mismatches [57]. These qualities have led to the use of SplintR ligase in highly specific detection assays targeting miRNA isoforms [58], SNPs [57], and splice variants [59]. Assays employing SplintR ligase have demonstrated remarkable sensitivity. One recent study showed that SplintR ligase could be used in conjunction with qPCR to specifically detect attogram quantities of a specific miRNA from within 10 ng of total RNA [58]. A padlock probe is a single stranded, linear DNA oligonucleotide that contains sequence complementary to a target molecule at the 5' and 3' ends, with a constant non-hybridizing backbone in between. A typical padlock probe architecture is shown in Figure 5.1. When a padlock probe binds to its

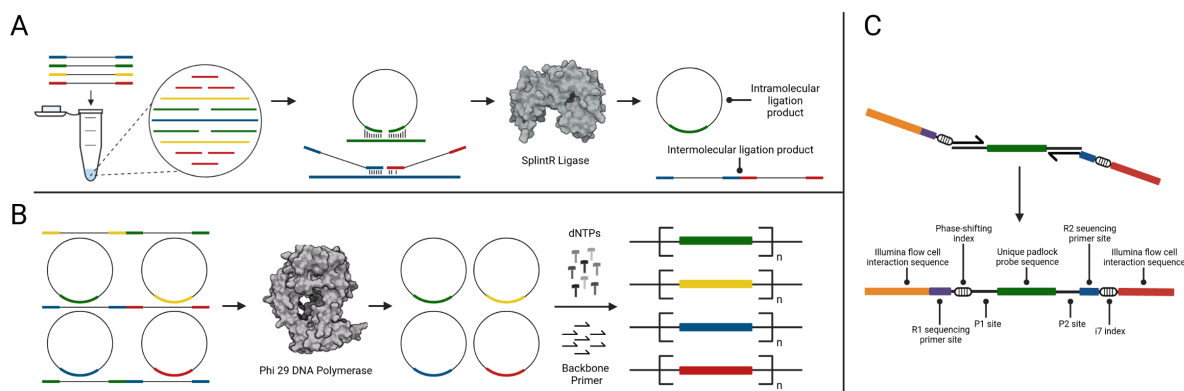


Figure 5.2: Overview of a Splintlock-seq library prep. **A.** A SplintR ligation reaction whereby padlock probes hybridized to mRNA targets are converted from linear, ssDNA molecules into circular ssDNA molecules. **B.** Φ 29 DNA polymerase is used to remove remaining single stranded nucleic acids from the ligation reaction and to amplify the information stored in the circularized padlock probes with rolling circle amplification. **C.** The rolling circle amplification product is converted into dual-indexed sequencing libraries compatible with TruSeq technology.

target, the two ends hybridize directly adjacent to one another. The ends can then be ligated together using an enzyme such as SplintR ligase, creating a circular ssDNA molecule in the process. Padlock probe-based detection of nucleic acids exhibits greater targeting specificity and sensitivity than paired probe strategies due to the cooperative annealing dynamics of the two probe ends[60]. Furthermore, circularized ssDNA molecules can be efficiently used as template for isothermal, rolling circle amplification with Φ 29 DNA polymerase. Φ 29 DNA polymerase exhibits very low bias, high fidelity, extremely high processivity, and strong multiple strand displacement activity [61]. By adding a primer targeting the padlock backbone, Φ 29 DNA polymerase can be used to initiate a rolling circle amplification, selectively amplifying intramolecular ligation products. This improves signal-to-noise ratios as non-specific intermolecular ligation products and unreacted probes will theoretically not be amplified. The degree to which rolling circle amplification boosts signal is remarkable. At 30 °C, Φ 29 DNA polymerase is capable of incorporating 2280 nt/min. The enzyme's average processivity is greater than 70 kb, indicating that a single 90 nt circularized padlock probe can be used as template to produce a concatamer containing over 775 copies in a little over a half an hour [62]. The sensitivity and high signal-to-noise ratios of padlock probe-templated rolling circle amplifications have been widely applied in the field of molecular diagnostics [63].

In this thesis, we introduce Splintlock-seq. A low-cost, quantitative transcriptomic library prep based on the activity of SplintR ligase and Φ 29 DNA polymerase on padlock probes. An overview of our approach to generating Splintlock-seq NGS libraries is shown in Figure 5.2. First, padlock probes designed with an automated pipeline are hybridized to mRNA transcripts representing a large subset of the target transcriptome. Next, padlock probes hybridized on their mRNA target are circularized by SplintR ligase. Off-target intermolecular ligation products are removed by adding Φ 29 polymerase without any dNTPs. After exonuclease treatment, dNTPs and an

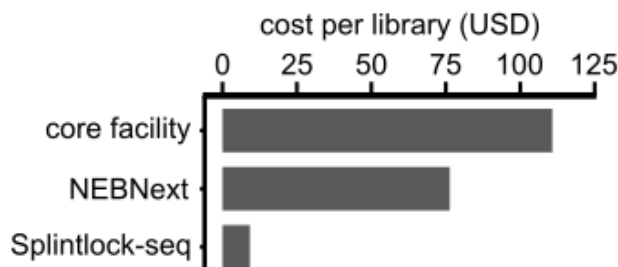


Figure 5.3: Bacterial transcriptomic library prep costs for traditional technologies and Splintlock-seq, from RNA isolation to final DNA purification. Data shown for core facilities assume a scale of >96 samples. A single Splintlock-seq library prep costs \$9.21, 12-fold cheaper than paying a sequencing core, and 8.4-fold cheaper than preparing libraries with NEBNext kits.

initiation primer are added to begin rolling circle amplification. This results in the formation of large quantities of concatemic ssDNA originating from the circularized padlock probes. Finally, the concatamer product is amplified with primers which add Illumina adapter sequences, making the library compatible with TruSeq technology. Importantly, a single Splintlock-seq library prep, starting from RNA isolation through the purification of the final sequencing library, costs only \$9.21, a mere fraction of the cost for current bacterial RNA-seq library preps (see Figure 5.3).

5.2 Automated design of padlock probe pools

Molecular probe-based approaches to measuring gene expression changes have traditionally been limited in their scope by the requirement that probes be manually designed. Previous studies using ligation-dependent strategies, like RASL-seq, have thus been limited to analyzing gene expression data for at most a few hundred transcript targets [55], [64]. As our goal is to profile gene expression changes across the entire annotated transcriptome of *S. mutans*, an automated approach is required. This was accomplished by building a custom, generalizable padlock probe design pipeline in R. The general approach implemented by this design pipeline is shown in Figure 5.4. The program accepts as an input a standard GenBank file (.gb or .gbk) containing annotation information for the microbe of interest. For bacteria with multiple chromosomes or associated extrachromosomal DNA, more than one GenBank file can be sequentially used as input.

The first step in probe design is the identification of off-limit sequences and candidate target sequences. A data structure containing off-limit sequences is created by extracting ribosomal RNA (rRNA) and transfer RNA (tRNA) sequences using the GenBank file(s) feature annotations. Similarly, a separate data structure is created to house sequence information (both sense and antisense) for every gene that is not an rRNA or tRNA gene. Therefore, protein-encoding genes and non-coding RNA genes are both considered as targets for probe design. Next, antisense sequences in both data structures (being the sequences capable of hybridizing to a given transcript) are hashed into k-mers, where k is the padlock probe target sequence length. By default, a total target sequence length of 30 nt is used, corresponding to two 15 nt targeting domains at the 5' and 3' end of each

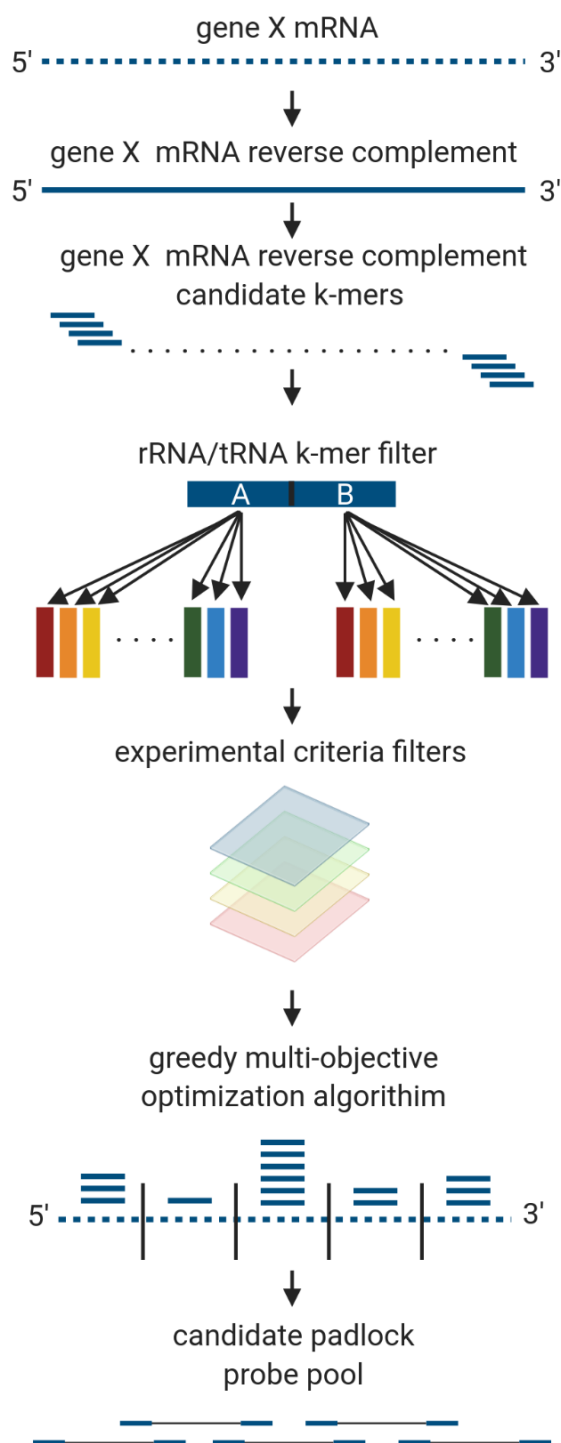


Figure 5.4: Overview of the top-down, automated padlock probe design pipeline. The pipeline starts with the set of unique mRNA target k -mers and gradually removes candidate sequences according to several filters including sequence affinity for rRNA and tRNA molecules and probe annealing temperature. A greedy, multi-objective optimization algorithm is then implemented to select the "best" set of padlock probes to target all remaining genes.

padlock probe. In the following text, candidate k-mers or candidate sequences refer to this 30-mer. For every gene target, any candidate k-mer which appears in the set of unique k-mers derived from rRNA and tRNA elements is immediately eliminated from consideration. Candidate k-mers which are not unique to a particular transcript are also eliminated.

Next, all candidate sequences were screened for theoretical binding affinity to off-limit sequences. For each candidate sequence, a fuzzy matching algorithm calculates the minimum Hamming distance between both constituent 15-mers and every unique 15-mer in the off-limit sequence set. If either of a candidate sequence’s constituent 15-mers are less than a user-defined edit distance from any of the off-limit sequences, that probe is eliminated from consideration. A global pandemic left me with a lot of time to think about the selection of this thresholding parameter. The result of those efforts is the mathematical derivation housed in Appendix A. Using that derivation, a threshold value of 3 was selected. This ensured candidate sequences were likely to miss the off-limit sequences while also retaining enough candidate sequences for future filtering.

Candidate k-mers were then put through the same fuzzy matching algorithm as described above but this time using hashed sequences derived from the reverse complement of the padlock probe backbone sequence. This ensures that padlock probes designed using this pipeline are not likely to efficiently bind to the backbone of other padlock probes during ligation reactions. While this sort of probe-probe hybridization scheme is unlikely to produce SplintR ligation products because both sequences are DNA, probe-probe interactions would soak up probes which might otherwise hybridize to their intended target.

Next, a series of experimental criteria filters were imposed on the remaining candidate probe sequences. Any sequence containing the recognition sites of BtsI or BspQI was discarded. The rationale behind this choice will be made evident in Section 5.4. Each of the component 15-mer sequences was also screened for the existence of stable secondary structures using an R interface to primer3. If stable secondary structures were identified, the candidate sequence was discarded. Any sequence containing homopolymer tracts longer than 4 nucleotides within the most interior 12 nucleotide stretch of the candidate sequence were discarded. This particular subdomain of the candidate sequence is referred to as the ligation junction and is composed of the six terminal nucleotides on each end of the padlock probe that come together on the target RNA. The melting temperatures of both component 15-mer sequences were calculated using the primer3 interface and candidate sequences whose 15-mers had a difference in melting temperature greater than 5 were discarded. Furthermore, any candidate sequence whose 15-mers failed to have a minimum melting temperature of 31°C was discarded. Finally, any candidate sequence with the most interior 2 nucleotides (referred to as the ligation pair) being GC were discarded. While having low bias overall, SplintR ligase does have a slight bias (79% ligation efficiency vs 95% ligation efficiency) against ligating adjacent DNA ends if the identity of nucleotide at 5’ end is a G and the identity of the nucleotide at 3’ end is a C [57].

For each transcript target, the smallest bin size to ensure uniform coverage by the remaining candidate sequences was calculated. A greedy multi-objective optimization algorithm was then

built and employed to select the "best" global set of candidate sequences by selecting the "best" local candidate sequence inside each bin. This was done by locally ranking each candidate sequence according to six selection criteria: the melting temperatures of the component 15-mers (higher melting temperatures are preferred), the melting temperature difference between component 15-mers (lower differences are preferred), ligation efficiency of the ligation pair (higher efficiency is preferred), ligation kinetics of the ligation pair (faster enzymatic rates are preferred), the maximum hexamer frequency score (lower scores are preferred), and the summed hexamer frequency score (lower scores are preferred). The first two criteria are readily understandable. The ligation efficiency and kinetics criteria were acquired from previous studies investigating the effects of nucleotide identity at the ligation junction [57]. No major biases were revealed from this study but small preferences for ligation junction nucleotides were identified and that information was incorporated here as a fine tuning parameter in probe selection. In order to ensure the candidate sequence is targeting a stretch of its transcript that is as unique as possible with regards to all other transcripts, a hexamer frequency approach was taken. The entire coding portion of the *S. mutans* UA159 genome, was hashed into hexamers and tabulated. The candidate sequence was then hashed into all component hexamers. The maximum frequency score for a candidate sequence is the number of times the most common component hexamer was observed within the coding regions of the *S. mutans* UA159 genome. The summed hexamer frequency score for a candidate sequence is simply the number of times all component hexamers were observed within the coding regions of the *S. mutans* UA159 genome. For each bin on each transcript, all remaining candidate sequences were ranked according to each of these criteria. A composite score was then calculated for each candidate sequence by taking the unweighted mean of all the criteria rankings. The probe with the best composite score for each bin was then selected for inclusion in the final pool. All candidate sequences were then converted into padlock probe sequences by concatenating the 5' hybridization 15-mer with the common backbone sequence (60 nt) and the 3' hybridization 15-mer. Padlock probe sequences were then converted into the form required by the padlock probe synthesis pipeline as described in Section 5.4. This was achieved by concatenating a universal 5' adapter sequence (AGGGAGAATGCACGCAGTG) with the reverse complement of the padlock probe sequence and a universal 3' adapter sequence (CGAAGAGCTGACTTGCGTC). The results of this operation were output as a simple CSV file.

This padlock probe design pipeline has been successfully used to design padlock probe pools for *S. mutans* UA159, *S. sobrinus* 6715-15 and *B. thailandensis* E264.

5.3 *S. mutans* UA159 Padlock Probe Pool v.1 Summary

The initial *S. mutans* padlock probe pool contained 12,007 padlock probes targeting 1,882 genes. Padlock probes were not able to be designed for 18 genes. Analysis of these "drop out" genes revealed that 12 of the 18 genes were associated with mobile genetic elements. The requirement that every candidate sequence be unique eliminated all padlock probes for any gene with multiple copies. Thus, these 12 genes immediately dropped out of consideration for padlock probe design.

As current transcriptomic analyses generally disregard reads mapping equally well to multiple genomic positions, this was not of great concern. The remaining six drop out genes were a 105 bp/35 AA pseudogene (SMU RS00345), a 98 bp ncRNA gene (SMU RS01520), a 204 bp/68 AA copper chaperone gene (SMU RS02080), two bacteriocin-like genes with lengths of 231 bp/77 AA (SMU RS02065) and 144 bp/48 AA (SMU RS09965), and one unknown integral membrane protein with a length of 138 bp/46 AA (SMU RS02065). All of these genes are short and thus had a small number of candidate sequences to begin with. Thus, any transcriptomic profiling accomplished with Splintlock-seq excludes these genes from consideration.

5.4 Padlock Probe Pool Synthesis

Massively parallel DNA synthesis technologies originally developed for DNA microarrays are capable of constructing large libraries consisting of more than 10,000 different oligonucleotide sequences. However, the amount of each oligonucleotide produced with these technologies (less than 1 fmol) and the formats required for synthesis limit their utility. In order to convert the small quantities of ssDNA derived from these synthesis techniques into large quantities of ssDNA padlock probes, we slightly modified the approach outlined by Murgha et al [65]. Briefly, sequences output by the automated probe design pipeline were synthesized on a DNA microarray (Custom Array). The resulting pool was then used as template in 96, 10 μ l 15-cycle PCR amplifications with primers targeting the universal adapter sequences. The primer targeting the 5' universal adapter has a non-hybridizing tail that adds sequence containing a BtsI recognition site and a T7 promoter. The primer targeting the 3' universal adapter sequence has a non-hybridizing tail that adds sequence containing a BspQI recognition site. Amplicons were purified and used as template for an IVT reaction using the HiScribe T7 High Yield RNA Synthesis Kit (NEB). The IVT product was purified and used as template for a reverse transcription reaction with the Protoscript II First Strand cDNA Synthesis kit (NEB). The primer used in this reaction targets the 3' universal adapter sequence and contains a biotin modification. The RNA is then subjected to alkaline hydrolysis and the cDNA is purified. cDNA—containing the padlock probe sequence flanked by universal adapters—is then used in a hybridization reaction with short oligonucleotides targeting the BtsI and BspQI recognition sites. The hybridization product is pulled down using hydrophobic streptavidin-covered magnetic beads (NEB) and sequentially digested on-bead first with Nb.BtsI and then with Nt.BspQI. Both of these enzymes are Type IIS nicking-variants. After the second digest, the adapter-free padlock probe pool enters the supernatant and can readily be purified using silica spin columns (Thermo Fisher Scientific).

5.5 Identification of highly non-specific probes in v.1 probe library

The fundamental challenge in developing Splintlock-seq is the simultaneous need for an optimized library prep protocol and an optimized padlock probe library. Of course,

one can't really be achieved without the other. In order to begin addressing this challenge, we recognized that the padlock probe design pipeline was likely not perfect, and that our current padlock probe pool probably contained some "problematic" probes. That is to say probes which generate non-specific signal, either due to hybridization with the rRNA or in the absence of any RNA. Because roughly 95% of the RNA within bacterial cells is rRNA and tRNA [66], [67], probes capable of efficiently hybridizing to the rRNA would generate large amounts of non-specific signal that would eat up sequencing depth and obscure the behavior of target-specific padlock probes. Therefore, the initial goal for Splintlock-seq was to determine if these probes exist, and if they do, to eliminate them from the padlock probe library.

5.5.1 Experimental design and methods

In order to identify problematic probes within the padlock probe library, sequencing libraries were prepared according to the initial protocol described in Section 5.5.1 using four different RNA inputs. The first input was total RNA isolated from *S. mutans*. The second was *S. mutans* total RNA with an rRNA depletion step performed using a MICROBExpress mRNA Enrichment kit (Invitrogen) according to the manufacturer's instruction. These two template types were optimistically used to see if on-target signal could be generated at equivalent levels regardless of the presence of rRNA. The third was a synthetic rRNA mixture. This mixture was derived by amplifying the 5S, 16S, and 23S genes off *S. mutans* genomic DNA using primers that add a T7 promoter to the 5' end of the amplicon. Those amplicons were purified and used as templates in IVT reactions with the HiScribe T7 High Yield RNA Synthesis Kit (NEB). The synthetic rRNA molecules were purified and pooled in an equimolar manner to create the final mix. Henceforth, anytime a Splintlock-seq library is said to be prepared using rRNA as template, it is referring to this synthetic mixture. The rRNA-templated library was included in order to identify probes that generate non-specific signal due to hybridization on rRNA. The fourth was a no RNA condition to detect probes who generated signal in the complete absence of RNA template.

Splintlock-seq library prep protocol v.1

A "starting point" Splintlock-seq library prep protocol was created relying heavily on manufacturer's recommendations for SplintR Ligase and literature reports for Φ 29 DNA polymerase used in conjunction with padlock probes [63], [68], [69]. However, all manufacturer recommendations and literature-derived protocols regarding ligation and rolling circle amplification were for small numbers of probes targeting small numbers of RNA templates. Therefore, several approximations regarding stoichiometry had to be made. These resulted in the initial library prep protocol below.

On ice, combine 500 ng of RNA template and 0.5 ng of padlock probe library in a sterile 200 μ L PCR tube. Add 2 μ L of 10x SplintR ligase buffer (NEB) and nuclease free water to a final volume of 19.5 μ L. Mix by pipetting. Heat the mixture at 80°C for 2 minutes followed by a -5°C/min ramp to 25°C using a preheated thermocycler to allow probes to hybridize to the RNA. Spike in 0.5 μ L of SplintR ligase (NEB, 25 units/ μ L), mix by pipetting, and incubate at 30°C for 1 hour. Heat kill the

ligase at 65°C for 20 minutes and put the reaction on ice. In a separate tube, combine 5 μ l of the ligation reaction with 6 μ l of 10x Φ 29 DNA polymerase buffer (NEB), 1.2 μ l of exonuclease-resistant RCA initiation primer (50 uM), 0.4 μ l of BSA (NEB, 20 mg/mL), 3 μ l dNTPs (NEB, 10mM each), 43.4 μ l of nuclease free water, and 1 μ l of Φ 29 DNA polymerase. The reaction is mixed by pipetting and then incubated at 30°C overnight. The next morning, heat kill Φ 29 DNA polymerase at 65°C for 10 minutes. The rolling circle amplification product is then converted into a sequencing library by the following PCR reaction. On ice, combine 0.5 μ l of RCA product, 2.5 μ l each of 10 uM stocks of the custom Illumina adapter-adding primers described in Section 3.11, 25 μ l of NEBNext UltraII Q5 Master Mix, and 19 μ l of nuclease free water. After 20 cycles of amplification, purify the 200 bp amplicons with 1.25x AMPure SPRI beads and elute in 20 μ l of nuclease free water. These dual-indexed libraries can then be pooled according to sequencing core specifications and ran on a single SP 1x100nt lane using an Illumina NovaSeq 6000, or any sequencing machine compatible with TruSeq technology.

Splintlock-seq NGS data processing

FASTQ files demultiplexed by core facilities according to the i7 index were demultiplexed again based on the custom phase-shifting barcode using `fastx.barcode.splitter` (1 mismatch allowed) and trimmed to only include the 30 nucleotides immediately downstream of the P1 site using `fastx.trimmer`. These 30 nucleotides represent the transcript sequence targeted by the associated padlock probe. The transcript sequences are then aligned to the *S. mutans* UA159 genome using `bowtie 1` with default settings [70]. The resulting SAM files are then read into a custom Python script that tabulates the observed read-index mappings. The resulting index count table is then output as a CSV and read into R for analysis. A key is generated using all observed index mappings for an experiment and the known indices of probe target sites. This key then allows index counts to be mapped to probe counts. Probe counts are then normalized into counts per million to account for library size effects.

5.5.2 Experimental results

The results of this experiment are shown in Figure 5.5. Probes whose count frequencies were highly correlated between total and rRNA-depleted libraries accounted for the majority of total signal and were also heavily represented in the rRNA-templated library. This indicates that the heavily represented probes in cases with mRNA present were not generating signal due to their target, but rather due to non-specific ligation on rRNA. Interestingly, some of these high frequency probes were also capable of generating high levels of signal in the absence of RNA all together. If a particular probe generated high amounts of signal in either the rRNA or no RNA cases they were labeled as "problem probes". Of the 12,007 padlock probes generated by the automated design pipeline, 1004 "problem probes" were identified and eliminated from consideration. This eliminated probes for an additional 12 genes: SMU RS01065 (hypothetical protein gene), SMU RS01135 (hypothetical protein gene), SMU RS03430 (hypothetical protein gene), SMU RS03465 (hypothetical protein

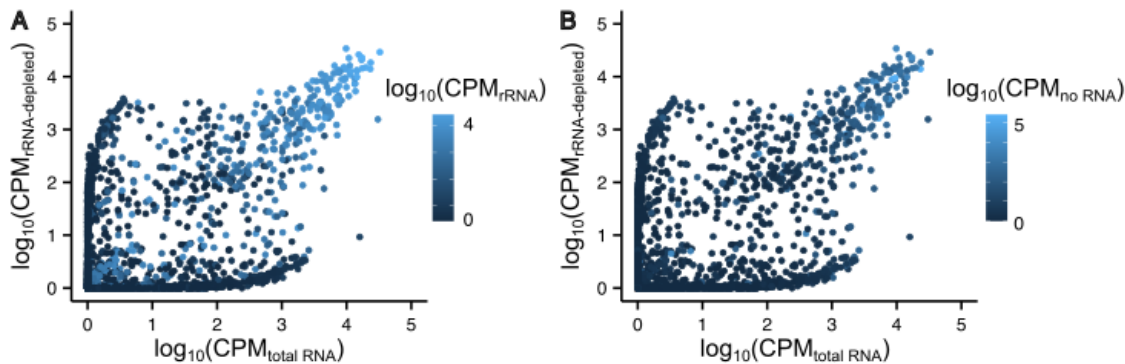


Figure 5.5: Non-specific probes within the initial padlock probe pool were identified by sequencing Splintlock-seq libraries prepared on total RNA, rRNA-depleted total RNA, rRNA only, or no RNA. **A.** shows the normalized probe count comparison between total RNA and rRNA-depleted libraries with data points colored to reflect their normalized abundance in an rRNA-templated library. **B.** shows the same data but with data points colored to reflect their normalized abundance in a library prepared using no RNA.

gene), SMU RS03545 (hypothetical protein gene), SMU RS04420 (*rplL*, ribosomal protein gene), SMU RS05320 (hypothetical protein gene), SMU RS07850 (*greA* gene), SMU RS08990 (putative transcriptional regulator gene), SMU RS09165 (rl30 ribosomal protein gene), SMU RS09400 (*ssrS* ncRNA gene, overlapping with tRNA-Lys gene), and SMU RS09920 (ncRNA overlapping with ribosomal protein gene). A second padlock probe library was synthesized without these probes, termed *S. mutans* UA159 padlock probe library v.2.

5.6 Splintlock-seq stoichiometry screen

With highly non-specific probes eliminated, we decided to turn our attention back to stoichiometry concerns. An additional motivating factor in this was that the libraries sequenced for the identification of non-specific probes all had a relatively low fraction of aligned reads (40-50%). This was indicative of an unoptimized library prep protocol.

5.6.1 Experimental design and methods

In order to determine the best stoichiometry for SplintR ligations with respect to total RNA and padlock probe library, a two-factor, three-level factorial design was adopted. Libraries were prepared with three different amounts of total RNA and three different amounts of padlock probe library. In order to sample a large portion of the stoichiometric space, factor levels were chosen to be spaced by two orders of magnitude. RNA levels were set at 0.5, 50, or 5000 ng; while padlock probe library levels were set at 0.001, 0.1, or 10 ng. The response metrics chosen to evaluate ligation stoichiometry performance were the percentage of NGS reads successfully aligned to the *S. mutans* genome and the Pearson correlation coefficient for probe frequency between replicate libraries.

5.6.2 Splintlock-seq library prep protocol v.2

An analysis of the unaligned reads from sequenced libraries revealed that about 6% of unaligned reads were correct intramolecular ligation products but originated from probes missing a single nucleotide at the 5' end or the 3' end. This resulted in a gapped read that **bowtie** was not capable of mapping to the genome. About 15% of unaligned reads were the result of the 3' universal adapter sequence not being removed during probe library synthesis, see Section 5.4. To address this, the amount of Nb.BtsI added and the duration of digestion were increased during probe library synthesis. Interestingly, about 55% of the unaligned reads were the result of chimeric sequences containing the 5' hybridization sequence of one probe and the 3' hybridization sequence of another probe. Thus, the majority of unaligned reads in surveyed libraries were due to high instances of intermolecular ligation. This led to two major changes in library prep protocol. First, the amount of enzyme added to the library prep protocol was decreased by a factor of 5 (2 μ l of a 1:20 dilution of SplintR ligase into NEB Diluent A). Second, an exonuclease treatment was added prior to rolling circle amplification. Due to Φ 29 DNA polymerase's high exonuclease activity [69], this modification simply required that dNTPs and the RCA initiation primer be withheld from the reaction for a set amount of time. qPCR experiments with primers targeting the backbone of probes and the ligation junction demonstrated that highly efficient removal of linear DNA can be accomplished in only 15 minutes and that circularized padlock probes are protected from exonuclease activity. These two changes are reflected in the second version of the Splintlock-seq library prep shown below. RNA and padlock probe library amounts are left undetermined to reflect the nature of the stoichiometry screen.

On ice, combine the desired amount of RNA template and desired amount of padlock probe library in a sterile 200 μ l PCR tube. Add 1 μ l of 10x SplintR ligase buffer (NEB) and nuclease free water to a final volume of 10 μ l. Mix by pipetting. Heat the mixture at 80°C for 2 minutes followed by a -1°C/min ramp to 25°C using a preheated thermocycler to allow probes to hybridize to the RNA. To this, add 10 μ l of a prepared ligase solution consisting of 1 μ l of 10x SplintR ligase buffer (NEB), 2 μ l of 1:20 diluted SplintR ligase (NEB, in Diluent A), and 7 μ l of nuclease free water. Mix by pipetting, and incubate at 30°C for 1 hour. Heat kill the ligase at 65°C for 20 minutes, denature RNA template from ligated padlock probes by heating at 95°C for 5 minutes, and put the reaction on ice. In a separate tube, combine 5 μ l of the ligation reaction with 5 μ l of 10x Φ 29 DNA polymerase buffer (NEB), 0.5 μ l of BSA (NEB, 20 mg/mL), 39 μ l of nuclease free water, and 0.5 μ l of Φ 29 DNA polymerase. The reaction is mixed by pipetting and then incubated at 30°C for 30 minutes. To this, add 10 μ l of prepared rolling circle amplification solution consisting of 4 μ l of nuclease free water, 1 μ l of 10x Φ 29 DNA polymerase buffer (NEB), 1.5 μ l of 50 μ M exonuclease-resistant RCA initiation primer, 3 μ l of dNTPs (NEB, 10mM each) and 0.5 μ l of Φ 29 DNA polymerase (NEB). Incubate the reaction at 30°C overnight. The next morning, heat kill Φ 29 DNA polymerase at 65°C for 10 minutes. The rolling circle amplification product is then converted into a sequencing library by the following PCR reaction. On ice, combine 5 μ l of 1:10 diluted RCA product (in nuclease free water), 2.5 μ l each of 10 μ M stocks of the custom Illumina adapter-adding primers described in Section 3.11, 25 μ l of NEBNext UltraII Q5 Master Mix, and 15 μ l of nuclease

free water. After 20 cycles of amplification, purify the 200 bp amplicons with 1.25x AMPure SPRI beads and elute in 20 μ l of nuclease free water. These dual-indexed libraries can then be pooled according to sequencing core specifications and ran on a single SP 1x100nt lane using an Illumina NovaSeq 6000, or any sequencing platform compatible with TruSeq technology.

5.6.3 Experimental results

Splintlock-seq libraries were made using the 9 different stoichiometric conditions discussed above and then replicated to produce a total of 18 libraries for sequencing. The results of this experiment are shown in Figure 5.6. The percentage of aligned reads depends heavily on the amount of RNA, while the correlation between replicates appears to depend primarily on the amount of padlock probe. To quantify these trends, a multivariate linear regression model was built to predict either the aligned read percentage or the Pearson correlation coefficient between replicate libraries. These models made predictions using the level of RNA, the level of padlock probe library, and the interaction term between RNA and padlock probe library.

The read alignment model was highly predictive and explained a large fraction of the observed variance ($R^2=0.79$) in read alignment percentage. This model confirmed that read alignment percentage solely depends on the amount of RNA. Within the stoichiometric space sampled, a 100-fold increase in the amount of RNA on average led to an increase of 27% in read alignment. The replicate correlation model was also highly predictive and explained a large fraction of the observed variance ($R^2=0.89$) in the correlation between replicate libraries. Interestingly, replicate correlation depends on the level of padlock probe library and the interaction term between the level of RNA and the level of padlock probe library. Within the stoichiometric space sampled, a 100-fold increase in padlock probe library levels resulted in, on average, a 0.34 increase in the value of the Pearson correlation coefficient for probe frequency between replicate libraries, with an additional increase, on average, of 0.08 if the RNA level and the padlock probe library level were 5000 ng and 10 ng respectively. With this information, the best choice for SplintR ligation stoichiometry was determined to be 5000 ng of total RNA with 10 ng of padlock probe library.

5.6.4 Splintlock-seq library prep protocol v.3

On ice, combine 5000 ng of RNA template and 10 ng of padlock probe library in a sterile 200 μ l PCR tube. Add 1 μ l of 10x SplintR ligase buffer (NEB) and nuclease free water to a final volume of 10 μ l. Mix by pipetting. Heat the mixture at 80°C for 2 minutes followed by a -1°C/min ramp to 25°C using a preheated thermocycler to allow probes to hybridize to the RNA. To this, add 10 μ l of a prepared ligase solution consisting of 1 μ l of 10x SplintR ligase buffer (NEB), 2 μ l of 1:20 diluted SplintR ligase (NEB, in Diluent A), and 7 μ l of nuclease free water. Mix by pipetting, and incubate at 30°C for 1 hour. Heat kill the ligase at 65°C for 20 minutes, denature RNA template from ligated padlock probes by heating at 95°C for 5 minutes, and put the reaction on ice. In a separate tube, combine 5 μ l of the ligation reaction with 5 μ l of 10x Φ 29 DNA polymerase buffer (NEB), 0.5 μ l of BSA (NEB, 20 mg/mL), 39 μ l of nuclease free water, and 0.5 μ l of Φ 29 DNA

polymerase. The reaction is mixed by pipetting and then incubated at 30°C for 30 minutes. To this, add 10 μ l prepared rolling circle amplification solution consisting of 4 μ l of nuclease free water, 1 μ l of 10x Φ 29 DNA polymerase buffer (NEB), 1.5 μ l of 50 uM exonuclease-resistant RCA initiation primer, 3 μ l of dNTPs (NEB, 10mM each) and 0.5 μ l of Φ 29 DNA polymerase (NEB). Incubate the reaction at 30°C overnight. The next morning, heat kill Φ 29 DNA polymerase at 65°C for 10 minutes. The rolling circle amplification product is then converted into a sequencing library by the following PCR reaction. On ice, combine 5 μ l of 1:10 diluted RCA product (in nuclease free water), 2.5 μ l each of 10 uM stocks of the custom Illumina adapter-adding primers described in Section 3.11, 25 μ l of NEBNext UltraII Q5 Master Mix, and 15 μ l of nuclease free water. After 20 cycles of amplification, purify the 200 bp amplicons with 1.25x AMPure SPRI beads and elute in 20 μ l of nuclease free water. These dual-indexed libraries can then be pooled according to sequencing core specifications and ran on a single SP 1x100nt lane using an Illumina NovaSeq 6000, or any sequencing platform compatible with TruSeq technology.

5.7 Padlock probe target specificity assay

Now that the Splintlock-seq library prep stoichiometry issue had been addressed, our attention was returned to quantifying padlock probe performance for the remaining 10,967 probes in the *S. mutans* UA159 padlock probe library v.2. We decided to assess padlock probe performance by investigating each probe’s ability to recognize its own target sequence, provided that target sequence is present, as well as it’s off-target recognition of other padlock probe target sequences and sequences in the rRNA. However, conducting this type of assessment at the single probe level would be practically infeasible. Therefore a pooled assessment approach was developed by synthesizing batches of 700 RNA padlock target sequences. Splintlock-seq libraries could then be prepared using synthetic RNA pools consisting of known RNA targets and profiled using the entire padlock probe library. By sequencing the resulting libraries and mapping reads back to specific probes, we can then determine the amount of signal generated in cases where that probe’s target was present and in cases where that probe’s target was absent.

5.7.1 RNA landing zone design and synthesis

Padlock probe target sequences were assigned a pool according to its target gene’s chromosome position. For example, the first gene by index position in the *S. mutans* UA159 genome is *dnaA* so all ten padlock target sequences for the *dnaA* transcript were assigned to pool 1. The sixteenth gene by index position is *mreD*, so the single padlock target sequence for *mreD* is assigned to pool 16. After the 16th gene by index position pool assignment starts back at 1. This rolling pool assignment strategy produced pools populated with 600-783 probe target sequences.

For every probe within the *S. mutans* UA159 padlock probe library v.2, a ssDNA construct was designed with the following architecture. At the 5’ end of the 115 nt sequence, I added one of sixteen random, dissimilar DNA sequences engineered to have annealing temperatures of 65°C

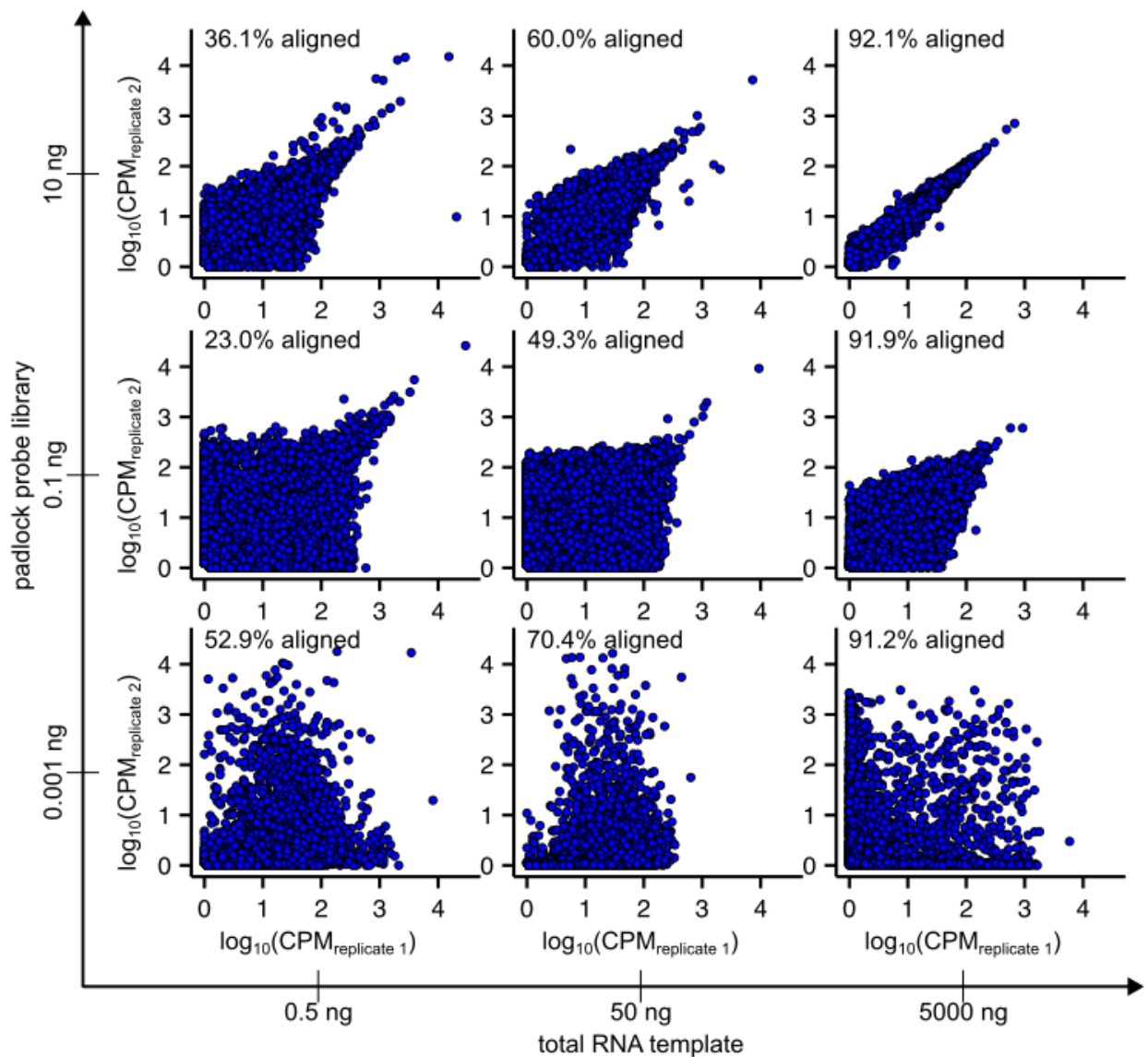


Figure 5.6: Nine different SplintR ligation conditions were tested with three different levels of total RNA and three different levels of padlock probe. High levels of correlation between replicates was achieved by using high amounts of probe while high levels of read alignment were achieved by using large amounts of RNA.

according to the NEB T_m Calculator for Q5 DNA polymerase. This sequence is referred to as the "pool tag". Downstream of the pool tag was a 70-nt variable region containing the 30-nt padlock probe target sequence as well as the transcript sequences extending 20-nt upstream and downstream of the target sequence. At the 3' end of the ssDNA construct, I added the reverse complement of the T7 promoter sequence (CCTATAGTGAGTCGTATTACAATTC). These 10,967 ssDNA constructs were synthesized (Custom Array) and used as template for 16 different PCRs. Each reaction utilized 5 ng of ssDNA construct library, 2.5 μ l of a 10 uM stock of a primer targeting the T7 promoter sequence, 2.5 μ l of a 10 uM stock of a primer targeting one of the sixteen pool tags, 15 μ l of nuclease free water and 25 μ l of 2x Q5 DNA polymerase master mix (NEB). After 20 cycles of amplification, dsDNA amplicons were purified and used as template for IVT reactions with the HiScribe T7 High Yield RNA Synthesis Kit according to the manufacturer's instructions (NEB). The resulting RNA molecules were referred to as "landing zones" because they each contain a 70-nt sequence mimicking the mRNA sequence targeted by one of the padlock probes within the v.2 library.

5.7.2 Experimental design and methods

The experimental design chosen for the padlock probe target specificity assay is shown in Figure 5.7 A. For each run, only one type of RNA template was used, with that RNA template being one of the RNA landing zone pools or the synthetic *S. mutans* rRNA mix. This design was executed in triplicate for a total of 51 libraries. For each run, a Splintlock-seq library was prepared according to the protocol in Section 5.6.4. It should be noted here that rRNA-templated libraries for this experiment required an additional 4 cycles of PCR relative to landing zone-templated libraries despite being prepared with the same mass of RNA. This seemed to indicate that the v.2 padlock probe pool was more orthogonal to rRNA on average than the v.1 pool. Counts for all probes in each of the library conditions were obtained by NGS and processed as described in Section 5.5.1.

5.7.3 Experimental results

Figure 5.7 B demonstrates that 94.8-96.9% of signal for all landing zone-templated libraries came from probes whose target was present in that library. This shows the v.2 padlock probe library overall has the ability to recognize targets if that target is present. We also wanted to understand individual probe specificity with regard to other probe target sequences and the rRNA. For each probe a specificity ratio was calculated as the sum of the CPM (counts per million) for the 3 libraries where that probe's target was present divided by the sum of the CPM for the 45 RNA landing zone-templated libraries where the probe's target was not present as well as the 3 rRNA-templated libraries. The median specificity score for padlock probes in the v.2 library was 27.3, with 85.0% of probes having a specificity ratio greater than or equal to 5. This indicated that the vast majority of probes within the v.2 library were heavily biased to their target sequence relative to all other padlock probe sequences and all rRNA sequences.

The majority of probes with a low specificity ratio were characterized as such because of off-target effects generated on the rRNA and not other padlock probe landing zones. Interestingly, a single

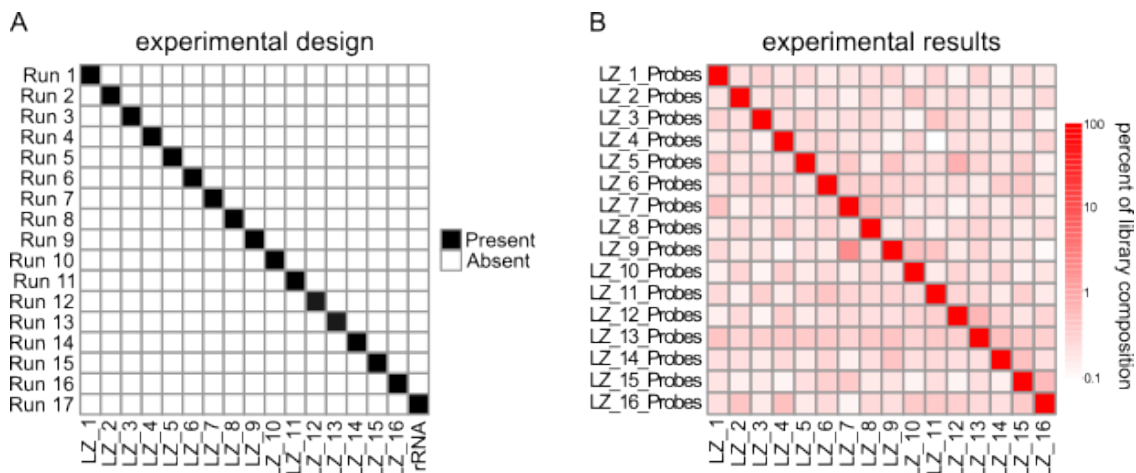


Figure 5.7: Design and results for the probe target specificity assay. **A.** illustrates the experimental design that was executed in triplicate. **B.** shows that 94.8-96.9% of signal for all landing zone-templated libraries came from probes whose target was present in that library. This demonstrates a global ability for target recognition at the experimental resolution of the RNA landing zones.

probe, SMU_RS05010_95_92, was responsible for 20.7% of the signal generated in rRNA-templated libraries, and a group of only 15 probes was responsible for over a third of the signal generated on rRNA-templated libraries. This indicated that the removal of only a small number of the remaining probes would greatly diminish the entire padlock probe library’s ability to generate signal with rRNA. The specificity ratio of all probes in the v.2 probe library was stored for reference during future padlock probe pool selection.

5.8 Padlock probe fold change quantification assay

While knowledge on a given padlock probe’s specificity with respect to rRNA is very useful, the probe’s ability to faithfully capture fold changes in target abundance is crucial in the context of transcriptomic profiling. For existing probe-based transcriptomic profiling technologies like RASL-seq [55], a candidate probes’ ability to measure fold-change is typically assessed individually with qPCR on biological samples with RNA-seq determined fold change values for the target gene [55], [64]. By combining direct control over the synthetic RNA landing zone pools with the throughput of NGS, we have developed an assay that allows us to compare actual and measured fold-changes for nearly 11,000 probes in only a small number of runs.

5.8.1 Experimental design

To simplify the nature of the fold-change experiments, the 16 RNA landing zone pools (1-16) were selectively combined in an equimolar manner to create 8 RNA pools (A-H). Landing zone pool 1 was combined with landing zone pool 9 to create RNA pool A, landing zone pool 2 was combined with landing zone pool 10 to create RNA pool B, landing zone 3 was combined with landing zone

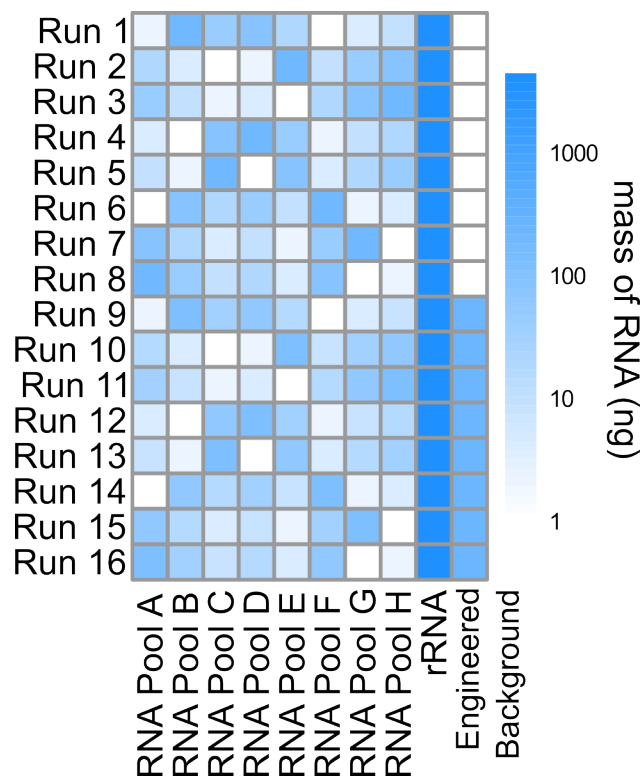


Figure 5.8: Experimental design for the probe fold change quantification assay. A Latin square design was utilized to ensure that for every run template from all 8 RNA pools were present, that every level of RNA pool was sampled, and the relationship between any two RNA pools was randomized across the entire experiment. Runs 1-8 correspond to low background conditions, while runs 9-16 correspond to high background conditions. This design was executed in triplicate.

pool 11 to create RNA pool C, and so forth. By controlling the relative amounts of each of these 8 RNA pools, we can simulate different *S. mutans* transcriptional states. 8 different levels of RNA abundance were selected: 2, 4, 8, 16, 32, 64, 128, and 256 ng. A Latin square design was employed in order to ensure that for any given run all 8 RNA pools were present, and that all 8 RNA levels were observed. This design also ensures that the relationship between any two RNA pools is randomized across the entire experiment. By including 8 different RNA pool levels, we enabled 28 pairwise fold change comparisons for each probe. As the goal of this experiment was to create synthetic representations of transcriptional states, and because mRNA only accounts for ~5% of the RNA present within a bacterial cell [66], [67], the *S. mutans* rRNA mixture was also included in every run.

The diversity of RNA sequences observed within the synthetic RNA pools and the rRNA mix is significantly lower than that in biological samples of total RNA. To introduce more RNA sequence diversity with the hopes of better emulating noise derived from off-target probe hybridization to non-landing zone mRNA sequences, an engineered background mixture of RNA was created in the following manner. Genomic DNA from *S. mutans* UA159 was sheared using a sonicator to an average length of 795 bp, roughly representative of the average gene length in *S. mutans* UA159. A KAPA HyperPrep DNA library prep (Roche) was then executed with this sheared genomic DNA using custom tailing adapters. In one of the adapter sequences, where the sequencing primer binding site would normally be, we substituted a T7 promoter site. Thus, at the end of the prep all amplicons would have a single T7 promoter. While this DNA library prep is not stranded, and results in the incorporation of sequences not found in *S. mutans* UA159 mRNA, RNA molecules produced using IVT reactions templated with this library greatly increase the RNA sequence diversity and should help synthetic RNA mixtures better emulate biological conditions. The RNA molecules resulting from IVT reactions using the HyperPrep DNA library are referred to as engineered background.

The experimental design for the padlock probe fold change quantification assay is shown in Figure 5.8. Runs 1-8 are considered "low background" and contain 510 ng of RNA pools A-H in various combinations determined by a Latin square design and 4500 ng of rRNA. Runs 9-16 are considered "high background" and contain halved amounts of RNA pools A-H (RNA levels are now 1, 2, 4, 8, 16, 32, 64, and 128 ng), 4500 ng of rRNA, and 254 ng of engineered background. Importantly, the relative mass amount of RNA landing zones in the high background condition emulates the approximate percentage of total bacterial RNA that is mRNA [66]. This design was executed in triplicate for a total of 48 Splintlock-seq libraries.

5.8.2 Experimental results

Splintlock-seq libraries were created according to the protocol in Section 5.6.4, and Illumina reads were processed as described in 5.5.1. Upon import into R, probe count data was organized by run number, where the resulting data structure contained the count information for every probe in the three replicate library prep conditions. Each padlock probe's target RNA pool (A-H) was then added to the data structure, along with its corresponding target's mass level for that experiment, and its

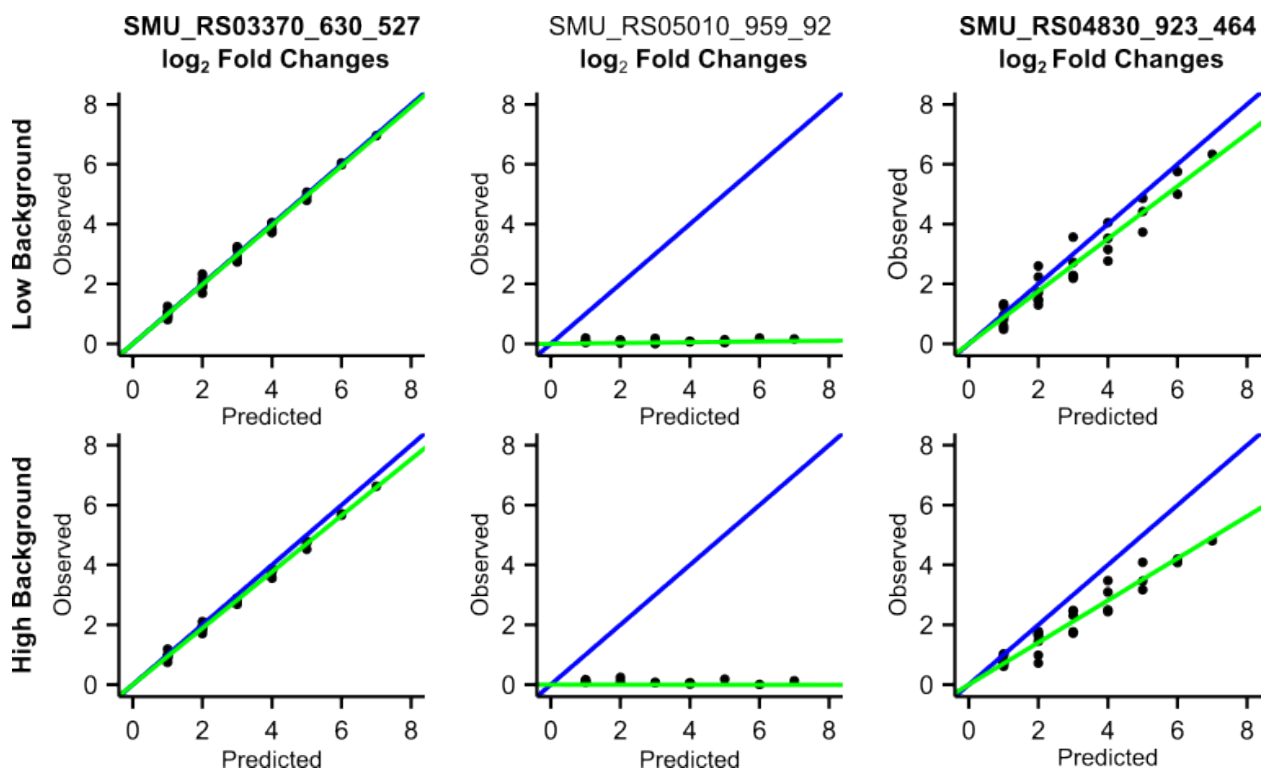


Figure 5.9: Results of the fold change quantification assay for three probes representing the good, the bad, and the typical. The blue line represents the line of equivalency, while the green line represents the univariate regression line of best fit for the data. The top row are results from low background experiments (rRNA only) and the bottom row are results from high background experiments (rRNA and engineered background). SMU_RS03370_630_527 represents an accurate and precise probe. SMU_RS05010_959_92 represents the worst type of probe for transcriptional profiling, one that is completely unresponsive to changes in target abundance. SMU_RS04830_923_464 has the median accuracy and precision metrics of a probe included in the v.3 probe library.

mean CPM. Two new data structures were then created from the existing data structures for every experiment within low background or high background conditions. In these data structures each observation corresponds to a padlock probe and the columns contain the predicted and observed fold changes for that probe. A representative visualization of three types of probe performance is depicted in Figure 5.9. For every probe, under low and high background conditions, accuracy was quantified as the mean fold change error, while precision was quantified as the residual standard error resulting from the univariate regression model predicting the base 2 logarithm of observed fold change from the base 2 logarithm of predicted fold change.

Of the 10,967 probes within the v.2 probe library, complete fold change assay data (all 28 comparisons were able to be made because the probe was observed in each of the 8 different runs) was measured for 9,321 probes in low background conditions; complete fold change assay data was measured for 10,213 probes in high background conditions; and complete fold change assay data was measured for 7,180 probes in both low and high background conditions. Under low background conditions, the mean percent fold change error for an typical probe was 34.6%. Under high background conditions, the mean percent fold change error for a typical probe was 52.7%. Under low background conditions, the residual standard error for a typical probe was 0.54. Under high background conditions, the residual standard error for a typical probe was 0.58.

In general, the decreased landing zone abundances and increased template complexity of libraries prepared with high background conditions resulted in a global decrease in probe accuracy. This could be due to the fact that increased sequence complexity leads to high levels of off-target hybridization which may or may not lead to ligation and signal generation. If ligation does occur, signal is artificially inflated which may obscure smaller changes in on-target signal. If ligation doesn't occur, precious probe is being sequestered away from its intended target, limiting the number of probes available to generate signal, thus reducing the ability to capture fold change information in an accurate manner. In the context of transcriptomic profiling, low accuracy probes may result in dampened fold-change approximation. Interestingly, the global precision metrics of probes did not change as much between low and high background conditions. The ability to detect differential expression of transcripts depends heavily on estimations of variance which are in turn effected by the precision of probe-based measurements. As Splintlock-seq was designed to complement RNA-seq as a low-cost profiling option, the ability to screen samples for differential expression places a high importance on probe precision.

5.9 Padlock probe library v.3 probe selection

With the information provided by the results of the specificity and the fold change assays, we felt informed enough to further refine the padlock probe pool with the aim of improving overall performance by eliminating low quality probes. First, any probe with a specificity ratio less than 5 was eliminated from the library. This removed 15% of the probes in the v.2 probe library but also removed the sources for 77% of off-target signal originating from rRNA, as determined by the

specificity assay. Next, all the probes were ranked according to their performance in the fold change assay. This ranking was achieved with a composite score calculated as the weighted average of a probe’s mean accuracy score (mean fold change error) between low background and high background conditions, and its mean precision score (residual standard error) between low background and high background conditions. The weighting used was 0.3 for the accuracy metric and 0.7 for the precision metric. This choice was made with the vision of Splintlock-seq being used as a screening technology in mind. While estimating fold change is very important, a higher priority was given to precision in the hopes of improving variance estimates which allow for better identification of differentially expressed genes.

The v.3 library was populated by selecting the best available probe for every gene according to the composite ranking and then back-filling with the best available probes until the global mean accuracy and mean precision metrics of the library began to worsen. This point occurred with a final library size of 2,404 probes targeting 1,809 genes (95.2%) of *S. mutans* UA159’s 1900 annotated genes. The v.3 probe library was synthesized in the manner described in 5.4.

5.10 Assessment of v.3 library rRNA orthogonality

With the v.3 probe library in hand, we wanted to quantify exactly how much intramolecular signal was generated on rRNA compared to total RNA. There are two types of signal noise that can be generated in a Splintlock-seq library prep. The first and less serious type is due to off-target intermolecular ligations. As long as these products are kept to manageable levels, as accomplished by the stoichiometry screen, they are easily identifiable in sequencing data because their chimeric sequences don’t align to the genome. The more nefarious type of noise is intramolecular ligation signal from off-target hybridization. This type of noise is more insidious because in biological experiments there is no way to distinguish between intramolecular probe sequences generated from the target transcript and intramolecular probe sequences generated from off-target hybridization. As the rRNA accounts for the vast majority (>90%) of cellular RNA [67], off-target hybridization to these molecules has traditionally represented the largest hurdle in probe design. We wanted to assess how our v.3 probe pool performed in this regard.

5.10.1 Experimental design and methods

Total RNA was isolated from four different cultures of *S. mutans* UA159 grown in TY media [71] supplemented with 0.5% glucose as described in Section 3.10. These four total RNA samples were then used to create four Splintlock-seq libraries with the v.3 padlock probes as described in Section 5.6.4. Four Splintlock-seq libraries were also prepared using rRNA only. Total signal levels in these libraries were measured by qPCR with primers amplifying across the ligation junction site. Intermolecular and intramolecular products were then distinguished by sequencing the libraries and aligning to the *S. mutans* UA159 genome, as described in Section 5.5.1. Total ligation signal values derived from qPCR were modified to reflect the aligned read fraction (intramolecular ligation

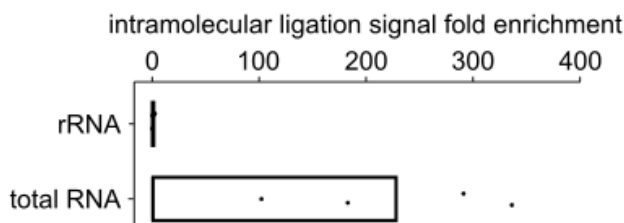


Figure 5.10: Intramolecular ligation signal fold enrichment for total RNA-templated Splintlock-seq libraries relative to rRNA-templated Splintlock-seq libraries. The presence of mRNA results higher levels of total signal and higher frequencies of intramolecular ligation. This in turn produces a mean 228-fold enrichment for intramolecular ligation signal using total RNA-templated libraries.

products) of each respective library. Intramolecular ligation signal fold enrichment was then determined by dividing the intramolecular signal for each library by the mean intramolecular signal for all rRNA libraries.

5.10.2 Experimental results

The results of this experiment are shown in Figure 5.10. On average, total signal levels from rRNA groups were 41-fold lower than total signal levels from total RNA groups. The mean aligned read percentage for rRNA-templated libraries was 6.9% while the mean aligned read percentage for total RNA-templated libraries was 95.4%. The intramolecular ligation signal fold enrichment for total RNA libraries ranged from 102.2-336.4, demonstrating high specificity for signal generation in the presence of mRNA. For reference, the current gold standard for "mRNA-enrichment" in bacterial RNA-seq preps actually relies on rRNA-depletion, and successfully enriches mRNA signal by 5-15 fold [72]–[74]. These results indicate that our combination of bioinformatics-guided design with empirical screening successfully led to a padlock probe pool that is highly orthogonal to rRNA.

5.11 Comparing Splintlock-seq and RNA-seq

While Splintlock-seq was originally designed to complement RNA-seq as a low-cost, targeted profiling method suitable for large-scale transcriptomic screens, we wanted to see if Splintlock-seq could be scaled to perform genome-wide screens for differentially-expressed genes. To do this, a carbon source utilization experiment was performed, whereby the results of a traditional RNA-seq pipeline could be compared directly to the results of a Splintlock-seq pipeline provided identical RNA inputs.

5.11.1 Experimental design and methods

Growing *S. mutans* on different carbon sources is known to produce transcriptional responses in a defined set of metabolism-related genes [71]. Therefore, we decided to profile and compare the transcriptional responses of *S. mutans* UA159 grown on TY media (3% tryptone and 0.1% yeast extract) supplemented with 0.5% glucose and *S. mutans* UA159 grown on TY media supplemented

with 0.5% fructose using RNA-seq and Splintlock-seq. Four biological replicates were included for each condition. Total RNA was isolated from these eight cultures as described in Section 3.10.

Splintlock-seq libraries were prepared using the eight total RNA samples and the v.3 padlock probe library as described in Section 5.6.4. Total RNA from the same eight total RNA samples were submitted to the Roy J. Carver Biotechnology Center for rRNA-depletion and RNA-seq library prep. All libraries were sequenced on SP-1 flow cells within an Illumina NovaSeq 6000 sequencing platform.

NGS data for Splintlock-seq libraries was processed as described in Section 5.5.1. RNA-seq data was quality filtered and trimmed to remove adapter sequences by the core facility. The resulting RNA-seq reads were directly aligned to the *S. mutans* UA159 genome as described in 5.5.1. Subsequent data analysis was performed in a custom R analysis pipeline.

A raw count matrix for Splintlock-seq was created with each observation corresponding to a padlock probe and each column corresponding to a library (Glucose 1-4 and Fructose 1-4). A raw count matrix for RNA-seq was created with each observation corresponding to a padlock probe and each column corresponding to a library (Glucose 1-4 and Fructose 1-4). These matrices were then input into DESeq2 for normalization, fold change and dispersion estimation, differential expression testing, and multiple testing correction using the default settings [75]. For direct gene-level comparison purposes, the Splintlock-seq probe producing the highest magnitude fold change for each gene was used. Only the 1806 genes with padlock probes in the v.3 library were used for comparison. It should be noted that fructose libraries were selected by DESeq2 to serve as the reference level. Thus, genes with a negative fold change are upregulated when *S. mutans* is grown on fructose, and genes with a positive fold change are downregulated when *S. mutans* is grown on fructose.

5.11.2 Experimental results

Comparisons of fold change estimations using RNA-seq libraries and Splintlock-seq libraries are shown in Figure 5.11. Both RNA-seq and Splintlock-seq show the vast majority of transcripts have no or very small magnitude fold changes between glucose and fructose-fuelled growth. This is expected as a previously defined set of *S. mutans* genes with large fold changes under similar glucose-fructose growth conditions represented only a small fraction of all genes [71]. Of particular interest is the general ability of Splintlock-seq to pull out genes with fold changes identified by RNA-seq. This is promising as it shows that Splintlock-seq is capable of picking up on differential transcript abundances in biological samples. However, the magnitude of the fold changes for those genes are generally under-predicted with Splintlock-seq. The range of $\log_2$ fold changes estimated by RNA-seq spans from -4.01 to 5.50. The range of $\log_2$ fold changes estimated by Splintlock-seq spans from -0.87 to 1.24. This discrepancy is interesting and could be due to a more pronounced version of the effect observed by introducing higher complexity background and reducing target levels in Section 5.8.2. It might be possible to bring these fold change estimates into better agreement by further optimizing the Splintlock-seq library prep protocol. Consider the case where a highly

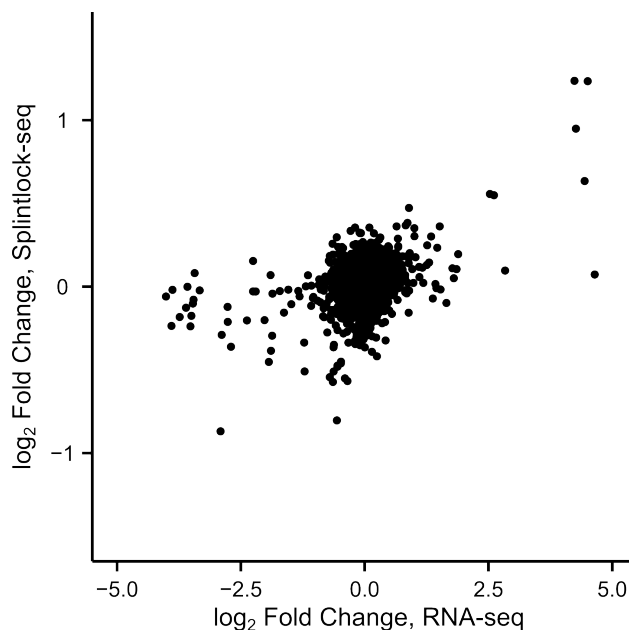


Figure 5.11: Fold change estimates for transcripts profiled using RNA-seq and Splintlock-seq. According to both technologies, most transcripts have no fold change between growth on glucose and growth on fructose. The observed range of fold changes for Splintlock-seq is smaller than the range observed with RNAseq.

expressed gene has a high magnitude fold change between two conditions, high levels of off-target sequestration of padlock probe might leave fewer probes available for on-target signal generation, thus dampening the fold change measurement’s magnitude. In this case, the addition of more padlock probe might be able to saturate off-target sites and leave more probes available to produce higher magnitude fold change estimates. Similarly, changes in ligation reaction temperature might boost the signal to noise ratio for probes allowing for better detection of high magnitude fold changes in low expression genes.

Volcano plots derived from the differential expression analyses for Splintlock-seq and RNA-seq are shown in Figure 5.12. These plots allow for the simultaneous visualization of differential expression hypothesis test results and the magnitude of estimated fold change for every gene. To aid in the comparison of Splintlock-seq and RNA-seq results, a set of seven genes from three operons previously characterized as differentially expressed in *S. mutans* grown on glucose versus fructose [71] were highlighted in both plots. These genes include: SMU_RS04035 (predicted transcriptional regulator of sugar metabolism), SMU_RS04040 (*pfkB*, 1-phosphofructokinase), SMU_RS07235 (PTS cellobiose transporter subunit IIC), SMU_RS07245 (PTS cellobiose transporter subunit IIA), SMU_RS08895 (PTS system fructose transporter subunit IID), SMU_RS08900 (PTS sugar transporter subunit IIC), and SMU_RS08910 (PTS fructose transporter subunit IIA). All of these genes appear in the “eruption tails” of the volcano plots for both Splintlock-seq and RNA-seq. **This indicates that Splintlock-seq has successfully identified a core set of differentially expressed genes.** However, in line with the discussion of fold change comparisons between Splintlock-seq and RNA-seq

above, the estimated fold changes with Splintlock-seq are of lower magnitude than RNA-seq. That said, the fold changes relative to each library prep technology's observed fold change range seem to have higher agreement. Taken together, these results demonstrate that Splintlock-seq is capable of scaling to accomplish genome-wide transcriptomic profiling and identify differential expression genes in agreement with the gold-standard technology, RNA-seq.

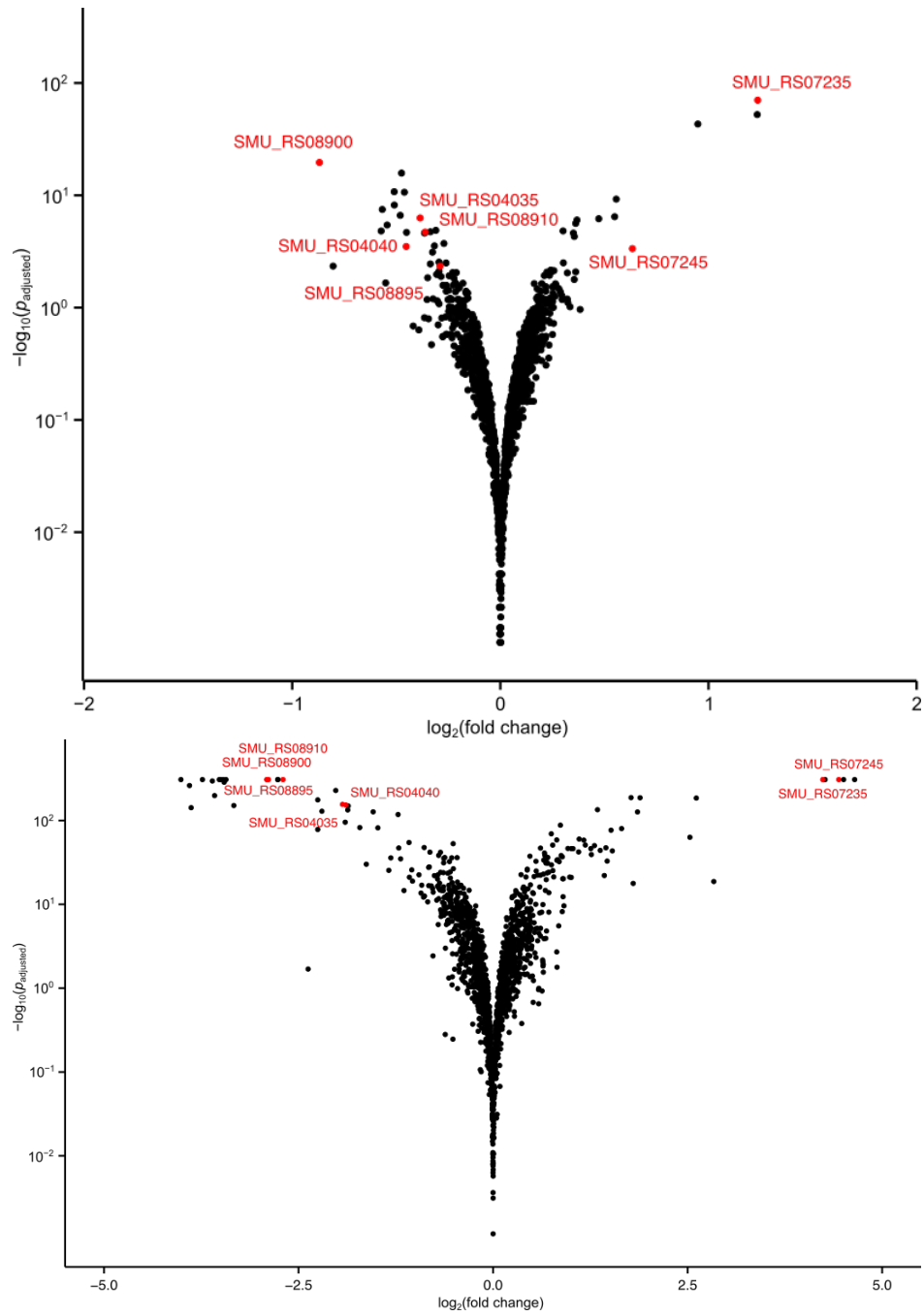


Figure 5.12: Volcano plots for Splintlock-seq (top) and RNA-seq (bottom). A set of seven genes known to be differentially expressed [71] when *S. mutans* is grown on glucose and fructose are colored in red. Splintlock-seq is able to identify these genes as differentially expressed, indicating promising utility in genome-wide transcriptional profiling contexts.

Chapter 6

Conclusions

6.1 Summary

In this dissertation, I present RACCET and Splintlock-seq, novel tools to aide in the genetic characterization of the oral streptococci.

Here, RACCET took a small number of uncharacterized promoters and RBSs and combined them with a novel modular assembly platform to produce tens of thousands of unique gene expression cassettes. These cassettes were then screened for various levels of transcriptional strength using a NGS molecular barcode counting assay. This resulted in a set of 12 characterized promoters spanning a 262-fold range of transcriptional strengths. These characterized promoters were then combinatorially combined with all RBSs and three different fluorescent reporter ORFs in a translational characterization screen. This produced a set of promoter-RBS pairs spanning a 426-fold range of gene expression strengths. Besides providing a robust set of valuable synthetic biology tools to researchers studying the oral streptococci, RACCET also revealed accurate prediction of gene expression strength in *S. mutans* requires an estimation of the interaction effects between promoter and RBS.

Splintlock-seq was developed in order to address the labor and cost issues associated with current bacterial RNA-seq library preps. The bottleneck requirement of manual probe design was eliminated with the development of an automated padlock probe design pipeline. A robust, library prep protocol was developed which ensured resulting sequencing libraries were free of intermolecular products and resulted in high degrees of technical reproducibility. High-throughput NGS assays were then developed and implemented to screen padlock probes for specificity, accuracy, and precision. Finally, Splintlock-seq was scaled to perform genome-wide profiling and succeeded in identifying a core set of differentially expressed genes within a biological experiment.

6.2 Future Directions

While *S. mutans* is a human pathogen of great importance in the context of dental disease, it is only one member of a rich microbial community that make up the oral streptococci. Indeed, recent studies have even called into question old assumptions of *S. mutans* being the primary causative agent for tooth decay, instead blaming larger ecological shifts that result in conditions which allow for multiple cariogenic bacteria to thrive [76], [77]. Within this framework, it is more critical than ever to understand the metabolic and genetic interactions of the oral microbiome, and specifically the oral streptococci. However, even compared to *S. mutans* many of these microbes have next to no characterized tools for manipulating gene expression. Given the phylogenetic proximity to *S. mutans*, I would be interested to see how the characterized expression toolkit derived here behaves in other members of the oral streptococci like *S. gordonii*, *S. sanguinus*, and *S. sobrinus*. Given that previous studies have shown that genomic context and expression vectors [13] can have an influence on gene expression strengths, I would also be interested in seeing what would change should the characterized expression parts from RACCET be expressed from pRW17-derived plasmids. Plasmid-based expression offers the advantage of a universal expression backbone and elviates the user from having to design new homology arms for different loci or different species. Copy number also offers another dimension on which synthetic biologists can control gene expression.

While great strides have been made since the development of Splintlock-seq began in the fall of 2019, many opportunities for improvement remain. The fundamental challenge in developing Splintlock-seq was, is, and always will be the simultaneous need for an optimized library prep protocol and an optimized padlock probe library. Of course, one can't really be achieved without the other. Instead, improvements in one area have been used to advance the other. This cyclical process is still ongoing. The fold change measurement issue observed in comparison with RNA-seq in 5.11.2 can be viewed as a probe issue but also as a library prep issue.

If the issue is simply that the probe pool we currently have doesn't contain probes of high enough quality, we have a simple fix: start with a much larger candidate pool. The 12,007 probes in v.1 library represent only 0.69% of the possible padlock probe sequences. Previously, we were limited by the capacity of the array technology used to produce the padlock probes; but since the project began synthesis options have been introduced for 96,000 or 8,000,000 oligonucleotide arrays. While the latter is overkill (the entire *S. mutans* mRNA sequence set only has 1,736,198 unique 30-mers), the former is an attractive option for broadening our search. It is completely feasible that the filtering parameters on the automated probe design pipeline were too strict and eliminated probes from consideration which might've proven to be excellent at capturing fold change information. Bioinformatic analysis of the problem probes identified in the v.1 library revealed no diagnostic features which would have allowed us to predict their off-target nature. This highlights the need for empirical validation. Given the ease and power of the fold-change quantification assay and our lab's expertise in deep learning, I would be very interested in conducting a massive top-down probe validation experiment where results of the fold change assay are fed into an AI-system which then tries to predict new, high performance probes from the available list of sequences.

However, if the fold change issue is due to library prep conditions, I think a systematic optimization of the SplintR ligation and rolling circle amplification should be conducted where the response is quantified as the global similarity to RNA-seq data. In doing so, I would suggest modulating several library prep variables while holding RNA constant to ensure NGS readouts are clean. These would include the concentration of padlock probe library, the concentration of SplintR ligase, the duration of ligation reaction, the temperature of ligation, and the duration of rolling circle amplification. Because this is fundamentally a multidimensional problem with likely interactions and higher order effects, response surface methodologies could be implemented to systematically characterize and optimize the Splintlock-seq library prep.

Bibliography

- [1] S. Klumpp, Z. Zhang, and T. Hwa, “Growth Rate-Dependent Global Effects on Gene Expression in Bacteria,” *Cell*, 2009, ISSN: 00928674. DOI: [10.1016/j.cell.2009.12.001](https://doi.org/10.1016/j.cell.2009.12.001).
- [2] J. H. Davis, A. J. Rubin, and R. T. Sauer, “Design, construction and characterization of a set of insulated bacterial promoters,” *Nucleic Acids Research*, 2011, ISSN: 03051048. DOI: [10.1093/nar/gkq810](https://doi.org/10.1093/nar/gkq810).
- [3] T. Hao, Z. Xie, M. Wang, *et al.*, “An anaerobic bacterium host system for heterologous expression of natural product biosynthetic gene clusters,” *Nature Communications*, 2019, ISSN: 20411723. DOI: [10.1038/s41467-019-11673-0](https://doi.org/10.1038/s41467-019-11673-0).
- [4] L. Keren, D. Van Dijk, S. Weingarten-Gabbay, *et al.*, “Noise in gene expression is coupled to growth rate,” *Genome Research*, 2015, ISSN: 15495469. DOI: [10.1101/gr.191635.115](https://doi.org/10.1101/gr.191635.115).
- [5] Z. Xie, F. Qi, and J. Merritt, “Development of a tunable wide-range gene induction system useful for the study of streptococcal toxin-antitoxin systems,” *Applied and Environmental Microbiology*, 2013, ISSN: 00992240. DOI: [10.1128/AEM.02320-13](https://doi.org/10.1128/AEM.02320-13).
- [6] D. J. LeBlanc, L. N. Lee, and A. Abu-Al-Jaibat, “Molecular, genetic, and functional analysis of the basic replicon of pVA380-1, a plasmid of oral streptococcal origin,” *Plasmid*, 1992, ISSN: 10959890. DOI: [10.1016/0147-619X\(92\)90044-B](https://doi.org/10.1016/0147-619X(92)90044-B).
- [7] C. Delorme, S. D. Ehrlich, and P. Renault, “Histidine biosynthesis genes in *Lactococcus lactis* subsp. *lactis*,” *Journal of Bacteriology*, 1992, ISSN: 00219193. DOI: [10.1128/jb.174.20.6571-6579.1992](https://doi.org/10.1128/jb.174.20.6571-6579.1992).
- [8] J. R. Van Der Ploeg, “Genome sequence of *Streptococcus mutans* bacteriophage M102,” *FEMS Microbiology Letters*, 2007, ISSN: 03781097. DOI: [10.1111/j.1574-6968.2007.00873.x](https://doi.org/10.1111/j.1574-6968.2007.00873.x).
- [9] W. R. Whitaker, E. S. Shepherd, and J. L. Sonnenburg, “Tunable Expression Tools Enable Single-Cell Strain Distinction in the Gut Microbiome,” *Cell*, 2017, ISSN: 10974172. DOI: [10.1016/j.cell.2017.03.041](https://doi.org/10.1016/j.cell.2017.03.041).
- [10] V. Potapov, J. L. Ong, R. B. Kucera, *et al.*, “Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly,” *ACS Synthetic Biology*, 2018, ISSN: 21615063. DOI: [10.1021/acssynbio.8b00333](https://doi.org/10.1021/acssynbio.8b00333).

- [11] S. T. Estrem, T. Gaal, W. Ross, and R. L. Gourse, "Identification of an UP element consensus sequence for bacterial promoters," *Proceedings of the National Academy of Sciences of the United States of America*, 1998, ISSN: 00278424. DOI: [10.1073/pnas.95.17.9761](https://doi.org/10.1073/pnas.95.17.9761).
- [12] R. K. Umarov and V. V. Solovyev, "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks," *PLoS ONE*, vol. 12, 2 Feb. 2017, ISSN: 19326203. DOI: [10.1371/journal.pone.0171410](https://doi.org/10.1371/journal.pone.0171410).
- [13] I. Biswas, J. K. Jha, and N. Fromm, "Shuttle expression plasmids for genetic studies in *Streptococcus mutans*," *Microbiology*, 2008, ISSN: 13500872. DOI: [10.1099/mic.0.2008/019265-0](https://doi.org/10.1099/mic.0.2008/019265-0).
- [14] C. Nieto, A. Puyet, and M. Espinosa, "MalR-mediated regulation of the *Streptococcus pneumoniae* malMP operon at promoter PM. Influence of a proximal divergent promoter region and competition between MalR and RNA polymerase proteins," *Journal of Biological Chemistry*, 2001, ISSN: 00219258. DOI: [10.1074/jbc.M010911200](https://doi.org/10.1074/jbc.M010911200).
- [15] W. Samarraï, D. X. Liu, A. M. White, *et al.*, "Differential responses of *Bacillus subtilis* rRNA promoters to nutritional stress," *Journal of Bacteriology*, 2011, ISSN: 00219193. DOI: [10.1128/JB.00708-10](https://doi.org/10.1128/JB.00708-10).
- [16] H. Chouayekh, P. Serror, S. Boudebouze, and E. Maguin, "Highly efficient production of the staphylococcal nuclease reporter in *Lactobacillus bulgaricus* governed by the promoter of the hlbA gene," *FEMS Microbiology Letters*, 2009, ISSN: 03781097. DOI: [10.1111/j.1574-6968.2009.01522.x](https://doi.org/10.1111/j.1574-6968.2009.01522.x).
- [17] P. R. Jensen and K. Hammer, "The sequence of spacers between the consensus sequences modulates the strength of prokaryotic promoters," *Applied and Environmental Microbiology*, 1998, ISSN: 00992240. DOI: [10.1128/aem.64.1.82-87.1998](https://doi.org/10.1128/aem.64.1.82-87.1998).
- [18] S. Ruiz-Cruz, V. Solano-Collado, M. Espinosa, and A. Bravo, "Novel plasmid-based genetic tools for the study of promoters and terminators in *Streptococcus pneumoniae* and *Enterococcus faecalis*," *Journal of Microbiological Methods*, 2010, ISSN: 01677012. DOI: [10.1016/j.mimet.2010.08.004](https://doi.org/10.1016/j.mimet.2010.08.004).
- [19] M. G. Shainheit, M. Mulé, and A. Camilli, "The core promoter of the capsule operon of *streptococcus pneumoniae* is necessary for colonization and invasive disease," *Infection and Immunity*, 2014, ISSN: 00199567. DOI: [10.1128/IAI.01289-13](https://doi.org/10.1128/IAI.01289-13).
- [20] M. B. Aspiras, K. M. Kazmerzak, P. E. Kolenbrander, R. McNab, N. Hardegen, and H. F. Jenkinson, "Expression of green fluorescent protein in *Streptococcus gordonii* DL1 and its use as a species-specific marker in coadhesion with *Streptococcus oralis* 34 in saliva-conditioned biofilms in vitro," *Applied and Environmental Microbiology*, 2000, ISSN: 00992240. DOI: [10.1128/AEM.66.9.4074-4083.2000](https://doi.org/10.1128/AEM.66.9.4074-4083.2000).

- [21] P. Slos, J. C. Bourquin, Y. Lemoine, and A. Mercenier, "Isolation and characterization of chromosomal promoters of *Streptococcus salivarius* subsp. *thermophilus*," *Applied and Environmental Microbiology*, 1991, ISSN: 00992240. DOI: [10.1128/aem.57.5.1333-1339.1991](https://doi.org/10.1128/aem.57.5.1333-1339.1991).
- [22] H. M. Salis, E. A. Mirsky, and C. A. Voigt, "Automated design of synthetic ribosome binding sites to control protein expression," *Nature Biotechnology*, 2009, ISSN: 10870156. DOI: [10.1038/nbt.1568](https://doi.org/10.1038/nbt.1568).
- [23] D. Omotajo, T. Tate, H. Cho, and M. Choudhary, "Distribution and diversity of ribosome binding sites in prokaryotic genomes," *BMC Genomics*, 2015, ISSN: 14712164. DOI: [10.1186/s12864-015-1808-6](https://doi.org/10.1186/s12864-015-1808-6).
- [24] R. L. Vellanoweth and J. C. Rabinowitz, "The influence of ribosome-binding-site elements on translational efficiency in *Bacillus subtilis* and *Escherichia coli* in vivo," *Molecular Microbiology*, 1992, ISSN: 13652958. DOI: [10.1111/j.1365-2958.1992.tb01548.x](https://doi.org/10.1111/j.1365-2958.1992.tb01548.x).
- [25] C. Unoson and E. G. H. Wagner, *Dealing with stable structures at ribosome binding sites: Bacterial translation and ribosome standby*, 2007. DOI: [10.4161/rna.4.3.5350](https://doi.org/10.4161/rna.4.3.5350).
- [26] F. Kunst, N. Ogasawara, I. Moszer, *et al.*, "The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*," *Nature*, 1997, ISSN: 00280836. DOI: [10.1038/36786](https://doi.org/10.1038/36786).
- [27] D. Ajdić, W. M. McShan, R. E. McLaughlin, *et al.*, "Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen," *Proceedings of the National Academy of Sciences of the United States of America*, 2002, ISSN: 00278424. DOI: [10.1073/pnas.172501299](https://doi.org/10.1073/pnas.172501299).
- [28] J. A. Lanie, W. L. Ng, K. M. Kazmierczak, *et al.*, "Genome sequence of Avery's virulent serotype 2 strain D39 of *Streptococcus pneumoniae* and comparison with that of unencapsulated laboratory strain R6," *Journal of Bacteriology*, 2007, ISSN: 00219193. DOI: [10.1128/JB.01148-06](https://doi.org/10.1128/JB.01148-06).
- [29] M. Van De Guchte, S. Penaud, C. Grimaldi, *et al.*, "The complete genome sequence of *Lactobacillus bulgaricus* reveals extensive and ongoing reductive evolution," *Proceedings of the National Academy of Sciences of the United States of America*, 2006, ISSN: 00278424. DOI: [10.1073/pnas.0603024103](https://doi.org/10.1073/pnas.0603024103).
- [30] M. M. Vickerman, S. Iobst, A. M. Jesionowski, and S. R. Gill, "Genome-wide transcriptional changes in *Streptococcus gordonii* in response to competence signaling peptide," *Journal of Bacteriology*, 2007, ISSN: 00219193. DOI: [10.1128/JB.01023-07](https://doi.org/10.1128/JB.01023-07).
- [31] D. S. Bindels, L. Haarbosch, L. Van Weeren, *et al.*, "MScarlet: A bright monomeric red fluorescent protein for cellular imaging," *Nature Methods*, 2016, ISSN: 15487105. DOI: [10.1038/nmeth.4074](https://doi.org/10.1038/nmeth.4074).
- [32] N. C. Shaner, G. G. Lambert, A. Chamma, *et al.*, "A bright monomeric green fluorescent protein derived from *Branchiostoma lanceolatum*," *Nature Methods*, 2013, ISSN: 15487091. DOI: [10.1038/nmeth.2413](https://doi.org/10.1038/nmeth.2413).

- [33] H. W. Ai, J. N. Henderson, S. J. Remington, and R. E. Campbell, “Directed evolution of a monomeric, bright and photostable version of *Clavularia* cyan fluorescent protein: Structural characterization and applications in fluorescence imaging,” *Biochemical Journal*, 2006, ISSN: 02646021. DOI: [10.1042/BJ20060874](https://doi.org/10.1042/BJ20060874).
- [34] J. D. Pédelacq, S. Cabantous, T. Tran, T. C. Terwilliger, and G. S. Waldo, “Engineering and characterization of a superfolder green fluorescent protein,” *Nature Biotechnology*, 2006, ISSN: 10870156. DOI: [10.1038/nbt1172](https://doi.org/10.1038/nbt1172).
- [35] J. Goedhart, D. Von Stetten, M. Noirclerc-Savoye, *et al.*, “Structure-guided evolution of cyan fluorescent proteins towards a quantum yield of 93%,” *Nature Communications*, 2012, ISSN: 20411723. DOI: [10.1038/ncomms1738](https://doi.org/10.1038/ncomms1738).
- [36] O. M. Subach, P. J. Cranfill, M. W. Davidson, and V. V. Verkhusha, “An enhanced monomeric blue fluorescent protein with the high chemical stability of the chromophore,” *PLoS ONE*, 2011, ISSN: 19326203. DOI: [10.1371/journal.pone.0028674](https://doi.org/10.1371/journal.pone.0028674).
- [37] P. Trieu-Cuot, G. De Cespedes, F. Bentorcha, F. Delbos, E. Gaspar, and T. Horaud, “Study of heterogeneity of chloramphenicol acetyltransferase (CAT) genes in streptococci and enterococci by polymerase chain reaction: Characterization of a new CAT determinant,” *Antimicrobial Agents and Chemotherapy*, 1993, ISSN: 00664804. DOI: [10.1128/AAC.37.12.2593](https://doi.org/10.1128/AAC.37.12.2593).
- [38] T. J. Lambert, “Fpbase: A community-editable fluorescent protein database,” *Nature Methods*, vol. 16, pp. 277–278, 4 Apr. 2019, ISSN: 15487105. DOI: [10.1038/s41592-019-0352-8](https://doi.org/10.1038/s41592-019-0352-8).
- [39] T. Ogawa, K. Kryukov, T. Imanishi, and K. Shiroguchi, “The efficacy and further functional advantages of random-base molecular barcodes for absolute and digital quantification of nucleic acid molecules,” *Scientific Reports*, 2017, ISSN: 20452322. DOI: [10.1038/s41598-017-13529-3](https://doi.org/10.1038/s41598-017-13529-3).
- [40] A. M. Smith, L. E. Heisler, R. P. St. Onge, *et al.*, “Highly-multiplexed barcode sequencing: An efficient method for parallel analysis of pooled samples,” *Nucleic Acids Research*, 2010, ISSN: 03051048. DOI: [10.1093/nar/gkq368](https://doi.org/10.1093/nar/gkq368).
- [41] P. A. Jensen, Z. Zhu, and T. van Opijnen, “Antibiotics Disrupt Coordination between Transcriptional and Phenotypic Stress Responses in Pathogenic Bacteria,” *Cell Reports*, 2017, ISSN: 22111247. DOI: [10.1016/j.celrep.2017.07.062](https://doi.org/10.1016/j.celrep.2017.07.062).
- [42] A. J. Storm and P. A. Jensen, “Designing randomized dna sequences free of restriction enzyme recognition sites,” *Biotechnology Journal*, vol. 13, p. 1700326, 1 Jan. 2018, ISSN: 18606768. DOI: [10.1002/biot.201700326](https://doi.org/10.1002/biot.201700326). [Online]. Available: <https://onlinelibrary.wiley.com/doi/10.1002/biot.201700326>.
- [43] A. M. Eren, J. H. Vineis, H. G. Morrison, and M. L. Sogin, “A filtering method to generate high quality short reads using illumina paired-end technology,” *PLoS ONE*, vol. 8, I. K. Jordan, Ed., e66643, 6 Jun. 2013, ISSN: 1932-6203. DOI: [10.1371/journal.pone.0066643](https://doi.org/10.1371/journal.pone.0066643). [Online]. Available: <https://dx.plos.org/10.1371/journal.pone.0066643>.

- [44] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: A flexible trimmer for illumina sequence data,” *Bioinformatics*, vol. 30, pp. 2114–2120, 15 Aug. 2014, ISSN: 1460-2059. DOI: [10.1093/bioinformatics/btu170](https://doi.org/10.1093/bioinformatics/btu170). [Online]. Available: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170>.
- [45] M. V. Matz, A. F. Fradkov, Y. A. Labas, *et al.*, “Fluorescent proteins from nonbioluminescent Anthozoa species,” *Nature Biotechnology*, 1999, ISSN: 10870156. DOI: [10.1038/13657](https://doi.org/10.1038/13657).
- [46] C. Kiss, J. Temirov, L. Chasteen, G. S. Waldo, and A. R. Bradbury, “Directed evolution of an extremely stable fluorescent protein,” *Protein Engineering, Design and Selection*, 2009, ISSN: 17410126. DOI: [10.1093/protein/gzp006](https://doi.org/10.1093/protein/gzp006).
- [47] T. J. Rudge, J. R. Brown, F. Federici, *et al.*, “Characterization of Intrinsic Properties of Promoters,” *ACS Synthetic Biology*, 2016, ISSN: 21615063. DOI: [10.1021/acssynbio.5b00116](https://doi.org/10.1021/acssynbio.5b00116).
- [48] S. Zhang, Z. Zou, J. Kreth, and J. Merritt, “Recombineering in *Streptococcus mutans* using direct repeat-mediated cloning-independent markerless mutagenesis (DR-CIMM),” *Frontiers in Cellular and Infection Microbiology*, 2017, ISSN: 22352988. DOI: [10.3389/fcimb.2017.00202](https://doi.org/10.3389/fcimb.2017.00202).
- [49] W. A. Kibbe, “Oligocalc: An online oligonucleotide properties calculator,” *Nucleic Acids Research*, vol. 35, SUPPL.2 Jul. 2007, ISSN: 03051048. DOI: [10.1093/nar/gkm234](https://doi.org/10.1093/nar/gkm234).
- [50] B. J. Haas, M. Chin, C. Nusbaum, B. W. Birren, and J. Livny, “How deep is deep enough for RNA-Seq profiling of bacterial transcriptomes?” *BMC Genomics*, 2012, ISSN: 14712164. DOI: [10.1186/1471-2164-13-734](https://doi.org/10.1186/1471-2164-13-734).
- [51] A. Hernandez and C. Wright, *Roy J. Carver Biotechnology Center - DNA Services Pricing*, 2020. [Online]. Available: <https://biotech.illinois.edu/htdna/pricing> (visited on 04/11/2020).
- [52] Wetterstrand KA, *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*, 2016.
- [53] R. Hrdlickova, M. Toloue, and B. Tian, *RNA-Seq methods for transcriptome analysis*, 2017. DOI: [10.1002/wrna.1364](https://doi.org/10.1002/wrna.1364).
- [54] A. A. Shishkin, G. Giannoukos, A. Kucukural, *et al.*, “Simultaneous generation of many RNA-seq libraries in a single reaction,” *Nature Methods*, 2015, ISSN: 15487105. DOI: [10.1038/nmeth.3313](https://doi.org/10.1038/nmeth.3313).
- [55] H. Li, J. Qiu, and X. D. Fu, “RASL-seq for Massively Parallel and Quantitative Analysis of Gene Expression,” *Current Protocols in Molecular Biology*, 2012, ISSN: 19343639. DOI: [10.1002/0471142727.mb0413s98](https://doi.org/10.1002/0471142727.mb0413s98).
- [56] V. Sriskanda and S. Shuman, “Specificity and fidelity of strand joining by *Chlorella* virus DNA ligase,” *Nucleic Acids Research*, 1998, ISSN: 03051048. DOI: [10.1093/nar/26.15.3536](https://doi.org/10.1093/nar/26.15.3536).
- [57] T. Krzywkowski and M. Nilsson, “Fidelity of RNA templated end-joining by *Chlorella* virus DNA ligase and a novel iLock assay with improved direct RNA detection accuracy,” *Nucleic Acids Research*, 2017, ISSN: 13624962. DOI: [10.1093/nar/gkx708](https://doi.org/10.1093/nar/gkx708).

- [58] J. Jin, S. Vaud, A. M. Zhelkovsky, J. Posfai, and L. A. McReynolds, “Sensitive and specific miRNA detection method using SplintR Ligase,” *Nucleic Acids Research*, 2016, ISSN: 13624962. DOI: [10.1093/nar/gkw399](https://doi.org/10.1093/nar/gkw399).
- [59] C. K. Roy, S. Olson, B. R. Graveley, P. D. Zamore, and M. J. Moore, “Assessing long-distance RNA sequence connectivity via RNA-templated DNA-DNA ligation,” *eLife*, 2015, ISSN: 2050084X. DOI: [10.7554/eLife.03700](https://doi.org/10.7554/eLife.03700).
- [60] B. C. Satterfield, *Cooperative primers, probes, and applications thereof*, 2013.
- [61] L. Blanco, A. Bernad, J. M. Lazaro, G. Martin, C. Garmendia, and M. Salas, “Highly efficient DNA synthesis by the phage Φ 29 DNA polymerase. Symmetrical mode of DNA replication,” *Journal of Biological Chemistry*, 1989, ISSN: 00219258. DOI: [10.13039/1000000002](https://doi.org/10.13039/1000000002).
- [62] M. S. Soengas, C. Gutiérrez, and M. Salas, “Helix-destabilizing activity of ϕ 29 single-stranded DNA binding protein: Effect on the elongation rate during strand displacement DNA replication,” *Journal of Molecular Biology*, 1995, ISSN: 00222836. DOI: [10.1006/jmbi.1995.0570](https://doi.org/10.1006/jmbi.1995.0570).
- [63] Y. Xing, P. Wang, Y. Zang, *et al.*, “A colorimetric method for H1N1 DNA detection using rolling circle amplification,” *Analyst*, 2013, ISSN: 13645528. DOI: [10.1039/c3an00100h](https://doi.org/10.1039/c3an00100h).
- [64] J. B. Li, Y. Gao, J. Aach, *et al.*, “Multiplex padlock targeted sequencing reveals human hypermutable cpg variations,” *Genome Research*, vol. 19, pp. 1606–1615, 9 Sep. 2009, ISSN: 10889051. DOI: [10.1101/gr.092213.109](https://doi.org/10.1101/gr.092213.109).
- [65] Y. E. Murgha, J. M. Rouillard, and E. Gulari, “Methods for the preparation of large quantities of complex single-stranded oligonucleotide libraries,” *PLoS ONE*, 2014, ISSN: 19326203. DOI: [10.1371/journal.pone.0094752](https://doi.org/10.1371/journal.pone.0094752).
- [66] A. J. Westermann, S. A. Gorski, and J. Vogel, “Dual rna-seq of pathogen and host,” *Nature Reviews Microbiology*, vol. 10, pp. 618–630, 9 Sep. 2012, ISSN: 17401526. DOI: [10.1038/nrmicro2852](https://doi.org/10.1038/nrmicro2852).
- [67] C. Rosenow, R. M. Saxena, M. Durst, and T. R. Gingeras, “Prokaryotic rna preparation methods useful for high density array analysis: Comparison of two approaches,” *Nucleic Acids Research*, vol. 29, e112, 22 Nov. 2020, ISSN: 13624962. DOI: [10.1093/nar/29.22.e112](https://doi.org/10.1093/nar/29.22.e112).
- [68] H. Liu, L. Li, L. Duan, *et al.*, “High specific and ultrasensitive isothermal detection of microRNA by padlock probe-based exponential rolling circle amplification,” *Analytical Chemistry*, 2013, ISSN: 00032700. DOI: [10.1021/ac401715k](https://doi.org/10.1021/ac401715k).
- [69] X. Y. Li, Y. C. Du, Y. P. Zhang, and D. M. Kong, “Dual functional Phi29 DNA polymerase-triggered exponential rolling circle amplification for sequence-specific detection of target DNA embedded in long-stranded genomic DNA,” *Scientific Reports*, 2017, ISSN: 20452322. DOI: [10.1038/s41598-017-06594-1](https://doi.org/10.1038/s41598-017-06594-1).

- [70] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short dna sequences to the human genome,” *Genome Biology*, vol. 10, R25, 3 Mar. 2009, ISSN: 14747596. DOI: [10.1186/gb-2009-10-3-r25](https://doi.org/10.1186/gb-2009-10-3-r25). [Online]. Available: <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25>.
- [71] L. Zeng and R. A. Burne, “Sucrose- and fructose-specific effects on the transcriptome of streptococcus mutans, as determined by rna sequencing,” *Applied and Environmental Microbiology*, vol. 82, pp. 146–156, 1 Jan. 2016, ISSN: 10985336. DOI: [10.1128/AEM.02681-15](https://doi.org/10.1128/AEM.02681-15).
- [72] A. Wahl, C. Huptas, and K. Neuhaus, “Comparison of rna depletion methods for efficient bacterial mrna sequencing,” *Scientific Reports*, vol. 12, pp. 1–11, 1 Dec. 2022, ISSN: 20452322. DOI: [10.1038/s41598-022-09710-y](https://doi.org/10.1038/s41598-022-09710-y).
- [73] N. Kumar, M. Lin, X. Zhao, *et al.*, “Efficient enrichment of bacterial mrna from host-bacteria total rna samples,” *Scientific Reports*, vol. 6, Oct. 2016, ISSN: 20452322. DOI: [10.1038/srep34850](https://doi.org/10.1038/srep34850).
- [74] P. H. Culviner, C. K. Guegler, and M. T. Laub, “A simple, cost-effective, and robust method for rna depletion in rna-sequencing studies,” *mBio*, vol. 11, 2 Mar. 2020, ISSN: 21507511. DOI: [10.1128/mBio.00010-20](https://doi.org/10.1128/mBio.00010-20).
- [75] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for rna-seq data with deseq2,” *Genome Biology*, vol. 15, p. 550, 12 Dec. 2014, ISSN: 1474760X. DOI: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8). [Online]. Available: <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>.
- [76] B. T. Rosier, M. D. Jager, E. Zaura, and B. P. Krom, “Historical and contemporary hypotheses on the development of oral diseases: Are we there yet?” *Frontiers in Cellular and Infection Microbiology*, vol. 4, JUL 2014, ISSN: 22352988. DOI: [10.3389/fcimb.2014.00092](https://doi.org/10.3389/fcimb.2014.00092).
- [77] B. Nyvad and N. Takahashi, “Integrated hypothesis of dental caries and periodontal diseases,” *Journal of Oral Microbiology*, vol. 12, p. 1710953, 1 Jan. 2020, ISSN: 2000-2297. DOI: [10.1080/20002297.2019.1710953](https://doi.org/10.1080/20002297.2019.1710953). [Online]. Available: <https://www.tandfonline.com/doi/full/10.1080/20002297.2019.1710953>.

Appendix A

Probability calculations for padlock probe design

What is the probability that a random DNA k -mer is at most $m_{\max}$ substitution mutations away from a random k -mer contained in a provided k -mer database if the database has a size d ?

For a given k , there are 4^k k -mers in total. The probability of randomly selecting a database k -mer is

$$p = \frac{d}{4^k}.$$

Let's now consider the following question: For a set number of allowed mismatches, $m_{\max}$, how many k -mers can be made given a starting k -mer? If $m_{\max} = 0$, the answer is 1, because the only k -mer you can make is the starting k -mer. If $m_{\max} = 1$, then you can make $3k$ new mutant sequences along with the starting k -mer, so the answer is $3k + 1$. The generalized expression that answers the above question is

$$n = \sum_{m=0}^{m_{\max}} \binom{k}{m} 3^m.$$

This expression is arrived at by realizing that there are k choose m possible combinations of positions to mutate in a k -mer, and at every chosen mutation site there three possible substitutions to make.

Now we have the probability of selecting a database k -mer at random, and the number of k -mers we can make provided a maximum number of substitution mutations, $m_{\max}$. We can now predict the probability of encountering exactly x database k -mers,

$$P(X = x) = \binom{n}{x} \left(\frac{d}{4^k}\right)^x \left(1 - \frac{d}{4^k}\right)^{n-x}.$$

Since

$$\lim_{a \rightarrow \infty} \binom{a}{b} p^b (1-p)^{a-b} = \frac{\lambda^b e^{-\lambda}}{b!}, \quad \lambda = pa$$

and because n grows rapidly with increases in k and $m_{\max}$, we can rewrite the probability of

encountering exactly x database k -mers as

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda = \frac{d}{4^k} n.$$

Now we want to know the probability that we encounter *any* number of database k -mers. This can be represented by the following expression

$$\sum_{x=1}^{\infty} P(X = x) = \sum_{x=1}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = 1 - P(X = 0) = 1 - e^{-\lambda}.$$

Therefore, given a random starting k -mer, the probability of encountering any number of k -mers from a database of size d , and allowing for a maximum number of mismatches, $m_{\max}$ is

$$P_{\text{hit}} = 1 - e^{-dn/4^k}.$$

The probability of avoiding any database k -mers is then

$$P_{\text{miss}} = 1 - P_{\text{hit}} = e^{-dn/4^k}.$$

Say you have a set of DNA sequences (query) and you want to select the subset which is most dissimilar to another set of DNA sequences (database). You do this by fuzzy matching your query set against the database but aren't sure how many mismatches you can allow. You can utilize the above expression to predict the fraction of candidate sequences which remain after the fuzzy matching. Let's say you're designing padlock probes for transcriptomic profiling and want to pick target DNA sequences that are as dissimilar to a set of "no-go" sequences (rRNA and tRNA) as possible. You develop a filter that eliminates any sequence within a certain number of edit distances from the "no-go" sequences but you don't know what filter threshold to use. Running real data through the filter with different threshold values is impractical because the filter is computationally demanding and your laptop is ancient. Padlock probes have two 15-mer ends and both sequences need to be dissimilar for the probe to pass the filter. So with $k = 15$ and $d = 6787$ (the number of unique 15-mers in the rRNA and tRNA sequences of *S. mutans* UA159), what is the maximum filter threshold for candidate sequences? Using the derivation above, and assuming that the sequences of the padlock probe ends are independent of each other, we can approximate probe retention for different thresholds using

$$P_{\text{miss}}^2 = (1 - P_{\text{hit}})^2 = e^{-2dn/4^k}.$$

As shown in Figure A.1, the probability theory derived here generates sequence retention rates that match very well with the empirical results! And now you know that you can allow for three mismatches and retain about 80% of your query set. Or you can choose a more stringent threshold of four mismatches and retain about 25% of your query set.

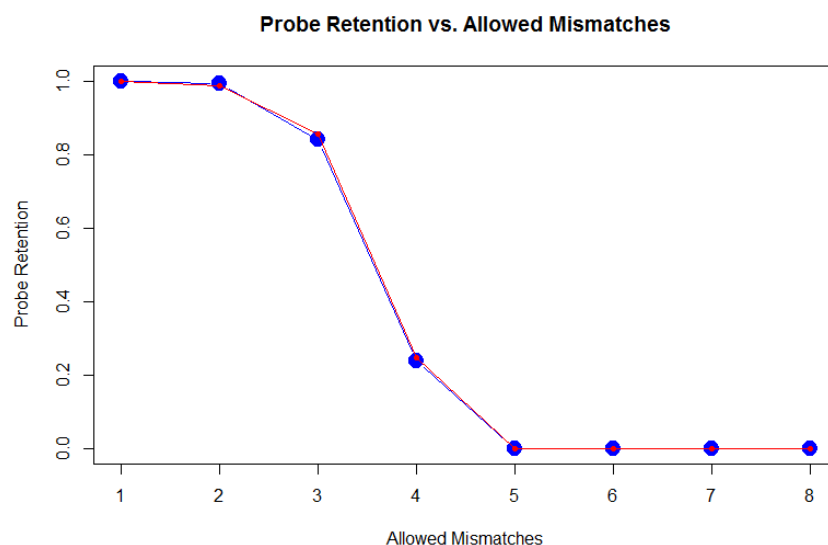


Figure A.1: Comparison of filtering results predicted using the derived probability equation to empirical results. The red line represents theoretical predictions and the blue line represents the empirical results painstakingly acquired by running the filter with real candidate probe sequences and real “no-go” sequences from the *S. mutans* UA159 genome over the course of many hours on a geriatric computer