Overcoming DNA Restriction-Modification systems to enable transformation and genetic engineering in non-model organisms

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Lauren Ann Riley

December 2020

Abstract

Many organisms naturally possess complex physiological traits that are of interest for biotechnology research. These non-model organisms are not routinely used though, due to a lack of foundational knowledge and genetic tools. The development of genetic tools is limited in part by the inability to efficiently transform DNA into these organisms. One of the major barriers to efficient DNA transformation is the native DNA Restriction-Modification systems within bacteria. DNA Restriction-Modification systems act as an immune system to cut heterologous DNA methylated differently than the host. To overcome these systems, DNA needs to be methylated in the same manner as the host organism prior to transformation. In order to determine the sites targeted for restriction we performed methylome analysis for non-model organisms of interest in collaboration with the Department of Energy's Joint Genome Institute. This information was used to choose methyltransferases for expression in *E. coli* to protect DNA for transformation of multiple organisms. Using this approach, we enabled the transformation of eight strains of bacteria including two strains of *Megasphaera elsdenii.* We are now developing *M. elsdenii* into a new bioengineering platform for the production of fuels and chemicals. *M. elsdenii* is a compelling platform organism due to the ability to efficiently produce C-4 and C-6 acids from lactate and glucose. This work acts as a blueprint for the rapid transformation and genetic engineering of non-model organisms.

Table of Contents

Chapter I: Genetic tool development in non-model microorganisms for biofuel and chemical production 1

Abstract 2

Introduction 2

Systems Biology in Non-Model Microorganisms 4

Barriers to Genetic Modification and Enabling Transformation 5

Chromosome Modification Tools in Non-model Microorganisms 13

Gene Expression Tools in Non-model Microorganisms 25

References 32

Chapter II: Rational development of transformation in *Clostridium thermocellum* ATCC 27405 via complete methylome analysis and evasion of native restriction-modification systems 44

Abstract 45

Introduction 46

Materials and Methods 49

Results 52

Discussion 56

References 61

Chapter III: Dynamic evolution of defense systems targeting foreign DNA in *Streptococcus thermophilus* strains 64

Abstract 65

Introduction 66

Materials and Methods 68

Results 72

Discussion 82

References 87

Chapter IV: Simple and rapid site-specific integration of multiple heterologous DNAs in *E. coli* 89

Abstract 90

Introduction 91
Materials and Methods..... 94
Results..... 99
Discussion 108
References 111
Chapter V: Developing *Megasphaera elsdenii* into a Platform Organism for Chemical Production from Lactate 113
Abstract..... 114
Introduction 115
Materials and Methods..... 118
Results..... 122
Discussion 129
References 132
Chapter VI: Conclusion..... 134
References 140
Appendix 141
Vita……. 145

List of Tables

Table 1.1: Commonly used antibiotics and their corresponding selectable markers..... 12

Table 1.2: Commonly used counter selectable markers and associated selections for homologous recombination in bacteria..... 17

Table 2.1: Methylated motifs in *C. thermocellum* ATCC27405..... 53

Table 2.2: Transformation efficiency of *C. thermocellum* ATCC 27405 using methylated plasmid DNA..... 57

Table 3.1 The target sequence used for each CRISPR system in both strains of *S. thermophilus* and how they were determined..... 71

Table 3.2 Methylome analysis of *S. thermophilus* JIM 8232 and DGCC 7710 75

Table 3.3 Methylome analysis of *E. coli* expressing an *L. lactis* UC063 Type I methyltransferase..... 77

Table 4.1 Strains utilized for the serine phage integration system 95

Table 4.2 Methylome analysis of *C. clariflavum* 4-2a 107

Table 4.3 Methylome analysis of *C. clariflavum* 4-2a *E. coli* methylation strain 107

Table 5.1 Methylome analysis and transformation efficiencies of *M. elsdenii* ATCC 25940 and NCIMB 70241 123

Table 5.2 Methylome analysis of the *E. coli* strains expressing *M. elsdenii* methyltransferases 125

Table 5.3 Antibiotic MIC for *M. elsdenii*..... 125

Table A.1 Methylome analysis in strains of non-model bacteria 141

Table A.2 Genotype of *E. coli* methylation strains..... 142

Table A.3 Transformation in non-model bacteria..... 144

List of Figures

Figure 1.1: Methods using homologous recombination for gene deletions 14

Figure 2.1: Confirmation of methyltransferase functionality in *E. coli* 55

Figure 3.1: Restriction modification systems in *S. thermophilus* as predicted by New England Biolabs REBASE 73

Figure 3.2: Pairwise comparison of homologous Type I RM systems from *S. thermophilus* strains and an *L. lactis* plasmid 77

Figure 3.3: The effect of RM systems on transformation efficiency in *S. thermophilus* 79

Figure 3.4: CRISPR systems within *Streptococcus thermophilus* strains 81

Figure 3.5: Activity of CRISPR systems in *S. thermophilus* 83

Figure 4.1: Site-specific recombinase system for DNA integration into the *E. coli* chromosome 100

Figure 4.2: Integration efficiencies of each recombinase 102

Figure 4.3: Integration efficiencies of each second-generation recombinase 105

Figure 5.1: Central Metabolism for *M. elsdenii* 117

Figure 5.2: Fermentation profile of *M. elsdenii* ATCC 25940 butanol producing strains from lactate 128

Chapter I

Genetic tool development in non-model microorganisms for biofuel and chemical production

A version of this chapter was submitted by Lauren A. Riley and Adam M. Guss.

Riley, L. A., Guss, A. M., (2020) Genetic tool development in non-model microorganisms for biofuel and chemical production. Review, Submitted

Lauren A. Riley researched this paper while Lauren Riley and Adam M. Guss wrote and revised.

Abstract

Non-model microorganisms often possess complex phenotypes that could be important for the future of biofuel and chemical production. They have received significant interest the last several years, but advancement is still slow due to the lack of a robust genetic toolbox in most organisms. Typically, "domestication" of a new non-model microorganism has been done on an *ad hoc* basis, and historically, it can take years to develop transformation and basic genetic tools. Here, we review the barriers and solutions to rapid development of genetic transformation tools in new hosts, with a major focus on Restriction-Modification systems, which are a well-known and significant barrier to efficient transformation. We further explore the tools and approaches used for efficient gene deletion, DNA insertion, and heterologous gene expression. Finally, more advanced and high throughput tools are now being developed in diverse non-model microbes, paving the way to rapid and multiplexed genome engineering for biotechnology.

Introduction

The world's energy and chemical demand is ever-increasing, and currently, the demand for fuels and chemicals is primarily met with fossil fuels. Fossil fuels like petroleum not only provide the raw materials to make liquid fuel for the transportation sector, they also provide the building blocks needed for a wide variety of heavily utilized chemicals and materials, including plastics, perfumes, paints, fertilizer, and detergents (4). However, fossil-derived carbon is inherently unsustainable, and a

promising alternative is microbial conversion of renewable feedstocks. Microorganisms often have the ability to make drop-in or functional replacements for the fuels and chemical building blocks used today, but metabolic engineering is typically required to reach the titers, rates, and yields needed for a commercial process as well as to diversify the chemicals that can be produced.

Currently, model organisms like *Saccharomyces cerevisiae* and *E. coli* are most commonly used for metabolic engineering research and are often used for industrial biochemical production. Model organisms are attractive because they are industrially robust, decades of research have enabled deep understanding of the organisms, and they have a large genetic toolbox to enable rapid and simple modifications. However, engineering of synthetic pathways into these organisms is often most successful with short heterologous pathways, while long synthetic and heterologous pathways are generally not as robust and can cause a large metabolic burden (5). To overcome the challenges with current industrial biochemical production and to make next-generation fuels and chemicals profitable, recent research has focused on non-model organisms that natively have phenotypes of interest for each unique feedstock and processing method (7).

Industrial bioproduction of fuels and chemicals initially focused on first generation feedstocks such as starch and sucrose. These feedstocks are easy to breakdown and convert, but potential competition with food crops, limited greenhouse gas emission benefits, and environmental sustainability concerns makes these feedstocks problematic (9). Alternate feedstocks such as the sugars and aromatics in lignocellulosic biomass (10), sunlight or renewable electricity coupled with CO_2 fixation (11, 12), syngas (15), and waste plastic (17) have the potential to be cheaper. These feedstocks are also more environmentally friendly sources of carbon for fuel and chemical production, but bioconversion of these substrates is much more challenging. However, many non-model microbes often have evolved to utilize these feedstocks efficiently, making them potentially attractive platforms for bioconversion.

Unfortunately, many of the organisms with these native capabilities have poorly characterized metabolisms and are challenging to modify genetically, hindering metabolic engineering efforts.

Other traits potentially beneficial to industrial processing methods include tolerance to low pH, tolerance to high salt, or the ability to grow at high temperatures. These traits could streamline chemical production by reducing contamination, reducing the need for pH control and buffering, and increasing enzymatic reaction speed. These abilities are multi-gene, complex phenotypes that are difficult or impossible to engineer into model organisms. The more widespread use of non-model organisms could leverage these phenotypes to usher in a new era of biotechnology for biofuel and chemical production if sufficiently robust genetic tools can be developed.

Systems Biology in Non-Model Microorganisms

In addition to the lack of genetic tools, limited knowledge of microbial physiology is a major barrier to the use of non-model organisms for metabolic engineering. Traditionally, physiological characterization of a new organism was time-consuming and labor-intensive, with much of the new knowledge generated through biochemistry and characterization of substrate utilization and product formation. It was difficult to gain additional in-depth physiological information to uncover cell metabolism and regulation (5, 24). Recently though, genomic knowledge has been easier to acquire due to the rise of next generation sequencing techniques and the dramatic drop in the cost of genome sequencing (26, 27). Using basic genomic knowledge, the groundwork can now be laid for systems-level biological analyses, or “omics”. Enabling tools like transcriptomics and proteomics assists in addressing fundamental questions and provide measurements for all cellular components, and they are straightforward to apply to new microorganisms (24). The cost and time associated with these tools are also dropping, and the major challenge is now turning these large datasets into usable knowledge. Genome scale models can allow for several predictions, including growth rate and advantageous gene

modifications but they are limited due to novel pathways and unpredictable biological responses. Integration of multi-omics data to models helps give a more comprehensive dataset and helps to link the genotypes associated with the complex phenotypes of interest (29, 30). Having comprehensive knowledge of a microorganisms' metabolic pathways and flux is important to successful metabolic engineering (14). With the increased availability of affordable -omics analyses and DNA synthesis, the major remaining barrier to widespread adoption of non-model microbes is the lack of robust genetic tools.

Barriers to Genetic Modification and Enabling Transformation

In order to develop non-model organisms for bioprocessing, genetic modification is typically needed. The development of genetic tools requires the ability to efficiently transform DNA into a target organism. There are four primary barriers to overcome for successful transformation of bacteria: cellular uptake of foreign DNA, evasion of native immune systems that degrade foreign DNA, selection for transformants, and stable maintenance of foreign DNA by the microbial host. Each poses a challenge, and the largest barrier is often the evasion of host defense systems.

DNA entry into the cell

The first challenge, introducing DNA into cells, requires DNA to get past one or two membranes, as well as other physical barriers, such as peptidoglycan, to reach the cytoplasm. Depending on the organism, a variety of techniques can be utilized, including commonly used methods such as electroporation, conjugation, and natural competence. For electroporation, an electrical field is applied to a mixture of DNA and washed cells, which opens holes in the membrane, and enables the uptake of DNA (30, 31). Electroporation works broadly across phylogenetic groups, including in diverse bacteria, Eukaryotes, and Archaea, and therefore is often the first method attempted by researchers trying to transform a new organism. Conjugation enables DNA transfer through direct cell to cell contact from a

donor cell to the recipient. This technique has been used in diverse organisms using a donor *E. coli* strain carrying broad host range conjugation machinery (32-34). Conjugation can be more challenging into organism that have a vastly different optimal growth conditions than *E. coli*, such as halophiles or thermophilic anaerobes. Natural competence, on the other hand, is a mechanism driven by suite of genes that allows some bacteria to uptake DNA from their environment (37). This DNA is transported into the cell as single stranded DNA (ssDNA), which is usually highly recombinogenic, making it an especially good method of transformation when homologous recombination is desired. Transformation using this technique is often relatively simple in strains containing the natural competence genes where the strain of interest is incubated with DNA. Prominent examples of hosts that use natural competence for transformation include the Firmicute *Bacillus subtilis* and the Gammaproteobacterium *Acinetobacter baylyi* (38, 39). In strains that are not naturally competent, conjugation and electroporation are the most common methods of transformation, with electroporation often enabling the highest transformation efficiencies (40). Though these three techniques are the most commonly used, there are many less commonly used techniques reported in the literature, including phage-enabled transfer (41), chemically induced competence (42), sonoporation (43), biolistic bombardment (44), liposome-mediated fusion (45), protoplast fusion (46), and nanofiber piercing (47). Because these methods are not used as widely, their application to new non-model organisms is relatively underexplored.

Host defense systems

Prokaryotic organisms have evolved multiple defenses against foreign DNA, which can present a major barrier to DNA transformation. Restriction-Modification (RM) systems act as innate immune systems and are one of the primary systems used by prokaryotes to protect themselves against foreign DNA. Cells recognize and degrade DNA that is methylated differentially from that of its own DNA (48). Almost 90% of prokaryotes encode RM systems, with most encoding two or more (49). Microbes often also encode adaptive immune systems, including clustered regularly interspaced short palindromic

repeats (CRISPR). CRISPR systems natively protect prokaryotes against foreign DNA, for instance from bacteriophage. Viral derived DNA sequences are acquired during an infection and inserted into the chromosome where they may be transcribed to prevent against repetitive infection (50). Native CRISPR systems are unlikely to pose a major barrier to DNA transformation in most cases because the specific sequences acquired by the host CRISPR system are unlikely to be present in the DNA that is being transformed. Therefore, CRISPR will only be further discussed in this review in relation to their use as genetic tools. Several other defense systems have been discovered, but they are rare and/or poorly characterized. These include abortive infection (Abi), BREX, Dnd, Dpd, DISARM, pAgos, and others (51, 52). The impact of these systems on genetic transformation is currently unknown, but not yet observed to present major barriers.

To enable genetic modification, transformed DNA must evade the native immune systems of the host. RM systems are the most important immune system that DNA needs to evade, and many studies have shown the importance of this for successful transformation (53-61). Of the four classes of RM systems, Types I, II, and III are typically comprised of two primary components, a restriction enzyme and a methyltransferase. The methyltransferase modifies a specific base within a specific motif throughout the host chromosome so that the genome is protected from restriction enzyme cleavage, with modifications including 6-methyladenine (m^6A), 4-methylcytosine (m^4C), or 5-methylcytosine (m^5C). The restriction enzyme cleaves the same motif in unmethylated DNA that enters the cell. Type IV systems, on the other hand, only consist of a nuclease that cleaves methylated DNA, with the recognized motifs different than those targeted by native methyltransferases from Types I, II, and III. The sequence specificity of Type IV systems is typically poorly characterized and is a ripe area for future research.

To evade RM systems, the first step is to identify the methylated motifs within a cell. Microbial methylation sites are most commonly determined through genome sequencing with single molecule real time (SMRT) sequencing on the PacBio platform (62, 63). SMRT sequencing routinely identifies m^6A

and m^4C motifs and, with a significantly lower efficiency, m^5C motifs through kinetic delays in nucleotide incorporation at modified bases during sequencing. While Eukaryotic methylome analysis often employs Whole Genome Bisulfite Sequencing (WGBS) to detect m^5C motifs, it has rarely been used in bacteria. However, robust detection of m^5C is critical for complete methylome analysis (54). Other emerging techniques, such as nanopore sequencing, may also be used for methylome analysis (64, 65). New England Biolabs (NEB) has created a database, REBASE, of all known RM systems (both experimentally determined and computationally predicted) and publicly available microbial methylomes to help in the identification of RM systems (66). Not all methyltransferases are associated with restriction enzymes; they can also play a regulatory role in DNA replication, gene expression, and DNA mismatch repair (67). However, the methylated motifs reveal the maximum number of sites that could be subject to restriction in a given host.

Once the methylated motifs are identified, several methods have been used to evade the corresponding RM systems. One approach is transforming DNA that lacks the identified motifs, or mutating DNA so that it no longer contains the motif (54, 68, 69). Recently, software has been designed to aid in the process of eliminating restriction sites (70). This can often be done for uncommon motifs, but it cannot always be used for shorter, more common motifs, like four base pair recognition sequences that may exist in plasmid origins of replication or other DNAs that require a specific sequence, such as those needed for homologous recombination.

Another way to overcome RM systems is by methylating DNA of interest in the same way as the target host prior to transformation, such that the organism does not recognize it as foreign. The most common approach to achieving methylation is to express the target organism's restriction-associated methyltransferases in *E. coli*, and then isolate the DNA of interest from this *E. coli* strain to properly methylate it prior to transformation. Early pioneering work demonstrated transformation of *Clostridium acetobutylicum* by first methylating plasmid DNA in *E. coli* with the methyltransferase from *B. subtilis*

phage phi3T (53). This approach was expanded with a method called Plasmid Artificial Modification (PAM), which introduced all the methyltransferases from a given strain into *E. coli* using plasmid-based expression, followed by isolation of the DNA of interest out of the PAM host to properly methylate it prior to transformation of the target organism (71). More recently, informed by methylome analysis, only those methyltransferases identified as important can be introduced to the *E. coli* chromosome to more stably express the enzymes and to only mimic the methylome data.

DNA can also be properly methylated through in vitro methylation (72). Some enzymes are commercially available, which can be especially useful for 4-base motifs that are targeted by commercially available methyltransferases. For those motifs that are not methylated by commercially available enzymes, cell-free extracts can be used, as demonstrated in *Heliobacter pylori*, to increase transformation efficiency (73).

RM systems can also be partially evaded by the transformation mechanism chosen. Often, electroporation is unsuccessful when using improperly methylated DNA and, in these cases, using conjugation can yield successful transformation. The conjugation machinery transfers single stranded DNA (ssDNA) that can avoid restriction enzyme cleavage until it has time to become properly methylated (74). Similarly, natural competence also imports ssDNA and can reduce the impact of restriction systems (75).

Another important consideration when transforming a new bacterial strain is the presence of Type IV restriction systems, which degrade methylated DNA motifs. Researchers typically isolate plasmid DNA from *E. coli* prior to transformation into their desired host, and commonly used laboratory *E. coli* strains encode two major DNA methylases, *dam* and *dcm*, targeting G(m^6A)TC and C(m^5C)WGG, respectively. Many potential target strains encode *dam*, either as part of an RM system or as a housekeeping methyltransferase, and a few encode *dcm*, but many do not. Therefore, it is important to

isolate plasmid DNA from an appropriate *E. coli* genetic background, using strains deleted for *dam* and/or *dcm* when the target organism does not methylate the same site. This prevents unnecessary DNA methylation and subsequent degradation by Type IV systems, which has been demonstrated in many strains (76-78). It is also important to utilize an *E. coli* strain that lacks the native Type IV systems (*mcrA*, *mcrBC*, and *mrr*) when expressing methyltransferases. Methylation of motifs recognized by the native type IV systems would cause *E. coli* to restrict its own chromosome, which would kill the *E. coli* strain.

RM systems are frequently acquired via horizontal gene transfer, and therefore methods to improve transformation in one strain typically do not directly translate to closely related strains. These systems tend to be hyper-variable within an environmental community of closely related strains, such that if a phage happens to avoid one cell's restriction systems and becomes methylated like that host, it does not eliminate the entire population (49). Therefore, methylome analysis and RM system evasion must be developed for each desired host, even if they are strains of the same species.

Maintaining and Selecting for DNA

Once DNA has entered the cell and evaded degradation, it needs to be maintained during cell division, and transformed cells must be selected from the untransformed cells that constitute the majority of the population. DNA can be maintained either through autonomous replication or through chromosomal integration. Plasmid-based autonomous replication is typically the best way to achieve the highest transformation efficiency for a new microbe and is the most commonly used method to demonstrate initial transformation. For metabolic engineering, however, plasmid-based gene expression has drawbacks, such as high copy number artifacts, variable plasmid copy number, the need for continuous antibiotic selection, and a high metabolic burden. Therefore, chromosomal integration is

also an important tool, and different approaches to developing DNA integration genetic tools will be discussed in more detail below.

Development of a plasmid-based DNA maintenance system relies on identification of an origin of replication that functions in the target organism. The most common vectors used for cloning and replication in *E. coli* use origins such as pUC, ColE1, and p15a, each with varying copy numbers (79), but often these origins do not replicate in non-model bacteria. Therefore, origins that function in diverse bacteria are needed. Broad host range plasmids are capable of transfer and maintenance in bacteria from different phylogenetic subgroups, examples include pBBR1, pRK2, pBC1, and many others (80). Plasmid libraries have been formed around the broad host-range origins of replication and used in a variety of organisms (81). Alternatively, origins can come from native plasmids of the strain or close relatives (82, 83). If none are present, replicating plasmids may be constructed by cloning the chromosomal origin of replication to create a mini-chromosome (84, 85).

Next, one must select for cells that were transformed and eliminate cells that were not. The most common positive selectable markers are antibiotic resistance genes (Table 1.1). To determine the marker(s) most likely to work, the minimum inhibitory concentration (MIC) is determined by exposing the host bacteria to increasing levels of a panel of antibiotics. This is most easily done in liquid growth medium, but higher concentrations of antibiotic may be needed for selection on agar plates. Another consideration when choosing antibiotic resistance markers is compatibility with an organism's growth conditions. For example, when working with thermophiles, one should focus on antibiotics that are more stable at higher temperatures like thiamphenicol and kanamycin (86) and thermotolerant selectable markers (87). A less commonly used alternative to antibiotics is nutritional selection. In this

Table 1.1: Commonly used antibiotics and their corresponding selectable markers

Antibiotic	Marker	Class
Kanamycin/Neomycin	*neo*	aminoglycoside
Chloramphenicol/Thiamphenicol	*cat*	chloramphenicol
Erythromycin	*ermB, ermF*	macrolide
Ampicillin/Carbenicillin	*bla*	beta-lactam
Tetracycline	*tet, others?*	tetracyclin
Bleomycin	*ble*	glycopeptide
Gentamicin	*aacC1*	aminoglycoside
Streptomycin and Spectomycin	*aadA*	aminoglycoside
Zeocin	*zeo*	glycopeptide
Apramycin	*apr*	aminoglycoside
Thiostrepton	*tsr*	cyclic oligopeptide
Puromycin	*pac*	aminonucleoside
Hygromycin	*hph*	aminoglycoside
Nourseothricin	*nat*	aminoglycoside

case, an auxotrophic strain is generated by deletion of an essential nutrient gene and transformed with a plasmid encoding the missing biosynthesis gene(s). The resulting transformants are selected on media lacking the target nutrient (88, 89). Libraries of modular plasmids combining antibiotic resistance markers, origins of replication, multiple cloning sites (MCS), and other genetic parts have been developed for different classes of organisms, and they are very useful for the rapid testing of genetic parts (90-92).

Chromosome Modification Tools in Non-model Microorganisms

After demonstration of initial transformation, more advanced tools can be developed to enable efficient genome editing and ultimately high throughput strain engineering. Genome editing can allow for gene deletions, insertion of heterologous pathways, point mutations, and altered gene regulation. Enabling genome integration and deletion tools facilitates rational metabolic engineering, where competing production pathways can be eliminated, and new pathways can be stably introduced. Development of more sophisticated tools is critical to the engineering of non-model organisms for high rates and titers of desired products.

Homologous Recombination

Homologous recombination is a naturally occurring mechanism essential for DNA repair in bacteria, and it can be leveraged to rationally introduce modifications to an organism's chromosome. The most commonly deployed method uses a non-replicating plasmid-based technique to create a scar-less mutation (93). Typically, this is mediated by nuclease-helicase complex, RecBCD, and the single stranded DNA-binding DNA repair protein, RecA (94). To use homologous recombination for gene deletion, a plasmid is required that contains a selectable maker, a counter selectable marker, and DNA that is homologous to the upstream and downstream regions of the gene targeted for deletion (Figure 1.1A). The lengths of homologous DNA regions are often around 500-1000 bp each. After transformation

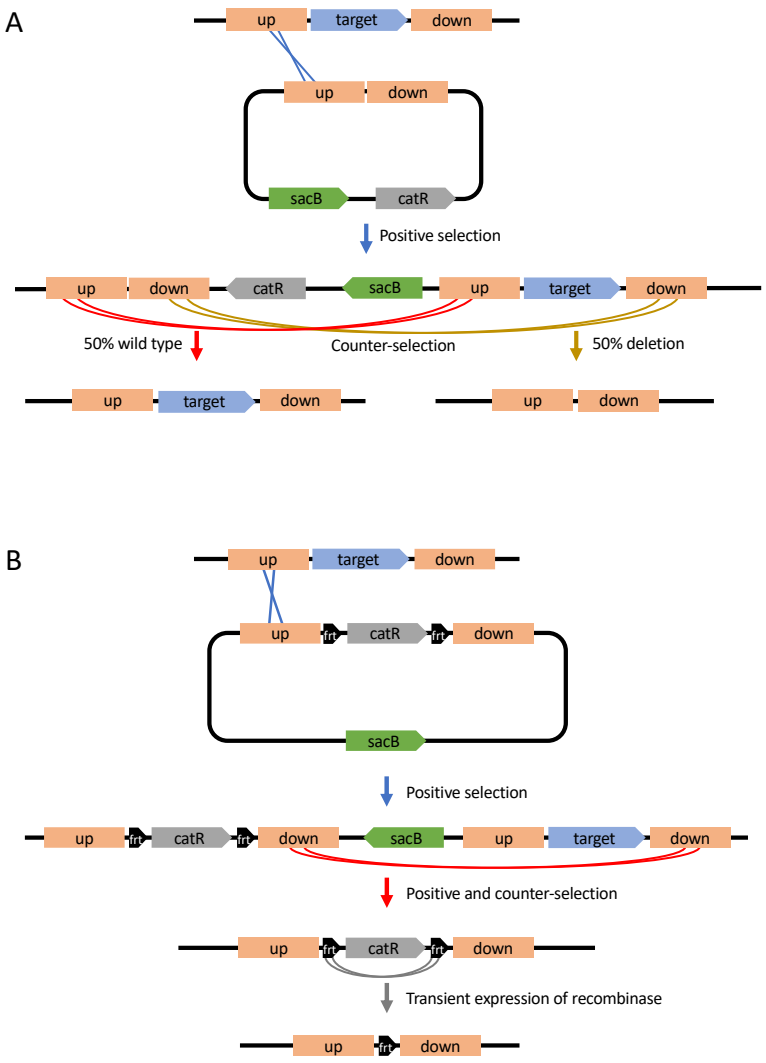


Figure 1.1: Methods using homologous recombination for gene deletions

Details for each are described in the text A) Basic homologous recombination using a non-replicating plasmid where 50% of colonies will be the desired deletion, B) Homologous recombination using the positive selectable marker between the homology arms, flanked with recombination sites, and C) Homologous recombination using a replicating vector and three homology regions. Orange, homology regions, with “up” representing the DNA region upstream of the target gene, and “down” representing the DNA region downstream of the target gene; blue, target gene for deletion; green, counter-selectable markers; grey, positive selectable marker; crossed lines, sites of recombination.

C

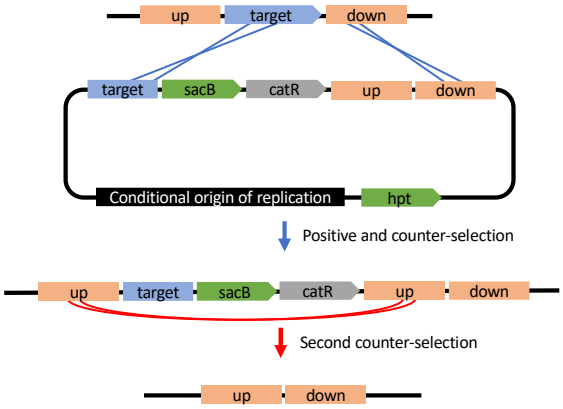


Figure 1.1: Continued

of the plasmid into the microbe, the plasmid recombines into the chromosome at one region of homology to integrate the entire plasmid, which is selected via the positive selectable marker contained on the plasmid. A second recombination event will resolve the merodiploid into either the parent chromosome or a gene deletion. If the first recombination event is repeated, the strain reverts to wild type, but if the recombination event occurs in the second region of homology, the targeted gene is deleted. When there is no fitness defect associated with the deletion, the frequency of deletion should be approximately 50%, which can be easily screened via PCR.

Selection for resolution of the merodiploid requires a counter-selectable marker, which allows selection for loss of DNA. Numerous counterselections have been demonstrated (Table 1.2), including *sacB* in Gram-negative bacteria and *pyrF* and *hpt* in Gram-positive bacteria (93, 95-98). Some of the markers are less desirable than others, though. Markers such as *sacB* often do not function well in Gram-positive bacteria (95). There are also markers such as *pyrF* that can act as both a selectable and counter-selectable marker making them more flexible (99-101), but in many cases, they also result in auxotrophy, requiring growth on minimal medium for prototrophic selection of transformants. Many genes used as counter-selectable markers are natively encoded in the host genome. In these cases, the native copy must first be deleted before it can be used as a counter-selectable marker. However, the presence of a native counter-selectable marker also presents an opportunity for metabolic engineering, as it can serve as a simple site for insertion of heterologous DNA with a direct selection for gene replacement. Importantly, inactivating mutations (e.g., frameshifts, active site mutations, etc) in counter-selectable markers have the same phenotype as recombinants and arise spontaneously, resulting in a background of cells that do not contain the targeted deletion. Therefore, counter-selections often require more screening to find correct clones relative to positive selection. In the absence of reliable counter-selectable markers, screens can be used, but this is much less efficient and should be avoided if possible (102).

Table 1.2: Commonly used counter selectable markers and associated selections for homologous recombination in bacteria.

Marker	Selection	Counterselection
sacB (1)		Sucrose
upp (2)		5-fluorouracil
hpt (3)		8-azahypoxanthine, others
tdk (6)		5-fluorodeoxuridine
pyrF/ura3 (8)	uracil prototrophy	5-fluoroortic acid
pheS(A294G) (11)*		*p*-chloro-phenylalanine
codA (12)		5-fluorocytosine
Inducible *mazF (13)*		N/A; Expression is toxic
galK (16)		2-deoxygalactose
apt (19)		2-fluoroadenine
trpC		5-fluoroanthranilic acid
rpsL(strA) (20)		streptomycin
tetAR (22)	Tetracycline resistance	Fusaric acid
thyA (25)		Trimethoprim
Inducible *ccdB (28)*		Expression is toxic
oroP (31)		5-fluoroorotate
pta-ack (32)		chloroacetate

Another approach is to place the antibiotic resistance marker between the homology arms, allowing for direct selection of the gene deletion (Figure 1.1B)(103). This approach is especially useful when the mutant phenotype is deleterious, which would make the approach in Figure 1.1A challenging. This approach can also work in the absence of a counter-selectable marker when using conjugation or natural competence and selecting directly for the double-recombination event that replaces the gene target with the antibiotic resistance gene. Often the antibiotic resistance marker is flanked by recombination sites, like *frt* or *lox*, so that the marker can be removed after integration into the chromosome by introduction of the corresponding recombinase, like Flp or Cre (103-105).

The frequency of homologous recombination varies widely in bacteria and in some cases is a rare event, making the transformation efficiency of integrating plasmids multiple orders of magnitude lower than for replicating plasmids. Therefore, replicating plasmid-based homologous recombination methods can be an attractive approach when the combination of transformation efficiency and homologous recombination in the target organism is low. However, because the plasmid replicates autonomously, it can be challenging to select chromosomal modifications; therefore, alternate approaches are needed.

One option is to use a conditional origin of replication, such as a temperature-sensitive plasmid. This allows the transformation event and the homologous recombination event to be uncoupled, where the plasmid can be transformed at a permissive temperature, generating a whole population of cells that contain the plasmid. Upon moving the culture to the non-permissive temperature, one can select for the cells that have performed recombination. Selection then can proceed as in Figure 1.1A.

Alternatively, introduction of a third region of homology and a second counter-selectable marker can allow for selection of each recombination event and plasmid loss (Figure 1.1C) (106). In this approach, the replicating plasmid is first transformed into the strain. Then, while maintaining selection for the antibiotic resistance gene, selection is performed against the counter-selectable marker that is

on the plasmid backbone. This selects for two recombination events to insert the antibiotic resistance onto the chromosome and simultaneous plasmid backbone loss. The resulting merodiploid can then be resolved by selecting against the second counter-selectable marker, resulting in the desired deletion.

Homologous recombination based genetic tools have been used in a variety of non-model organisms to increase biochemical production and enable non-native carbon catabolism though the deletion of competing pathways and insertion of heterologous genes (107-110).

Recombineering

Methods to increase the efficiency of homologous recombination have been developed, often called recombination-mediated genetic engineering, or recombineering (111, 112). By increasing the recombination efficiency, construction of gene deletions, insertions, and point mutations is more efficient. Recombineering uses proteins like those derived from the lambda bacteriophage Red complex, Beta, Exo, and Gam, to integrate single stranded and linear DNA into the genome, often using short homology arms to guide recombination. Recombineering is widely used in *E. coli*, and the genes associated with lambda Red have also been demonstrated to function in other proteobacteria, like *Pseudomonas putida* (113). Recombineering can be difficult to develop in non-model organisms because lambda Red genes are often not functional in those hosts. Considerable effort has gone towards identifying genes analogous to lambda Red and the recombination protein RecT in phylogenetically diverse organisms (114, 115). Using this approach, recombineering has been demonstrated in a variety of other strains, including *Clostridium acetobutylicum* (116), *Lactococcus lactis* (117), *Clostridium thermocellum* (118), and others (114, 119, 120). The recombinases can then be used to enable ssDNA recombination for point mutations, increased efficiency of vector integration, and increased CRISPR-Cas editing efficiencies. Another benefit of recombineering is often the ability to use much shorter

homology arms. Native homologous recombination techniques often requires 300-1000 bp of homology for efficient recombination, while recombineering techniques can shorten this to as few as 20 bp.

Random DNA Insertion

Non-homology-based approaches are also useful when engineering non-model microbes. One such approach is transposition. Transposons allow random integration of DNA segments throughout the genome with no required homology. Therefore, transposition is traditionally used to create a library of insertional disruptions that can be screened for particular phenotypes. These phenotypes can then be correlated with genotypes by identifying the site of insertion (121). Commonly used transposons include Tn5, which inserts mostly randomly but is known to favor "hotspots", and *Himar1*, which can insert randomly between any TA dinucleotide (122, 123). Recently, random transposon mutagenesis has been combined with high throughput sequencing for genome-scale analyses, called TnSeq (124). Further combination with randomly barcoded transposons simplifies reverse genetics, where a pooled library of barcoded insertional mutants can be created, and the bar codes can be mapped to the genome to identify disrupted genes (125). A barcoded library can be an incredibly powerful tool for genome-scale analyses. A list of putatively essential genes can be compiled by identifying the genes that were not disrupted in the library. Furthermore, genes related to phenotypes such as solvent and organic acid resistance, product formation, and carbon assimilation can be identified by finding barcodes that increase or decrease in abundance under selective conditions (124, 126, 127).

Transposons can also be used to insert heterologous pathways into bacteria (128), which has been applied to non-model microbes such as *C. ljungdahlii* for acetone production (129) and *Acidothiobacillus ferroxidans* for isobutyric acid biosynthesis (130). Transposons have also been combined with the Cre-*lox* site-specific recombination system to insert a landing pad, followed by site-specific insertion of heterologous pathways (131). Transposons are especially useful for application in

organisms that lack other genetic tools such as homologous recombination for integrating DNA into the chromosome. Integrating DNA randomly has its drawbacks, however. The transposon often integrates into coding regions of the genome, which can interfere with the fitness of the strain, and identifying the location of transposition requires additional effort. Random integration also does not enable comparison of a construct across genetic backgrounds because the transposition event occurs in different loci, which can result in differing levels of gene expression for the construct (132).

Site-specific DNA Integration

Site-specific recombinases have been a valuable tool for DNA insertions and excisions where the enzyme catalyzes a recombination event between two specific DNA sequences, sometimes called attachment (*att*) sites. The most commonly used site-specific recombinases are the Cre and Flp that enable a reversable recombination (integration or excision) reaction at identical sites, *loxP* and *frt* respectively (133). Both systems have been used to remove antibiotic resistance markers by integrating a cassette into the chromosome with *loxP* or *frt* flanking the marker followed by transient expression of the corresponding recombinase (111). For example, Flp/*frt* is widely used in the *E. coli* lambda Red system to remove the antibiotic resistance marker, and it has also been used in a variety of non-model organisms, including thermophiles, methanotrophs, and biofuel producing *Clostridium* species (96, 103, 110, 134). One drawback to using Cre and Flp, however, is that the *loxP* or *frt* scar left behind can cause genetic instability when multiple modifications are stacked into a single strain, leaving behind multiple, identical copies of the scar that can all serve as future substrates for the recombinase. Using mutated sites can help overcome the genome instability (135, 136).

A second group enables a site-specific recombination event at non-identical attachment sites, typically called *attB* (bacterial) and *attP* (phage), reflective of the fact that these recombinases often derive from bacteriophage that lysogenize their hosts. Recombination creates two new sites, *attL* (left)

and *attR* (right). The integration event is unidirectional because the *attL* and *attR* are not substrates for recombination, so the recombinase cannot excise the construct from the genome (137). Some unidirectional recombinases, including the λ integrase, require host machinery to function, which limits use outside of their native host or very close relatives. λ phage also has relatively large attachment sites of 350-450 bp each. Alternatively, the large serine recombinases, as exemplified by the archetypical ΦC31 integrase, can function in a broad-host range because they do not require host machinery. This has led to their development as genetic tools in a wide variety of microbial hosts (138-140). Due to their utility for rapid and high efficiency integration of DNA, they have been used to test genetic libraries at single copy on the chromosome, such as promoter libraries (138), and to integrate heterologous DNA (141, 142). Genetic tools developed with serine integrases can rely on native or pseudo-*att* sites that are present in the genome (143) or non-native *att* sites that have been previously integrated into the genome (141, 144, 145). One of the most commonly used integrases is ΦC31, which was identified from *Streptomyces* and is commonly used in a DNA integration system in this genus (146). Multiple orthogonal recombinases have been identified and characterized (137), setting the stage for them to be used much more broadly. Another type of site-specific recombinase includes transposases that target highly conserved sequences like Tn7 and Tn1545 (132). Tn7 has been demonstrated to function in more than 20 bacterial species. It enables integration into the highly conserved *attTn7* site downstream of the *glmS* gene. This makes the use of site-specific transposons more desirable for some applications than the random transposons for integration of heterologous constructs (132).

CRISPR/Cas

A recent and potentially universal tool for engineering non-model microbes uses Clustered Regularly Interspaced Palindromic Repeats (CRISPR) and CRISPR associated (Cas) proteins. Native CRISPR-Cas systems are comprised of a CRISPR array containing spacers separated by short repeats. Spacers become incorporated into the array when the microorganism encounters invading DNA to

enable the microorganism to recognize and attack that same sequence when they encounter it again (147). Many of the systems cleave DNA at this sequence using a Cas nuclease that is directed by a guide RNA (gRNA). Cas9 was first identified in *Streptococcus pyrogenes* (spCas9), and it introduces a double stranded break (DSB) at a specific sequence guided by a sgRNA adjacent to a protospacer adjacent motif (PAM) (148, 149). Natively, the gRNA is derived from the CRISPR array, and in the case of the most utilized type II Cas nuclease, Cas9, the array is transcribed and processed (crRNA) with trans-acting RNA (tracrRNA) into a single guide RNA (sgRNA)(150). In heterologous systems, the sgRNA can be transcribed as a single RNA, simplifying the number of genetic parts needed for engineering. If a template containing the desired modification is also included, the DSB can be repaired, resulting in an edited genome.

CRISPR-Cas systems have been leveraged in a wide variety of organisms to introduce scarless point mutations, insertions, and deletions in the genome. The most widely used CRISPR nuclease is spCas9 because, unlike many of the other Cas proteins, it is a single enzyme that only requires the introduction of a synthetic gRNA and the Cas9 nuclease to enable genome editing. CRISPR-spCas9 was first introduced as a genome engineering tool for bacteria in *E. coli* (149, 151) and has now been implemented in a wide variety of species using a plasmid for the repair template and an organism's native homologous recombination machinery, as recently reviewed in (149). Cas9-based editing has been further developed by combination with recombineering machinery to enhance the efficiency of DSB repair, and allowing use of a linear recombineering template with homology to the edited region (151, 152). The combination of CRISPR and recombineering can increase the rate of mutation close to 100% due to the lethality of the DSB. Functionally, this makes CRISPR-Cas a potent and targetable counter-selectable marker. The combination of recombineering and CRISPR has been used in many proteobacteria, but it can take substantial effort to transfer to organisms where transformation efficiency is low and recombineering tools do not yet exist.

The nuclease spCas9 is most commonly used for genome editing across many types of bacteria. However, Cas9 expression can be toxic in a variety of hosts (145, 153). Toxicity can be overcome by placing the gene under an inducible promoter so that it is only expressed when needed (154), but this is only possible in organisms where inducible promoters have been developed. Toxicity can also be overcome in some organisms by mutating one of the active sites, resulting in Cas9n, so that it can only nick one strand of DNA rather than creating a DSB (155, 156). Another issue with the commonly used Cas9 enzymes is that they are unable to function in thermophiles; therefore, thermophilic Cas9 enzymes have been identified and implemented along with thermophilic recombineering machinery for engineering thermophiles (118). Although spCas9 is the most widely used nuclease, other CRISPR-Cas systems have been identified to overcome problems with spCas9. These nucleases also have different cleavage patterns and target different PAM sequences. Cpf1, a member of the Cas12a group, has been used in several organisms for genome editing, including those where spCas9 did not function (145, 157, 158). Native CRISPR systems have also been leveraged for genome editing (159, 160). There is significant interest in CRISPR based tools for non-model organisms, and it has been reviewed in more depth elsewhere (148, 149, 161, 162).

CRISPR/Cas for gene regulation

While CRISPR-Cas9 has been widely demonstrated for DNA insertions and deletions, it can also be used for gene regulation by silencing or activating transcription using a catalytically inactive Cas9 (dCas9). In this case the dCas9 is unable to act as a nuclease but is still able to bind DNA at a targeted sequence. In CRISPR interference (CRISPRi), a gRNA-targeted dCas9 binds to the promoter of the desired gene, therefore blocking either transcription initiation or elongation and inhibiting gene expression (163). This enables gene knockdowns in a regulated and reversable way, and expression of multiple gRNAs can allow multiplexing of gene knockdowns (164). CRISPRi has been used a wide variety of non-

model microorganisms to both study the activity of essential genes and to knockdown expression of competing pathways to increase production of fuels and chemicals (165-168).

While CRISPR can be a powerful genome engineering and gene regulation tool, it is often difficult to optimize in non-model bacteria. Therefore, research has gone into developing a modular CRISPRi system that can be deployed across phylogenetically diverse bacteria. Mobile-CRISPRi has demonstrated the gene knockdown efficacy in a diverse set of pathogenic microbes, though this system has yet to be used in organisms of interest for industrial use (169). While CRISPRi represses transcription, dCas9 can also be used to activate transcription via CRISPRa. In CRISPRa the dCas9 is fused to a transcriptional activator so that when dCas9 binds upstream of a gene it recruits RNA polymerase leading to increased transcription. While CRISPRa has been used in *E. coli* using both PolII (170) and SoxS (171), this tool is still in the early stages of being applied to other prokaryotes (172).

Gene Expression Tools in Non-model Microorganisms

The ability to reliably fine tune gene expression is an important synthetic biology tool and can play an important role in increasing biofuel and chemical yields from engineered pathways. Transcription and translation levels can be controlled by several different components including promoters, riboswitches, ribosome binding sites (RBS), and terminators. Genes can be constitutively expressed, where protein production levels are largely determined by the strengths of the promoter and RBS, or genes can be regulated where expression can be turned “on” or “off” by inducible promoters and riboswitches. Reporter genes can provide an easily assayed output to help determine the impact of different genetic parts on expression levels, and they also having applications in the development of biosensors for the detection and control metabolite levels during bioconversion. Developing these gene expression tools for non-model microbes is critical to enabling rational metabolic pathway optimization for improved product formation.

Constitutive gene expression

The simplest way to modulate gene transcription is through the evaluation of promoters with varying activity levels. Often, strong characterized promoters from model organisms, like P_{tac} from *E. coli*, are used but they do not always confer robust gene expression in non-native hosts. Identifying and implementing native promoters can overcome this issue. Strong promoters often drive expression of genes in central metabolism and identifying those genes in a target organism may help in identifying useful promoters. RNAseq can be employed on cells grown under desired culture conditions (173-175), where genes with high transcript levels may be indicative of a strong promoter. Further characterization of the identified promoters through a time course study and under various growth conditions may identify those that are constitutive (175).

Overexpression with strong promoters is not always desirable, especially if an intermediate in the pathway is toxic or if a protein is membrane-bound. Furthermore, when many genes need to be expressed, the metabolic burden of protein production can become substantial (176). Therefore, the development of a promoter library with a range of promoter strengths is also useful. Typically, the closer the -35 and -10 sequences are to consensus, the stronger the promoter is. Mutating each of the bases in the sequence and varying the space between them has been showed to vary promoter strength (177, 178)**.** Mutated promoters can also be obtained through error prone PCR or through site directed mutagenesis (174). Promoter libraries have been characterized in a wide variety of non-model organisms (179-182). Other regions that impact gene expression and can be varied include the UP element and the ribosome binding site (RBS). UP elements are sequences upstream of the -35 box in the promoter that bind to the alpha subunit of the RNA polymerase, greatly increasing expression levels (183). While underexplored for many non-model microbes, UP elements can bring an additional boost to expression when very high transcription levels are desired (184-186). The RBS recruits the ribosome and it is an important control point for translation initiation. The RBS is commonly used to tune gene

expression through alterations of the sequence and the space between the sequence and the start codon. Libraries of RBSs have been created in combination with promoters to fine tune gene expression for fuel and chemical production (181, 187, 188). An RBS calculator has been created to aid in the creation of RBS libraries and to predict translation initiation rates (189).

Regulated gene expression

The ability to regulate gene expression is an important tool for both strain engineering and bioproduction. Applications include expression of a toxic gene (e.g., Cas9) for a short period (154, 190), balancing of metabolic flux with a biosensor (191), or turning on a production pathway once cell growth has reached stationary phase to reduce the metabolic burden (192-194). They are also important to determine whether a gene is essential under the conditions being tested. One mechanism of creating inducible promoters involves the use of transcription factors. Well characterized promoters from *E. coli* such as lactose- and arabinose-inducible promoters have been transferred, sometimes with mutations to enable functionality, to a wide variety of other organisms, including species of *Clostridium*, *Pseudomonas*, *Bacillus*, and *Ralstonia* (81, 195). Other inducers have also been commonly used such as tetracycline (196), xylose (197), and nisin (96). Newer promoters such as the Jungle Express expression system for proteobacteria (198), and less commonly used ones such as laminaribiose (199) also have the potential to be useful for engineering various non-model organisms. However, transferring well characterized promoters into non-model bacteria can still be challenging. Many result in "leaky" expression or lack regulatory proteins, such as repressors or activators, not present in heterologous hosts. In addition, even if regulatory proteins are transferred with the promoter, transport of the inducer molecule may be an issue.

Gene expression can also be regulated post-transcriptionally. Riboswitches are regulatory mRNA elements in the 5' untranslated region that are capable of regulating gene expression through small

molecule-induced structural switching (200). They add a layer of regulation but are less explored for the engineering of non-model microbes relative to model organisms like *E. coli*. In this mechanism mRNA forms two distinct shapes depending on the presence of a specific small molecule, which either blocks or allows translation. Examples of the use of riboswitches for engineering of non-model microbes include theophylline for inducible production of biofuels and other compounds (201), lysine to balance flux for competing pathways (200, 202), and flavonoids such as naringenin for increased production of flavonoids (203). Recently, thermophilic riboswitches have also been identified and used to regulate biofuel production (204). Riboswitches are particularly beneficial for regulating gene expression in non-model microbes because they function independently of host-associated proteins and without the need for organism-specific promoters. For example, a set of theophylline riboswitches have been demonstrated functional in a many species including multiple cyanobacteria, Mycobacterium, and Streptomyces. These switches have also proven to be superior for regulation of gene expression when compared to traditional IPTG-inducible promoters (201).

Terminators

Transcriptional terminators in *E. coli* have been well characterized and are an important tool for gene expression. Terminators can impact mRNA stability and expression levels of adjacent genes (205) Simple terminators that do not require additional termination factors have been used widely in diverse genera. Some terminators, like those from the T7 phage, can be inefficient at termination, causing read through on multigene constructs. However, several commonly used terminators contain repeats that increase the stability of transcriptional pausing and increase reliability, though this can cause issues when using the same terminator repeatedly use within a single construct (174). Efficient and reliable control of multigene constructs for metabolic engineering of bacteria therefore requires several well characterized terminators. Libraries (175) and terminator prediction programs (176) have been

developed, but terminator research in non-model bacteria lags behind other parts for gene expression like promoters.

Reporter Genes

Reporter genes encode proteins that can be tracked or quantified, often visually or spectrophotometrically. They have many applications in flow cytometry, microscopy, protein localization, and microbial co-cultures, and they are particularly useful when screening promoter and gene expression libraries. Enzymes can act as reporter genes, including common ones such as *lacZ* (β-galactosidase) and *uidA/gusA* (β-glucuronidase), where a substrate (e.g., X-gal) is cleaved to generate a colored compound, and enzyme activity is proportional to enzyme abundance. Other enzymes include chloramphenicol acetyltransferase (*catP*) and alcohol dehydrogenase (*adhE*). Determining enzymatic activity can be laborious, which makes these enzymatic reporter genes less useful, and they are typically used only as end-point assays (206).

Fluorescent reporter genes are the most commonly used type of reporter gene because they can provide real-time measurements of gene expression. In aerobic microbes, several different genes encoding fluorescent protein have been engineered, including the green fluorescent protein, *gfp*, and the red fluorescent protein mCherry. Each of the fluorescent proteins have different characteristics from color to brightness to stability (207). These genes have been used in a wide variety of applications. The fluorescent genes are used in transcriptional fusions to indicate the strength of promoters and expression vectors during bacterial cell growth (177, 208). The simplicity and non-invasive nature of fluorescent proteins allows easier tracking of the dynamics of gene expression and allow much higher throughput. This higher throughput enables the use of these genes as biosensors by fusing the fluorescent reporter to regulated promoters that respond to the presence of metabolites, such as the target product molecule, metabolic intermediate, or toxic compound (209-211). One caveat is that

reporters are only a proxy for gene expression, and the expression level of heterologous genes can depend on the exact genetic context, especially if mRNA secondary structure or stability changes. While these fluorescent reporter genes are incredibly useful in aerobic bacteria, they require O_2 to form the fluorescent chromophore.

Many biotechnology-relevant organisms are strict anaerobes, where O_2-dependent reporters can be more challenging to use. The use of such reporter genes requires that anaerobic cultures be brought into an aerobic environment, which can be lethal, and the cells often need to be washed in a time-consuming process. Therefore, they cannot be used in real-time growth experiments anaerobically. Flavin binding fluorescent proteins (FbFPs) do not require O_2 for fluorescence, making them potentially useful tools for anaerobes. Examples of FbFPs include iLOV, BsFbFP, PpFbFP, and EcFbFP which have been demonstrated in a variety of organisms including many species of *Clostridia* (212). However, FbFPs are not as bright as aerobic fluorescent genes, which make them more difficult to quantify. To overcome the limitation of FbFPs, a fluorescence-activating and absorption-shifting tag (FAST) protein has been developed. FAST relies on the presence of an exogenously added ligand that is only fluorescent when bound to FAST. Two color signals have been developed, a green yellow (YFAST) signal where 4-hydroxy-3-methylbenzylidine-rhodanine (HMBR) binds FAST and a red signal (rFAST) where 4-hydroxy-3,5-dimethoxybenzylidene-rhodanine (HBR3,5DOM) binds. FAST produces a fluorescence signal similar to that of GFP and has recently been used in a few organisms (213, 214). This tool has the potential to revolutionize the use of reporter genes in anaerobic microbes. This includes the ability to monitor real-time gene expression, screen promoter libraries, study protein localization, and develop other high through-put tools such as biosensors.

Often development of a new non-model microorganism is done on a case by case basis and it can take years to develop transformation and basic genetic tools. A systematic pipeline for enabling knowledge through omics, developing transformation, and developing genetic tools for a new organism

is needed to increase the speed of development. As next generation tools, like CRISPR-Cas, are developed more rapid and advanced metabolic engineering can occur, better enabling the future of non-model organisms. Introducing new tools into non-model microorganisms comes with challenges but as more tools are developed for phylogenically diverse organisms this process will become easier and more systematic. CRISPR-Cas, recombineering, and site-specific recombinases each have numerous applications, allowing for engineering of diverse phenotypes and helping to further the utility of non-model microorganisms in bioprocessing.

References

1. Gay P, Le Coq D, Steinmetz M, Berkelman T, Kado CI. Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. J Bacteriol. 1985;164(2):918-21.
2. Fabret C, Dusko Ehrlich S, Noirot P. A new mutation delivery system for genome-scale approaches in *Bacillus subtilis*. Molecular microbiology. 2002;46(1):25-36.
3. Pritchett MA, Zhang JK, Metcalf WW. Development of a Markerless Genetic Exchange Method for *Methanosarcina acetivorans* C2A and Its Use in Construction of New Genetic Tools for Methanogenic Archaea. Applied and environmental microbiology. 2004;70(3):1425.
4. Naik SN, Goud VV, Rout PK, Dalai AK. Production of first and second generation biofuels: A comprehensive review. Renewable and Sustainable Energy Reviews. 2010;14(2):578-97.
5. Czajka J, Wang Q, Wang Y, Tang YJ. Synthetic biology for manufacturing chemicals: constraints drive the use of non-conventional microbial platforms. Applied microbiology and biotechnology. 2017;101(20):7427-34.
6. Argyros DA, Tripathi SA, Barrett TF, Rogers SR, Feinberg LF, Olson DG, et al. High Ethanol Titers from Cellulose by Using Metabolically Engineered Thermophilic, Anaerobic Microbes. Applied and environmental microbiology. 2011;77(23):8288-94.
7. Gowen CM, Fong SS. Applications of systems biology towards microbial fuel production. Trends in Microbiology. 2011;19(10):516-24.
8. Husson RN, James BE, Young RA. Gene replacement and expression of foreign DNA in mycobacteria. J Bacteriol. 1990;172(2):519-24.
9. Ralph Sims MT. From 1st-to 2nd-generation biofuel technologies: An overview of current industry and RD&D activities. IEA Bioenergy. 2008.
10. Olson DG, McBride JE, Shaw AJ, Lynd LR. Recent progress in consolidated bioprocessing. Curr Opin Biotechnol. 2012;23(3):396-405.
11. Knoot CJ, Ungerer J, Wangikar PP, Pakrasi HB. Cyanobacteria: Promising biocatalysts for sustainable chemical production. The Journal of biological chemistry. 2018;293(14):5044-52.
12. Karthikeyan R, Singh R, Bose A. Microbial electron uptake in microbial electrosynthesis: a mini-review. Journal of industrial microbiology & biotechnology. 2019;46(9-10):1419-26.
13. Kast P. pKSS--a second-generation general purpose cloning vector for efficient positive selection of recombinant clones. Gene. 1994;138(1-2):109-14.
14. van der Geize R, de Jong W, Hessels GI, Grommen AWF, Jacobs AAC, Dijkhuizen L. A novel method to generate unmarked gene deletions in the intracellular pathogen *Rhodococcus equi* using 5-fluorocytosine conditional lethality. Nucleic acids research. 2008;36(22):e151-e.
15. Devarapalli M, Atiyeh HK. A review of conversion processes for bioethanol production with a focus on syngas fermentation. Biofuel Research Journal. 2015;2(3):268-80.
16. Zhang X-Z, Yan X, Cui Z-L, Hong Q, Li S-P. mazF , a novel counter-selectable marker for unmarked chromosomal manipulation in Bacillus subtilis. Nucleic Acids Research. 2006;34(9):e71-e.
17. Pathak VM, Navneet. Review on the current status of polymer degradation: a microbial approach. Bioresources and Bioprocessing. 2017;4(1):15.
18. Ueki T, Inouye S, Inouye M. Positive-negative KG cassettes for construction of multi-gene deletions using a single drug marker. Gene. 1996;183(1-2):153-7.
19. Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. Non-model organisms, a species endangered by proteogenomics. Journal of Proteomics. 2014;105:5-18.

20. Eldem V, Zararsiz G, Taşçi T, Duru YBIP, Bakir Y, Erkan M. Transcriptome Analysis for Non-Model Organism: Current Status and Best-Practices. Applications of RNA-Seq and Omics Strategies - From Microorganisms to Human Health2017.
21. Ellegren H. Genome sequencing and population genomics in non-model organisms. Trends in ecology & evolution. 2014;29(1):51-63.
22. Zhang C, She Q, Bi H, Whitaker RJ. The *apt*/6-methylpurine counterselection system and its applications in genetic studies of the hyperthermophilic archaeon *Sulfolobus islandicus*. Appl Environ Microbiol. 2016;82(10):3070-81.
23. Dean D. A plasmid cloning vector for the direct selection of strains carrying recombinant plasmids. Gene. 1981;15(1):99-102.
24. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets. Nature reviews Molecular cell biology. 2006;7(3):198-210.
25. Fondi M, Liò P. Multi -omics and metabolic modelling pipelines: Challenges and tools for systems microbiology. Microbiological Research. 2015;171:52-64.
26. Maloy SR, Nunn WD. Selection for loss of tetracycline resistance by *Escherichia coli*. Journal of bacteriology. 1981;145(2):1110-1.
27. Yan Q, Fong SS. Challenges and Advances for Genetic Engineering of Non-model Bacteria and Uses in Consolidated Bioprocessing. Frontiers in Microbiology. 2017;8(2060).
28. Stacey KA, Simson E. IMPROVED METHOD FOR THE ISOLATION OF THYMINE-REQUIRING MUTANTS OF *ESCHERICHIA COLI*. Journal of bacteriology. 1965;90(2):554-5.
29. Bernard P, Gabant P, Bahassi EM, Couturier M. Positive-selection vectors using the F plasmid *ccdB* killer gene. Gene. 1994;148(1):71-4.
30. Calvin N, Hanawalt P. High-efficiency transformation of bacterial cells by electroporation. Journal of bacteriology. 1988;170(6):2796-801.
31. Dower WJ, Miller JF, Ragsdale CW. High efficiency transformation of *E. coli* by high voltage electroporation. Nucleic acids research. 1988;16(13):6127-45.
32. Trieu-Cuot P, Carlier C, Martin P, Courvalin P. Plasmid transfer by conjugation from *Escherichia coli* to Gram-positive bacteria. FEMS Microbiology Letters. 1987;48(1-2):289-94.
33. Mazodier P, Petter R, Thompson C. Intergeneric conjugation between *Escherichia coli* and *Streptomyces* species. Journal of Bacteriology. 1989;171(6):3583-5.
34. Tominaga Y, Ohshiro T, Suzuki H. Conjugative plasmid transfer from *Escherichia coli* is a versatile approach for genetic transformation of thermophilic *Bacillus* and *Geobacillus* species. Extremophiles. 2016;20(3):375-81.
35. Solem C, Defoor E, Jensen PR, Martinussen J. Plasmid pCS1966, a New Selection/Counterselection Tool for Lactic Acid Bacterium Strain Construction Based on the *oroP* Gene, Encoding an Orotate Transporter from *Lactococcus lactis*. Applied and environmental microbiology. 2008;74(15):4772.
36. Shaw AJ, Covalla SF, Hogsett DA, Herring CD. Marker removal system for *Thermoanaerobacterium saccharolyticum* and development of a markerless ethanologen. Applied and environmental microbiology. 2011;77(7):2534-6.
37. Blokesch M. Natural competence for transformation. Current Biology. 2016;26(21):R1126-R30.
38. Dubnau D. Genetic competence in *Bacillus subtilis*. Microbiol Rev. 1991;55(3):395-424.
39. Vaneechoutte M, Young DM, Ornston LN, De Baere T, Nemec A, Van Der Reijden T, et al. Naturally transformable *Acinetobacter* sp. strain ADP1 belongs to the newly described species *Acinetobacter baylyi*. Appl Environ Microbiol. 2006;72(1):932-6.
40. Pyne ME, Bruder M, Moo-Young M, Chung DA, Chou CP. Technical guide for genetic advancement of underdeveloped and intractable *Clostridium*. Biotechnology advances. 2014;32(3):623-

41. Thomason LC, Costantino N, Court DL. *E. coli* genome manipulation by P1 transduction. Current protocols in molecular biology. 2007;Chapter 1:Unit 1.17.
42. Mandel M, Higa A. Calcium-dependent bacteriophage DNA infection. Journal of molecular biology. 1970;53(1):159-62.
43. Song Y, Hahn T, Thompson IP, Mason TJ, Preston GM, Li G, et al. Ultrasound-mediated DNA transfer for bacteria. Nucleic acids research. 2007;35(19):e129-e.
44. Shark KB, Smith FD, Harpending PR, Rasmussen JL, Sanford JC. Biolistic transformation of a procaryote, *Bacillus megaterium*. Applied and environmental microbiology. 1991;57(2):480-5.
45. Kawata Y, Yano S, Kojima H. *Escherichia coli* can be transformed by a liposome-mediated lipofection method. Bioscience, biotechnology, and biochemistry. 2003;67(5):1179-81.
46. Katsumata R, Ozaki A, Oka T, Furuya A. Protoplast transformation of glutamate-producing bacteria with plasmid DNA. Journal of bacteriology. 1984;159(1):306-11.
47. Wilharm G, Lepka D, Faber F, Hofmann J, Kerrinnes T, Skiebe E. A simple and rapid method of bacterial transformation. J Microbiol Methods. 2010;80(2):215-6.
48. Arber W, Linn S. DNA modification and restriction. Annu Rev Biochem. 1969;38:467-500.
49. Vasu K, Nagaraja V. Diverse Functions of Restriction-Modification Systems in Addition to Cellular Defense. Microbiology and Molecular Biology Reviews. 2013;77(1):53.
50. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. Science (New York, NY). 2010;327(5962):167-70.
51. Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. Science (New York, NY). 2018;359(6379).
52. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. Nature Reviews Microbiology. 2020;18(2):113-9.
53. Mermelstein LD, Papoutsakis ET. In vivo methylation in *Escherichia coli* by the *Bacillus subtilis* phage phi 3T I methyltransferase to protect plasmids from restriction upon transformation of *Clostridium acetobutylicum* ATCC 824. Applied and environmental microbiology. 1993;59(4):1077-81.
54. Riley LA, Ji L, Schmitz RJ, Westpheling J, Guss AM. Rational development of transformation in *Clostridium thermocellum* ATCC 27405 via complete methylome analysis and evasion of native restriction-modification systems. J Ind Microbiol Biotechnol. 2019.
55. Chung D, Farkas J, Westpheling J. Overcoming restriction as a barrier to DNA transformation in *Caldicellulosiruptor* species results in efficient marker replacement. Biotechnol Biofuels. 2013;6(1):82.
56. Mermelstein LD, Welker NE, Bennett GN, Papoutsakis ET. Expression of Cloned Homologous Fermentative Genes in *Clostridium acetobutylicum* ATCC 824. Bio/Technology. 1992;10(2):190-5.
57. Chung D, Farkas J, Huddleston JR, Olivar E, Westpheling J. Methylation by a Unique α-class N4-Cytosine Methyltransferase Is Required for DNA Transformation of *Caldicellulosiruptor bescii* DSM6725. PLOS ONE. 2012;7(8):e43844.
58. Zhang G, Wang W, Deng A, Sun Z, Zhang Y, Liang Y, et al. A mimicking-of-DNA-methylation-patterns pipeline for overcoming the restriction barrier of bacteria. PLoS Genet. 2012;8(9):e1002987-e.
59. Monk IR, Shah IM, Xu M, Tan M-W, Foster TJ. Transforming the Untransformable: Application of Direct Transformation To Manipulate Genetically Staph*ylococcus aureus* and *Staphylococcus epidermidis*. mBio. 2012;3(2):e00277-11.
60. Ando T, Xu Q, Torres M, Kusugami K, Israel DA, Blaser MJ. Restriction–modification system differences in *Helicobacter pylori* are a barrier to interstrain plasmid transfer. Molecular microbiology. 2000;37(5):1052-65.

61. Chen Q, Fischer JR, Benoit VM, Dufour NP, Youderian P, Leong JM. In vitro CpG methylation increases the transformation efficiency of *Borrelia burgdorferi* strains harboring the endogenous linear plasmid lp56. Journal of bacteriology. 2008;190(24):7885-91.
62. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-89.
63. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010;7(6):461-5.
64. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nature protocols. 2015;10(3):475-83.
65. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nat Methods. 2017;14(4):411-3.
66. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 2015;43(Database issue):D298-9.
67. Mohapatra SS, Biondi EG. DNA Methylation in Prokaryotes: Regulation and Function. In: Krell T, editor. Cellular Ecophysiology of Microbe. Cham: Springer International Publishing; 2017. p. 1-21.
68. Rachkevych N, Sybirna K, Boyko S, Boretsky Y, Sibirny A. Improving the efficiency of plasmid transformation in *Shewanella oneidensis* MR-1 by removing ClaI restriction site. J Microbiol Methods. 2014;99:35-7.
69. Kim JY, Wang Y, Park MS, Ji GE. Improvement of transformation efficiency through in vitro methylation and SacII site mutation of plasmid vector in *Bifidobacterium longum* MG1. Journal of microbiology and biotechnology. 2010;20(6):1022-6.
70. Johnston CD, Cotton SL, Rittling SR, Starr JR, Borisy GG, Dewhirst FE, et al. Systematic evasion of the restriction-modification barrier in bacteria. Proceedings of the National Academy of Sciences. 2019;116(23):11454-9.
71. Yasui K, Kano Y, Tanaka K, Watanabe K, Shimizu-Kadota M, Yoshikawa H, et al. Improvement of bacterial transformation efficiency using plasmid artificial modification. Nucleic Acids Res. 2009;37(1):e3.
72. Jennert KCB, Tardif C, Young DI, Young M. Gene transfer to *Clostridium cellulolyticum* ATCC 35319. Microbiology. 2000;146(12):3071-80.
73. Donahue JP, Israel DA, Peek Jr RM, Blaser MJ, Miller GG. Overcoming the restriction barrier to plasmid transformation of *Helicobacter pylori*. Molecular Microbiology. 2000;37(5):1066-74.
74. Purdy D, O'Keeffe TA, Elmore M, Herbert M, McLeod A, Bokori-Brown M, et al. Conjugative transfer of clostridial shuttle vectors from *Escherichia coli* to *Clostridium difficile* through circumvention of the restriction barrier. Mol Microbiol. 2002;46(2):439-52.
75. Mell JC, Redfield RJ. Natural Competence and the Evolution of DNA Uptake Specificity. Journal of Bacteriology. 2014;196(8):1471.
76. Guss AM, Olson DG, Caiazza NC, Lynd LR. Dcm methylation is detrimental to plasmid transformation in *Clostridium thermocellum*. Biotechnology for biofuels. 2012;5(1):30.
77. Kolek J, Sedlar K, Provaznik I, Patakova P. Dam and Dcm methylations prevent gene transfer into *Clostridium pasteurianum* NRRL B-598: development of methods for electrotransformation, conjugation, and sonoporation. Biotechnology for Biofuels. 2016;9(1):14.
78. González-Cerón G, Miranda-Olivares OJ, Servín-González L. Characterization of the methyl-specific restriction system of *Streptomyces coelicolor* A3(2) and of the role played by laterally acquired nucleases. FEMS Microbiology Letters. 2009;301(1):35-43.
79. Baker TA, Wickner SH. Genetics and enzymology of DNA replication in *Escherichia coli*. Annual review of genetics. 1992;26:447-77.
80. Jain A, Srivastava P. Broad host range plasmids. FEMS Microbiology Letters. 2013;348(2):87-96.

81. Bi C, Su P, Müller J, Yeh Y-C, Chhabra SR, Beller HR, et al. Development of a broad-host synthetic biology toolbox for *Ralstonia eutropha* and its application to engineering hydrocarbon biofuel production. Microbial Cell Factories. 2013;12(1):107.
82. Chung D, Cha M, Farkas J, Westpheling J. Construction of a stable replicating shuttle vector for *Caldicellulosiruptor* species: use for extending genetic methodologies to other members of this genus. PloS one. 2013;8(5):e62881-e.
83. Liu D, Pakrasi HB. Exploring native genetic elements as plug-in tools for synthetic biology in the cyanobacterium *Synechocystis* sp. PCC 6803. Microbial Cell Factories. 2018;17(1):48.
84. Lee S-W, Browning GF, Markham PF. Development of a replicable oriC plasmid for Mycoplasma gallisepticum and *Mycoplasma imitans*, and gene disruption through homologous recombination in *M. gallisepticum*. Microbiology. 2008;154(9):2571-80.
85. Zakrzewska-Czerwińska J, Majka J, Schrempf H. Minimal requirements of the *Streptomyces lividans* 66 oriC region and its transcriptional and translational activities. Journal of Bacteriology. 1995;177(16):4765.
86. Peteranderl R, Shotts EB, Jr., Wiegel J. Stability of antibiotics under growth conditions for thermophilic anaerobes. Applied and environmental microbiology. 1990;56(6):1981-3.
87. Hoseki J, Yano T, Koyama Y, Kuramitsu S, Kagamiyama H. Directed evolution of thermostable kanamycin-resistance gene: a convenient selection marker for *Thermus thermophilus*. Journal of biochemistry. 1999;126(5):951-6.
88. Vidal L, Pinsach J, Striedner G, Caminal G, Ferrer P. Development of an antibiotic-free plasmid selection system based on glycine auxotrophy for recombinant protein overproduction in *Escherichia coli*. Journal of biotechnology. 2008;134(1-2):127-36.
89. Schneider JC, Jenings AF, Mun DM, McGovern PM, Chew LC. Auxotrophic markers *pyrF* and *proC* can replace antibiotic markers on protein production plasmids in high-cell-density *Pseudomonas fluorescens* fermentation. Biotechnology progress. 2005;21(2):343-8.
90. Heap JT, Pennington OJ, Cartman ST, Minton NP. A modular system for *Clostridium* shuttle plasmids. J Microbiol Methods. 2009;78(1):79-85.
91. Wicke N, Radford D, Verrone V, Wipat A, French CE. BacilloFlex: A modular DNA assembly toolkit for *Bacillus subtilis* synthetic biology. bioRxiv. 2017:185108.
92. Anthony WE, Carr RR, DeLorenzo DM, Campbell TP, Shang Z, Foston M, et al. Development of *Rhodococcus opacus* as a chassis for lignin valorization and bioproduction of high-value compounds. Biotechnology for Biofuels. 2019;12(1):192.
93. Song CW, Lee J, Lee SY. Genome engineering and gene expression control for bacterial strain development. Biotechnology Journal. 2015;10(1):56-68.
94. Zhang Y, Buchholz F, Muyrers JPP, Stewart AF. A new logic for DNA engineering using recombination in *Escherichia coli*. Nature Genetics. 1998;20(2):123-8.
95. Reyrat JM, Pelicic V, Gicquel B, Rappuoli R. Counterselectable markers: untapped tools for bacterial genetics and pathogenesis. Infect Immun. 1998;66(9):4011-7.
96. Van Zyl WF, Dicks LMT, Deane SM. Development of a novel selection/counter-selection system for chromosomal gene integrations and deletions in lactic acid bacteria. BMC Molecular Biology. 2019;20(1):10.
97. Zhang X-Z, Yan X, Cui Z-L, Hong Q, Li S-P. *mazF*, a novel counter-selectable marker for unmarked chromosomal manipulation in *Bacillus subtilis*. Nucleic acids research. 2006;34(9):e71-e.
98. Dubeau M-P, Ghinet MG, Jacques P-E, Clermont N, Beaulieu C, Brzezinski R. Cytosine deaminase as a negative selection marker for gene disruption and replacement in the genus *Streptomyces* and other actinobacteria. Applied and environmental microbiology. 2009;75(4):1211-4.

99. Boeke JD, La Croute F, Fink GR. A positive selection for mutants lacking orotidine-5'-phosphate decarboxylase activity in yeast: 5-fluoro-orotic acid resistance. Molecular and General Genetics MGG. 1984;197(2):345-6.
100. Tripathi SA, Olson DG, Argyros DA, Miller BB, Barrett TF, Murphy DM, et al. Development of *pyrF*-based genetic system for targeted gene deletion in *Clostridium thermocellum* and creation of a *pta* mutant. Appl Environ Microbiol. 2010;76(19):6591-9.
101. Galvao TC, de Lorenzo V. Adaptation of the yeast URA3 selection system to gram-negative bacteria and generation of a Δ*betCDE Pseudomonas putida* strain. Applied and environmental microbiology. 2005;71(2):883-92.
102. Khasanov FK, Zvingila DJ, Zainullin AA, Prozorov AA, Bashkirov VI. Homologous recombination between plasmid and chromosomal DNA in *Bacillus subtilis* requires approximately 70 bp of homology. Molecular & general genetics : MGG. 1992;234(3):494-7.
103. Marx CJ, Lidstrom ME. Broad-Host-Range *cre-lox* System for Antibiotic Marker Recycling in Gram-Negative Bacteria. BioTechniques. 2002;33(5):1062-7.
104. Lee S-H, Kim HJ, Shin Y-A, Kim KH, Lee SJ. Single crossover-mediated markerless genome engineering in *Clostridium acetobutylicum*. Journal of microbiology and biotechnology. 2016;26(4):725-9.
105. Tan X, Liang F, Cai K, Lu X. Application of the FLP/FRT recombination system in cyanobacteria for construction of markerless mutants. Applied microbiology and biotechnology. 2013;97(14):6373-82.
106. Olson DG, Lynd LR. Transformation of *Clostridium thermocellum* by electroporation. Methods in enzymology. 2012;510:317-30.
107. Tian L, Papanek B, Olson DG, Rydzak T, Holwerda EK, Zheng T, et al. Simultaneous achievement of high ethanol yield and titer in *Clostridium thermocellum*. Biotechnology for Biofuels. 2016;9(1):116.
108. Tian L, Conway PM, Cervenka ND, Cui J, Maloney M, Olson DG, et al. Metabolic engineering of *Clostridium thermocellum* for n-butanol production from cellulose. Biotechnology for Biofuels. 2019;12(1):186.
109. Basen M, Geiger I, Henke L, Muller V. A Genetic System for the Thermophilic Acetogenic Bacterium *Thermoanaerobacter kivui*. Applied and environmental microbiology. 2018;84(3).
110. Li Z, Xiong B, Liu L, Li S, Xin X, Li Z, et al. Development of an autotrophic fermentation technique for the production of fatty acids using an engineered *Ralstonia eutropha* cell factory. Journal of Industrial Microbiology & Biotechnology. 2019;46(6):783-90.
111. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. Proceedings of the National Academy of Sciences of the United States of America. 2000;97(12):6640-5.
112. Ellis HM, Yu D, DiTizio T, Court DL. High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. Proceedings of the National Academy of Sciences. 2001;98(12):6742-6.
113. Luo X, Yang Y, Ling W, Zhuang H, Li Q, Shang G. *Pseudomonas putida* KT2440 markerless gene deletion using a combination of λ Red recombineering and Cre/*loxP* site-specific recombination. FEMS microbiology letters. 2016;363(4).
114. Sun Z, Deng A, Hu T, Wu J, Sun Q, Bai H, et al. A high efficiency recombineering system with PCR-based ssDNA in *Bacillus subtilis* mediated by the native phage recombinase GP35. Applied microbiology and biotechnology. 2015;99:5151-62.
115. van Kessel JC, Hatfull GF. Recombineering in *Mycobacterium tuberculosis*. Nat Methods. 2007;4(2):147-52.

116. Dong H, Tao W, Gong F, Li Y, Zhang Y. A functional *recT* gene for recombineering of *Clostridium*. Journal of biotechnology. 2014;173:65-7.
117. van Pijkeren JP, Britton RA. High efficiency recombineering in lactic acid bacteria. Nucleic Acids Res. 2012;40(10):e76.
118. Walker JE, Lanahan AA, Zheng T, Toruno C, Lynd LR, Cameron JC, et al. Development of both type I–B and type II CRISPR/Cas genome editing systems in the cellulolytic bacterium *Clostridium thermocellum*. Metabolic Engineering Communications. 2020;10:e00116.
119. Corts AD, Thomason LC, Gill RT, Gralnick JA. A new recombineering system for precise genome-editing in *Shewanella oneidensis* strain MR-1 using single-stranded oligonucleotides. Scientific Reports. 2019;9(1):39.
120. van Kessel JC, Marinelli LJ, Hatfull GF. Recombineering mycobacteria and their phages. Nature reviews Microbiology. 2008;6(11):851-7.
121. Hensel M, Shea JE, Gleeson C, Jones MD, Dalton E, Holden DW. Simultaneous identification of bacterial virulence genes by negative selection. Science (New York, NY). 1995;269(5222):400-3.
122. Lampe DJ, Churchill ME, Robertson HM. A purified mariner transposase is sufficient to mediate transposition in vitro. EMBO J. 1996;15(19):5470-9.
123. Simon R, Priefer U, Pühler A. A broad host range mobilization system for in vivo genetic engineering: transposon mutagenesis in gram negative bacteria. Bio/technology. 1983;1(9):784.
124. van Opijnen T, Bodi KL, Camilli A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods. 2009;6(10):767-72.
125. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, et al. Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. mBio. 2015;6(3):e00306-15.
126. Calero P, Jensen SI, Bojanovič K, Lennen RM, Koza A, Nielsen AT. Genome-wide identification of tolerance mechanisms toward p-coumaric acid in *Pseudomonas putida*. Biotechnology and bioengineering. 2018;115(3):762-74.
127. Curtis PD. Essential genes predicted in the genome of *Rubrivivax gelatinosus*. Journal of bacteriology. 2016;198(16):2244-50.
128. Yomano LP, York SW, Zhou S, Shanmugam KT, Ingram LO. Re-engineering *Escherichia coli* for ethanol production. Biotechnology Letters. 2008;30(12):2097-103.
129. Philipps G, de Vries S, Jennewein S. Development of a metabolic pathway transfer and genomic integration system for the syngas-fermenting bacterium *Clostridium ljungdahlii*. Biotechnology for Biofuels. 2019;12(1):112.
130. Inaba Y, Banerjee I, Kernan T, Banta S. Transposase-mediated chromosomal integration of exogenous genes in *Acidithiobacillus ferrooxidans*. Appl Environ Microbiol. 2018;84(21):e01381-18.
131. Wang G, Zhao Z, Ke J, Engel Y, Shi Y-M, Robinson D, et al. CRAGE enables rapid activation of biosynthetic gene clusters in undomesticated bacteria. Nature Microbiology. 2019;4(12):2498-510.
132. Choi K-H, Gaynor JB, White KG, Lopez C, Bosio CM, Karkhoff-Schweizer RR, et al. A Tn7-based broad-range bacterial cloning and expression system. Nat Methods. 2005;2(6):443-8.
133. Groth AC, Calos MP. Phage integrases: biology and applications. Journal of molecular biology. 2004;335(3):667-78.
134. Chou Y-C, Linger J, Yang S, Zhang M. Genetic engineering and improvement of a *Zymomonas mobilis* for arabinose utilization and its performance on pretreated corn stover hydrolyzate. Journal of Biotechnology & Biomaterials. 2015;5(NREL/JA-5100-64639).

135. Lambert JM, Bongers RS, Kleerebezem M. Cre-*lox*-Based System for Multiple Gene Deletions and Selectable-Marker Removal in *Lactobacillus plantarum*. Applied and environmental microbiology. 2007;73(4):1126-35.
136. Langer SJ, Ghafoori AP, Byrd M, Leinwand L. A genetic screen identifies novel non-compatible *loxP* sites. Nucleic Acids Res. 2002;30(14):3067-77.
137. Brown WRA, Lee NCO, Xu Z, Smith MCM. Serine recombinases as tools for genome engineering. Methods. 2011;53(4):372-9.
138. Elmore JR, Furches A, Wolff GN, Gorday K, Guss AM. Development of a high efficiency integration system and promoter library for rapid modification of *Pseudomonas putida* KT2440. Metabolic Engineering Communications. 2017;5:1-8.
139. Combes P, Till R, Bee S, Smith MC. The *Streptomyces* genome contains multiple pseudo-*attB* sites for the (phi)C31-encoded site-specific recombination system. J Bacteriol. 2002;184(20):5746-52.
140. Guss AM, Rother M, Zhang JK, Kulkarni G, Metcalf WW. New methods for tightly regulated gene expression and highly efficient chromosomal integration of cloned genes for *Methanosarcina* species. Archaea (Vancouver, BC). 2008;2(3):193-203.
141. Huang H, Chai CS, Yang S, Jiang WH, Gu Y. Phage serine integrase-mediated genome engineering for efficient expression of chemical biosynthetic pathway in gas-fermenting *Clostridium ljungdahlii*. Metabolic Engineering. 2019;52:293-302.
142. Ko B, D'Alessandro J, Douangkeomany L, Stumpf S, deButts A, Blodgett J. Construction of a new integrating vector from actinophage φOZJ and its use in multiplex *Streptomyces transformation*. Journal of Industrial Microbiology & Biotechnology. 2019.
143. Bilyk B, Luzhetskyy A. Unusual site-specific DNA integration into the highly active pseudo-*attB* of the *Streptomyces albus* J1074 genome. Applied microbiology and biotechnology. 2014;98(11):5095-104.
144. Zhang JJ, Moore BS, Tang X. Engineering *Salinispora tropica* for heterologous expression of natural product biosynthetic gene clusters. Applied microbiology and biotechnology. 2018;102(19):8437-46.
145. Jiang Y, Qian F, Yang J, Liu Y, Dong F, Xu C, et al. CRISPR-Cpf1 assisted genome editing of *Corynebacterium glutamicum*. Nature communications. 2017;8:15179.
146. Bierman M, Logan R, O'Brien K, Seno ET, Nagaraja Rao R, Schoner BE. Plasmid cloning vectors for the conjugal transfer of DNA from *Escherichia coli* to *Streptomyces* spp. Gene. 1992;116(1):43-9.
147. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. CRISPR provides acquired resistance against viruses in prokaryotes. Science (New York, NY). 2007;315(5819):1709-12.
148. Javed MR, Noman M, Shahid M, Ahmed T, Khurshid M, Rashid MH, et al. Current situation of biofuel production and its enhancement by CRISPR/Cas9-mediated genome engineering of microbial cells. Microbiological Research. 2019;219:1-11.
149. Vento JM, Crook N, Beisel CL. Barriers to genome editing with CRISPR in bacteria. J Ind Microbiol Biotechnol. 2019;46(9-10):1327-41.
150. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011;471(7340):602-7.
151. Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat Biotechnol. 2013;31(3):233-9.
152. Pyne ME, Moo-Young M, Chung DA, Chou CP. Coupling the CRISPR/Cas9 System with Lambda Red Recombineering Enables Simplified Chromosomal Gene Replacement in *Escherichia coli*. Applied and environmental microbiology. 2015;81(15):5103-14.

153. Cho S, Choe D, Lee E, Kim SC, Palsson B, Cho BK. High-Level dCas9 Expression Induces Abnormal Cell Morphology in *Escherichia coli*. ACS synthetic biology. 2018;7(4):1085-94.
154. Wasels F, Jean-Marie J, Collas F, Lopez-Contreras AM, Lopes Ferreira N. A two-plasmid inducible CRISPR/Cas9 genome editing tool for *Clostridium acetobutylicum*. J Microbiol Methods. 2017;140:5-11.
155. Li K, Cai D, Wang Z, He Z, Chen S. Development of an efficient genome editing tool in *Bacillus licheniformis* using CRISPR-Cas9 nickase. Appl Environ Microbiol. 2018;84(6):e02608-17.
156. Ran FA, Hsu Patrick D, Lin C-Y, Gootenberg Jonathan S, Konermann S, Trevino AE, et al. Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. Cell. 2013;154(6):1380-9.
157. Zetsche B, Gootenberg Jonathan S, Abudayyeh Omar O, Slaymaker Ian M, Makarova Kira S, Essletzbichler P, et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell. 2015;163(3):759-71.
158. Ungerer J, Pakrasi HB. Cpf1 is a versatile tool for CRISPR genome editing across diverse species of cyanobacteria. Scientific reports. 2016;6(1):1-9.
159. Pyne ME, Bruder MR, Moo-Young M, Chung DA, Chou CP. Harnessing heterologous and endogenous CRISPR-Cas machineries for efficient markerless genome editing in *Clostridium*. Scientific reports. 2016;6(1):1-15.
160. Hidalgo-Cantabrana C, Goh YJ, Pan M, Sanozky-Dawes R, Barrangou R. Genome editing using the endogenous type I CRISPR-Cas system in *Lactobacillus crispatus*. Proceedings of the National Academy of Sciences. 2019;116(32):15774-83.
161. Mougiakos I, Bosma EF, de Vos WM, van Kranenburg R, van der Oost J. Next Generation Prokaryotic Engineering: The CRISPR-Cas Toolkit. Trends in Biotechnology. 2016;34(7):575-87.
162. Mougiakos I, Bosma EF, Ganguly J, van der Oost J, van Kranenburg R. Hijacking CRISPR-Cas for high-throughput bacterial metabolic engineering: advances and prospects. Current Opinion in Biotechnology. 2018;50:146-57.
163. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nature protocols. 2013;8(11):2180-96.
164. Yao L, Cengic I, Anfelt J, Hudson EP. Multiple gene repression in cyanobacteria using CRISPRi. ACS synthetic biology. 2016;5(3):207-12.
165. Cleto S, Jensen JVK, Wendisch VF, Lu TK. *Corynebacterium glutamicum* Metabolic Engineering with CRISPR Interference (CRISPRi). ACS synthetic biology. 2016;5(5):375-85.
166. Huang C-H, Shen CR, Li H, Sung L-Y, Wu M-Y, Hu Y-C. CRISPR interference (CRISPRi) for gene regulation and succinate production in cyanobacterium *S. elongatus* PCC 7942. Microbial Cell Factories. 2016;15(1):196.
167. Liu X, Gallay C, Kjos M, Domenech A, Slager J, van Kessel SP, et al. High-throughput CRISPRi phenotyping identifies new essential genes in *Streptococcus pneumoniae*. Molecular systems biology. 2017;13(5).
168. Tao W, Lv L, Chen GQ. Engineering *Halomonas* species TD01 for enhanced polyhydroxyalkanoates synthesis via CRISPRi. Microb Cell Fact. 2017;16(1):48.
169. Peters JM, Koo B-M, Patino R, Heussler GE, Hearne CC, Qu J, et al. Enabling genetic analysis of diverse bacteria with Mobile-CRISPRi. Nature Microbiology. 2019;4(2):244-50.
170. Bikard D, Jiang W, Samai P, Hochschild A, Zhang F, Marraffini LA. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res. 2013;41(15):7429-37.
171. Dong C, Fontana J, Patel A, Carothers JM, Zalatan JG. Synthetic CRISPR-Cas gene activators for transcriptional reprogramming in bacteria. Nature communications. 2018;9(1):2489.

172. Liu Y, Wan X, Wang B. Engineered CRISPRa enables programmable eukaryote-like gene activation in bacteria. Nature communications. 2019;10(1):3693.
173. Yang Y, Shen W, Huang J, Li R, Xiao Y, Wei H, et al. Prediction and characterization of promoters and ribosomal binding sites of *Zymomonas mobilis* in system biology era. Biotechnology for Biofuels. 2019;12(1):52.
174. Jin L-Q, Jin W-R, Ma Z-C, Shen Q, Cai X, Liu Z-Q, et al. Promoter engineering strategies for the overproduction of valuable metabolites in microbes. Applied microbiology and biotechnology. 2019;103(21):8725-36.
175. Luo Y, Zhang L, Barton KW, Zhao H. Systematic Identification of a Panel of Strong Constitutive Promoters from *Streptomyces albus*. ACS synthetic biology. 2015;4(9):1001-10.
176. Glick BR. Metabolic load and heterologous gene expression. Biotechnology advances. 1995;13(2):247-61.
177. Siegl T, Tokovenko B, Myronovskyi M, Luzhetskyy A. Design, construction and characterisation of a synthetic promoter library for fine-tuned gene expression in actinomycetes. Metabolic Engineering. 2013;19:98-106.
178. Rytter JV, Helmark S, Chen J, Lezyk MJ, Solem C, Jensen PR. Synthetic promoter libraries for *Corynebacterium glutamicum*. Applied microbiology and biotechnology. 2014;98(6):2617-23.
179. Mordaka PM, Heap JT. Stringency of Synthetic Promoter Sequences in *Clostridium* Revealed and Circumvented by Tuning Promoter Library Mutation Rates. ACS synthetic biology. 2018;7(2):672-81.
180. Yang G, Jia D, Jin L, Jiang Y, Wang Y, Jiang W, et al. Rapid Generation of Universal Synthetic Promoters for Controlled Gene Expression in Both Gas-Fermenting and Saccharolytic *Clostridium* Species. ACS synthetic biology. 2017;6(9):1672-8.
181. Englund E, Liang F, Lindberg P. Evaluation of promoters and ribosome binding sites for biotechnological applications in the unicellular cyanobacterium *Synechocystis* sp. PCC 6803. Scientific Reports. 2016;6(1):36640.
182. Zobel S, Benedetti I, Eisenbach L, de Lorenzo V, Wierckx N, Blank LM. Tn7-Based Device for Calibrated Heterologous Gene Expression in *Pseudomonas putida*. ACS synthetic biology. 2015;4(12):1341-51.
183. Ross W, Gosink KK, Salomon J, Igarashi K, Zou C, Ishihama A, et al. A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. Science (New York, NY). 1993;262(5138):1407.
184. Presnell KV, Flexer-Harrison M, Alper HS. Design and synthesis of synthetic UP elements for modulation of gene expression in *Escherichia coli*. Synthetic and Systems Biotechnology. 2019;4(2):99-106.
185. Ross W, Aiyar SE, Salomon J, Gourse RL. *Escherichia coli* promoters with UP elements of different strengths: modular structure of bacterial promoters. J Bacteriol. 1998;180(20):5375-83.
186. Phan TTP, Nguyen HD, Schumann W. Development of a strong intracellular expression system for *Bacillus subtilis* by optimizing promoter elements. Journal of biotechnology. 2012;157(1):167-72.
187. Zhang B, Zhou N, Liu Y-M, Liu C, Lou C-B, Jiang C-Y, et al. Ribosome binding site libraries and pathway modules for shikimic acid synthesis with *Corynebacterium glutamicum*. Microbial cell factories. 2015;14:71-.
188. Sun T, Miao L, Li Q, Dai G, Lu F, Liu T, et al. Production of lycopene by metabolically-engineered *Escherichia coli*. Biotechnology Letters. 2014;36(7):1515-22.
189. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol. 2009;27(10):946-50.

190. Altenbuchner J. Editing of the *Bacillus subtilis* genome by the CRISPR-Cas9 system. Appl Environ Microbiol. 2016;82(17):5421-7.
191. Morgan S-A, Nadler DC, Yokoo R, Savage DF. Biofuel metabolic engineering with biosensors. Curr Opin Chem Biol. 2016;35:150-8.
192. Lo T-M, Chng SH, Teo WS, Cho H-S, Chang MW. A Two-Layer Gene Circuit for Decoupling Cell Growth from Metabolite Production. Cell Systems. 2016;3(2):133-43.
193. Chubukov V, Desmarais JJ, Wang G, Chan LJG, Baidoo EE, Petzold CJ, et al. Engineering glucose metabolism of *Escherichia coli* under nitrogen starvation. NPJ systems biology and applications. 2017;3(1):1-7.
194. Burg JM, Cooper CB, Ye Z, Reed BR, Moreb EA, Lynch MD. Large-scale bioprocess competitiveness: the potential of dynamic metabolic control in two-stage fermentations. Current Opinion in Chemical Engineering. 2016;14:121-36.
195. Zhang J, Liu Y-J, Cui G-Z, Cui Q. A novel arabinose-inducible genetic operation system developed for *Clostridium cellulolyticum*. Biotechnology for Biofuels. 2015;8(1):36.
196. Rodríguez-García A, Combes P, Pérez-Redondo R, Smith MCA, Smith MCM. Natural and synthetic tetracycline-inducible promoters for use in the antibiotic-producing bacteria *Streptomyces*. Nucleic Acids Research. 2005;33(9):e87-e.
197. Girbal L, Mortier-Barrière I, Raynaud F, Rouanet C, Croux C, Soucaille P. Development of a sensitive gene expression reporter system and an inducible promoter-repressor system for *Clostridium acetobutylicum*. Appl Environ Microbiol. 2003;69(8):4985-8.
198. Ruegg TL, Pereira JH, Chen JC, DeGiovanni A, Novichkov P, Mutalik VK, et al. Jungle Express is a versatile repressor system for tight transcriptional control. Nature communications. 2018;9(1):1-13.
199. Mearls EB, Olson DG, Herring CD, Lynd LR. Development of a regulatable plasmid-based gene expression system for *Clostridium thermocellum*. Applied microbiology and biotechnology. 2015;99(18):7589-99.
200. Nshogozabahizi JC, Aubrey KL, Ross JA, Thakor N. Applications and limitations of regulatory RNA elements in synthetic biology and biotechnology. Journal of Applied Microbiology. 2019;127(4):968-84.
201. Ma AT, Schmidt CM, Golden JW. Regulation of Gene Expression in Diverse Cyanobacterial Species by Using Theophylline-Responsive Riboswitches. Applied and environmental microbiology. 2014;80(21):6704-13.
202. Zhou LB, Zeng AP. Exploring lysine riboswitch for metabolic flux control and improvement of L-lysine synthesis in *Corynebacterium glutamicum*. ACS synthetic biology. 2015;4(6):729-34.
203. Jang S, Jang S, Xiu Y, Kang TJ, Lee S-H, Koffas MAG, et al. Development of Artificial Riboswitches for Monitoring of Naringenin In Vivo. ACS synthetic biology. 2017;6(11):2077-85.
204. Marcano-Velazquez JG, Lo J, Nag A, Maness P-C, Chou KJ. Developing Riboswitch-Mediated Gene Regulatory Controls in Thermophilic Bacteria. ACS synthetic biology. 2019;8(4):633-40.
205. Jin DJ, Burgess RR, Richardson JP, Gross CA. Termination efficiency at *rho*-dependent terminators depends on kinetic coupling between RNA polymerase and *rho*. Proceedings of the National Academy of Sciences of the United States of America. 1992;89(4):1453-7.
206. Platteeuw C, Simons G, De Vos W. Use of the *Escherichia coli* beta-glucuronidase (*gusA*) gene as a reporter gene for analyzing promoters in lactic acid bacteria. Appl Environ Microbiol. 1994;60(2):587-93.
207. Thorn K. Genetically encoded fluorescent tags. Mol Biol Cell. 2017;28(7):848-57.
208. Zaslaver A, Bren A, Ronen M, Itzkovitz S, Kikoin I, Shavit S, et al. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. Nat Methods. 2006;3(8):623-8.

209. Su L, Jia W, Hou C, Lei Y. Microbial biosensors: A review. Biosensors and Bioelectronics. 2011;26(5):1788-99.
210. Jha RK, Narayanan N, Pandey N, Bingen JM, Kern TL, Johnson CW, et al. Sensor-Enabled Alleviation of Product Inhibition in Chorismate Pyruvate-Lyase. ACS synthetic biology. 2019;8(4):775-86.
211. Alvarez-Gonzalez G, Dixon N. Genetically encoded biosensors for lignocellulose valorization. Biotechnology for Biofuels. 2019;12(1):246.
212. Charubin K, Bennett RK, Fast AG, Papoutsakis ET. Engineering *Clostridium* organisms as microbial cell-factories: challenges & opportunities. Metabolic Engineering. 2018;50:173-91.
213. Streett HE, Kalis KM, Papoutsakis ET. A Strongly Fluorescing Anaerobic Reporter and Protein-Tagging System for *Clostridium* Organisms Based on the Fluorescence-Activating and Absorption-Shifting Tag Protein (FAST). Applied and environmental microbiology. 2019;85(14):e00622-19.
214. Monmeyran A, Thomen P, Jonquière H, Sureau F, Li C, Plamont M-A, et al. The inducible chemical-genetic fluorescent marker FAST outperforms classical fluorescent proteins in the quantitative reporting of bacterial biofilm dynamics. Scientific Reports. 2018;8(1):10336.

Chapter II

Rational development of transformation in *Clostridium thermocellum* ATCC 27405 via complete methylome analysis and evasion of native restriction-modification systems

This chapter was originally published by Lauren A. Riley, Lexiang Ji, Robert J. Schmitz, Janet Westpheling, and Adam M. Guss:

Riley LA, Ji L, Schmitz RJ, Westpheling J, Guss AM. (2019) Rational development of transformation in *Clostridium thermocellum* ATCC 27405 via complete methylome analysis and evasion of native restriction-modification systems. *JIMB*. 46(9), 1435-1443.

Lexiang Ji and Robert J. Schmitz performed the Whole Genome Bisulfite sequencing. Lauren A. Riley performed the methyltransferase expression and *Clostridium thermocellum* transformation. Lauren A. Riley, Adam M. Guss, and Janet Westpheling wrote and revised this paper.

Abstract

A major barrier to both metabolic engineering and fundamental biological studies is the lack of genetic tools in most microorganisms. One example is *Clostridium thermocellum* ATCC 27405^T, where genetic tools are not available to help validate decades of hypotheses. A significant barrier to DNA transformation is restriction-modification systems, which defend against foreign DNA methylated differently than the host. To determine the active restriction-modification systems in this strain, we performed complete methylome analysis via Single Molecule, Real-Time sequencing to detect 6-methyladenine and 4-methylcytosine and the rarely used whole genome bisulfite sequencing to detect 5-methylcytosine. Multiple active systems were identified, and corresponding DNA methyltransferases were expressed from the *Escherichia coli* chromosome to mimic the *C. thermocellum* methylome. Plasmid methylation was experimentally validated and successfully electroporated into *C. thermocellum* ATCC 27405. This combined approach enabled genetic modification of the *C. thermocellum* type strain and acts as a blueprint for transformation of other non-model microorganisms.

Introduction

Most microbial metabolic engineering for biotechnology research is performed in model organisms because they are well studied and have a large toolbox to enable genetic modifications (1). Non-model organisms, on the other hand, are an attractive alternative as potential industrial platforms for bioconversion because they often possess complex phenotypes that are not currently feasible to engineer into model organisms, such as the ability to grow at extreme pH, extreme robustness/toxicity tolerance, or the ability to catabolize less common substrates such as syngas, methane, cellulose, or lignin (2, 3). However, work with non-model organisms is limited due to a lack tools and knowledge of the organism. A major barrier to genetic tools development is the inability to efficiently transform DNA, and to routinely allow for the use of non-model organisms, a systematic process for developing transformation is needed.

One of the major barriers to successful transformation of bacteria is native DNA restriction-modification (RM) systems, which act as a bacterial immune system to cut DNA that is methylated differently than the host (4). RM systems are classified into four types. In Types I, II, and III, a restriction enzyme typically cuts DNA that is unmethylated at a specific recognition sequence, and a corresponding DNA methyltransferase adds a methyl group to the host's DNA to protect these same sequences from restriction enzyme activity (4). Type I systems are comprised of three subunits: a DNA methyltransferase, a DNA recognition subunit, and a restriction enzyme. This type of system recognizes two motifs of 3-4 bases separated by any 5-8 bases (5), such as the EcoKI system that targets A**A**CNNNNNNG**T**GC (N is any base), and motifs are methylated at the N-6 position of one adenine per DNA strand to form N^6–methyladenine (m^6A). Type II and III systems are typically comprised of a methyltransferase and a restriction enzyme subunit. Type II systems are largely studied and commonly used as tools in molecular biology. Their recognition systems are often palindromic, and they can methylate bases to form m^6A, N^4-methylcytosine (m^4C), or 5-methylcytosine (m^5C) (6, 7). Type III

systems recognize non-palindromic motifs and typically methylate to form m^6A (8). The last group, Type IV, is only comprised of a restriction enzyme, which recognizes motifs that are methylated differentially than the host (8, 9).

To successfully evade most native RM systems, DNA needs to be methylated in the same way as the target organism, and many studies have shown that overcoming these systems is important for efficient DNA transformation (10-14). To rationally evade RM systems, the targeted motifs first need to be identified. Microbes protect their chromosomal DNA from restriction via DNA methylation; therefore, methylome analysis can be used to identify these motifs. While not all DNA methyltransferases are associated with restriction enzymes (15), methylome analysis does reveal all potentially active Type I, II, and III RM systems present in an organism. A common approach to microbial methylome analysis uses Single Molecule Real Time (SMRT) sequencing on the PacBio platform (16-19), which identifies m^6A and m^4C motifs based on kinetic delays in nucleotide incorporation when a base is methylated (20). This approach has been specifically used to help overcome restriction barriers to genetic transformation in a number of organisms (12, 14, 21). SMRT sequencing can also identify m^5C motifs but the signal is less strong, so it requires immense coverage of the genome or modification of the methylated base via Tet1 oxidation (22), approaches that are not common practice. An alternate approach to detect m^5C and m^4C is Methyl-C sequencing for whole genome bisulfite sequencing (WGBS), in which the cytosine (but not m^5C) is deaminated to uracil, followed by polymerase chain reaction (PCR) to convert the resulting uracils to thymines. The resulting PCR-amplified DNA can then be sequenced, and remaining cytosines in the sequence were previously methylated. This approach has not been routinely used in bacteria, and we have only identified a few studies that utilized WGBS for bacterial methylome analysis for characterization of RM systems (23-25).

Clostridium thermocellum is an anaerobic, thermophilic bacterium that efficiently deconstructs lignocellulosic biomass via large cell surface-associated enzyme complexes called cellulosomes. The

resulting soluble sugars are fermented into products such as organic acids and ethanol (26). There is particular interest in *C. thermocellum* due to its potential for biofuel production from lignocellulose via a process called consolidated bioprocessing (CBP), in which cellulolytic enzyme production and fermentation occur in a single reactor (27, 28). The cellulosome was first discovered in the *Clostridium thermocellum* type strain, ATCC 27405 (29), and a substantial amount of work has been done on this strain to understand carbon metabolism and the genes related to cellulosome production (30-34). All of these studies examined the wild type strain using omics tools such as transcriptomics and proteomics, but to date, this strain has not been genetically modified. This is in stark contrast to *C. thermocellum* DSM 1313, where transformation and genetics are readily available (35) and extensive metabolic engineering has been achieved (36-40). Transformation was enabled in this strain through the isolation of DNA out of the proper *E. coli* strain (*dam+dcm-*), without the need for further methylation.

The lack of genetic tools in strain ATCC 27405 has hindered both fundamental studies and development of this strain for bioengineering. Thus far, one publication has demonstrated transformation of this strain by using a custom-made electroporator (41), and additional attempts to transform *C. thermocellum* ATCC 27405 using commercially-available equipment has been unsuccessful. One other publication demonstrated a low transformation rate of 2.5±1.5 colonies per microgram of DNA with a single plasmid using large cell and DNA volumes, which makes the process hard to utilize in the future (42).

One possible reason for the difference in success of transformation between strains DSM 1313 and ATCC 27405 is differences in RM systems, which are understudied in *C. thermocellum* ATCC 27405. New England Biolabs Restriction Enzyme Database (REBASE) (43) predicts the RM systems encoded in both strains of *C. thermocellum*, where ATCC 27405 encodes eight restriction systems and DSM 1313 only encodes five. Of the eight potential RM systems in ATCC 27405, one is a putative Type I system, genes Cthe_1144-1145, though it seems to be missing the restriction enzyme subunit. There are six

putative Type II systems encoded: Cthe_1511-1513, Cthe_1638-1639, Cthe_1748-1749, Cthe_2470-2471, Cthe_2319-2320, and Cthe_1748-1749. Cthe_2740 and Cthe_1728-1729 do not have annotated associated restriction enzymes, but the other four encode putative restriction enzymes. The genome also encodes one Type III RM system, Cthe_0518-0519, with Cthe_0518 being a predicted restriction enzyme. No apparent Type IV systems are encoded in the genome. Previously, one of these RM systems was shown to be active in cell extracts of strain ATCC 27405, and extracts exhibited endonuclease activity targeting the motif G**A**TC, similar to the MboI restriction enzyme (44). Here, we report the methylome of *C. thermocellum* ATCC 27405, identify and express the four active methyltransferases, validate the expression for in vivo methylation, and successfully transform *C. thermocellum* ATCC 27405.

Materials and Methods

Strains, Plasmids, and Growth Conditions

Escherichia coli Top 10 Δ*dcm::frt* was created by deleting *dcm* in Top10 (Invitrogen) with the lambda Red recombinase system as previously described (45). *E. coli* strains were grown aerobically in LB at 37°C, and chloramphenicol was added for plasmid selection at a final concentration of 15 µg/mL. *C. thermocellum* ATCC 27405 was grown in CTFUD medium (37), which is comprised of (per L) 3 g sodium citrate tribasic dehydrate, 1.3 g ammonium sulfate, 1.43 g potassium phosphate monobasic, 1.8 g potassium phosphate dibasic trihydrate, 0.5 g cysteine-HCL, 10.5 g MOPS sodium salt, 6 g glycerol-2-phosphate disodium, 5 g cellobiose, 4.5 g yeast extract, 0.13 g calcium chloride dehydrate, 2.6 g magnesium chloride hexahydrate, 0.0001 g ferrous sulfate heptahydrate, and 0.5 ml 0.2% (w/v) resazurin. The pH is adjusted to 7.0 after addition of MOPS with 45% (w/v) potassium hydroxide. *C. thermocellum* was grown at 50 or 55°C, as indicated, in a Coy anaerobic chamber. Thiamphenicol was added to the medium when appropriate at a final concentration of 5 µg/mL.

Two plasmids, pNJ020 and pAMG216, were used for transformation of *C. thermocellum,* each containing the pNW33N origin of replication for *C. thermocellum* and the *cat* gene for thiamphenicol selection. Plasmid pNJ020 contains the p15a origin for medium copy number replication in *E. coli*. Plasmid pAMG216 is derived from pAMG205 with the yeast machinery deleted, and it has a high copy number pUC origin of replication for *E. coli* (46). Methyltransferases were codon optimized for *E. coli* and synthesized with the T5Lac promoter by GenScript Biotech Corp (New Jersey, USA), amplified by PCR, and cloned into CRIM integration vectors (47) using Gibson assembly (New England Biolabs, NEB). The native *C. thermocellum* gene Cthe_0519 was cloned into pAH55 (47). The bifunctional methyltransferase phi3TI, from *Bacillus phage phi3T,* was cloned into pAH144. Each plasmid was integrated, using the CRIM system (47), into the λ and HK022 phage *attB* sites of *E. coli* Top 10 *Δdcm::frt*, resulting in strain AG2006 (genotype: Top10 *Δdcm::frt* λ::Cthe_0519 HK022::phi3TI).

SMRT Sequencing Methylome Analysis

Genomic DNA from *C. thermocellum* ATCC 27405 was isolated using the Genomic Tip kit (Qiagen according to the manufacturer's instructions and sent to Expression Analysis (Durham, NC) for sequencing on a Pacific Biosciences (PacBio) instrument. Single Molecule Real-Time (SMRT) sequencing was performed using the PacBio RS technology with four SMRT cells. Methylated sequences were determined by Expression Analysis using the SMRT Analysis software (48).

Whole Genome Bisulfite Sequencing Analysis

MethylC-seq libraries were prepared and Illumina sequencing was performed (Genomics & Bioinformatics Core, University of Georgia) using an Illumina NextSeq 500 instrument. For data processing, raw reads were trimmed for adapters and preprocessed to remove low quality reads using cutadapt 1.9.dev1 (49). Qualified reads were aligned to the *C. thermocellum* ATCC 27405 reference genome (50). Fully unmethylated lambda DNA was used as a control to calculate the sodium bisulfite

reaction non-conversion rate of unmodified cytosines. Only cytosine sites with a minimum coverage (set as 3) were selected for subsequent analysis. Binomial test coupled with Benjamini-Hochberg correction was adopted to determine the methylation status of each cytosine. The m^5C motifs were identified as previously described in (23).

Determining functionality of the methyltransferases

Plasmids pNJ020 and pAMG216 were transformed into *E. coli* strains AG2006 and Top10 *Δdcm::frt* and grown in liquid culture with 0.1 mM IPTG to induce methyltransferase expression. Plasmid pNJ020 was isolated and digested in three separate reactions with NlaIII, TseI and HindIII, and HaeIII and HindIII (New England Biolabs, Ipswich, MA) according the manufacturer's instructions. To determine the amount of DNA methylated, digested plasmid was separated via agarose gel electrophoresis and quantified using BioRad Gel Imager software.

Transformation of C. thermocellum ATCC 27405

Three 5 mL cultures were inoculated with *C. thermocellum* ATCC 27405 and grown overnight. Three 500 mL cultures were inoculated with 1% of the overnight cultures and grown at 55°C. Cultures were harvested at an optical density (O.D.) of ~1.0, transferred to a 500 mL centrifuge bottle, placed on ice for 30 minutes, and then centrifuged at 4 °C at 5000 x g for 15 minutes. The supernatant was decanted and cells were washed with 250 mL cold electroporation buffer (10% glycerol, 250 mM sucrose), which was added without disrupting/resuspending the cell pellet. Cells were spun again, and the wash was repeated two more times. After the last wash, the electroporation buffer was completely removed, and the cell pellets were resuspended in 200 µL electroporation buffer and transferred to a microcentrifuge tube. Using fresh electrocompetent cells from each batch, 20 µL of cells were transformed with 1 µg of DNA. Square wave electroporation was performed in a 1 mM electroporation cuvette at 1200 v with a 1.5 ms pulse. Cells were resuspended in 1 mL of CTFUD liquid medium and

mixed with molten CTFUD supplemented with 1.5% agar and thiamphenicol, allowed to solidified, and incubated at 50 °C for 3-5 days, when colonies were counted.

Results

Complete methylome analysis of C. thermocellum ATCC 27405

To determine which of the eight RM systems, predicted by REBASE, are active, methylome analysis was performed. All methylated motifs in *C. thermocellum* ATCC 27405 were determined through two sequencing techniques, PacBio SMRT sequencing and MethylC-sequencing for WGBS. PacBio SMRT sequencing detected three m^6A motifs and WGBS detected two m^5C motifs (Table 2.1). Based on the type of motif and homology to known enzymes, each motif was tentatively assigned to a DNA methyltransferase (Table 2.1). *Cthe_2470 and Cthe_1511* are both annotated as Dam methyltransferases, and *Cthe_1511* is encoded in a putative operon with the MboI-type restriction enzyme, suggesting that these enzymes target G**AT**C. The CNC**A**NNNNNNN**T**TC motif is consistent with Type I RM system motifs (two 3-4 base specific sites separated by 5-8 Ns), and so this is presumably targeted by the only Type I enzyme encoded in the genome – Cthe1144-1145. The non-palindromic sequence GTCAT is consistent with a Type III system, and so is likely targeted by the only encoded Type III enzyme, Cthe_0519, which is experimentally validated below. Cthe_2320 is annotated as a HaeIII family restriction endonuclease, which targets G**GC**C, suggesting that *Cthe_2321* methylates G**GC**C. The only remaining m^5C methyltransferase encoded in the genome is *Cthe_1749*, suggesting that it targets the remaining m^5C site – GCWGC. These results are consistent with the predictions from REBASE.

DNA Methyltransferases were functionally expressed in E. coli

To methylate plasmid DNA in the same way as *C. thermocellum* ATCC 27405, we engineered *E. coli* to mimic the *C. thermocellum* ATCC 27405 methylome. One of the motifs, G**A**TC, is natively methylated by the Dam methyltransferase in *E. coli.* Two additional methyltransferases were expressed

Table 2.1: Methylated motifs in *C. thermocellum* ATCC27405.

Methylated bases are in bold. In cases where T or G are bold, the methylation is on the A or C of the complementary strand, respectively. % m^6A motifs were detected by PacBio SMRT sequencing, and m^5C motifs were detected by WGBS. Modified is the percentage of these motifs in the genome that were detected as methylated. # of motifs in the genome is the number of times each motif appears in the genome. The "Predicted Methyltransferase" is the most likely *C. thermocellum* gene responsible for each methylation. (N=Any base, W= A or T)

Motif	Type	% Modified	# Motifs in the Genome	Predicted Methyltransferase
G**AT**C	m^6A	74	8234	*Cthe_2470 and 1511*
CNC**A**NNNNNN**T**TC	m^6A	57.1	1775	*Cthe_1144-1145*
GTC**A**T	m^6A	49.6	6945	*Cthe_0519*
G**C**W**G**C	m^5C	100	6283	*Cthe_1749*
G**GC**C	m^5C	100	12192	*Cthe_2321*

from the *E. coli* chromosome. First, Cthe_0519 was integrated the chromosome to target GTC**A**T. Next, a previously characterized bifunctional *Bacillus* phage methyltransferase, *phi3TI* (51), was added to the chromosome. The *Phi3TI* methyltransferase is known to methylate both G**C**W**G**C and G**GC**C, the same sites methylated in *C. thermocellum* by Cthe_1749 and 2321, so expressing this gene enabled methylation of both sites via the expression of a single gene. Cthe_0519 and *phi3TI* were added to the chromosome of an *E. coli dam*$^{+}$ *dcm*$^{-}$ strain, so that Dam natively methylates GATC, a methylated motif observed in *C. thermocellum*, but Dcm does not methylate C**C**WGG, which was not seen in the *C. thermocellum* methylome analysis. The last methylated motif, CNC**A**NNNNNN**T**TC, does not occur on pNJ020, and therefore the corresponding methyltransferase was not expressed from the *E. coli* chromosome.

The in vivo activity of Cthe_0519 and Phi3TI methyltransferases was determined by restriction enzyme digestion to examine the extent of methylation by our heterologously expressed methyltransferases. Many commercially-available restriction enzymes are blocked by overlapping DNA methylation. Therefore, a restriction enzyme was chosen for each motif that either partially overlaps the motif of interest or, when possible, targets the exact same motif. If the plasmid is properly methylated, then the enzyme will be blocked by the methylation and no cutting will occur. For instance, restriction enzyme NlaIII cuts unmethylated CATG, and a fraction of these NlaIII sites will overlap with GTCAT (when the sequence GTCATG occurs) and will not be cut (Figure 2.1A). For pNJ020, methylation of GTC**A**T by Cthe_0519 blocks NlaIII-cutting between a 393 bp and 62 bp band, resulting in the generation of 455 bp band instead (Figure 2.1A). Complete blockage of this cut site, and therefore compete methylation, was observed for Cthe_0519 (Figure 2.1B, Lane 1, 2), where complete disappearance of the 393 bp band and generation of the new 455 bp band is observed. For the G**GC**C motif, restriction enzyme HaeIII targets the same sequence, so we would expect complete blockage of restriction activity if methylation of this sequence occurs in *E. coli*. We linearized the plasmid with HindIII and digested with HaeIII (Figure 2.1).

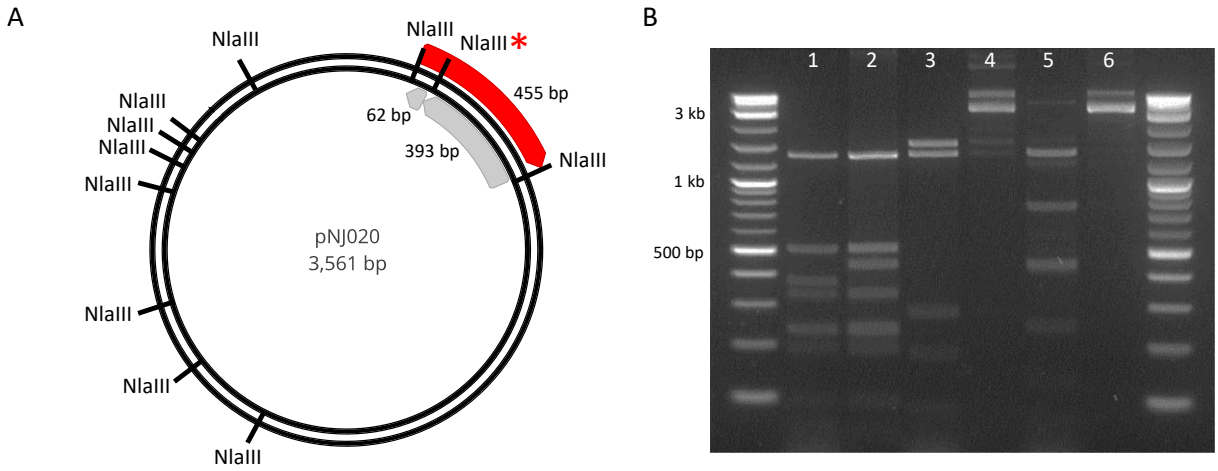


Figure 2.1: Confirmation of methyltransferase functionality in *E. coli*.

A) pNJ020 plasmid map with NlaIII restriction sites, including NlaIII* that overlaps with the GTCAT methylation site. When blocked by methylation, NlaIII* is not cut by NlaIII, resulting in a 455 bp band rather than 393 and 62 bp bands. B) Agarose gel of restriction digests of plasmid pNJ020 isolated from *E. coli* Top10 (Lanes 1,3, and 5) and from methylating strain AG2006 (Lanes 2, 4, and 6). Methylation at GTC**A**T is indicated by a band shift from 393 bp to 455 bp (Lane 1 vs 2) or complete blockage of digestion by G**GC**C (Lane 3 vs 4) and G**C**W**G**C (Lane 5 vs 6).

Unmethylated DNA (Lane 3) was completely digested, while Phi3TI methylation mostly blocked HaeIII digestion (Lane 4), suggesting nearly complete methylation of G**GC**C by Phi3TI. For G**C**W**G**C methylation, enzyme TseI targets the same motif, so the plasmid was linearized with HindIII and digested with TseI. Unmethylated DNA (Lane 5) was completely digested, while Phi3TI methylation fully blocked TseI digestion (Lane 6), suggesting complete methylation of G**C**W**G**W by Phi3TI.

Methylated DNA allows transformation of C. thermocellum ATCC 27405

To determine if targeted DNA methylation would allow transformation of *C. thermocellum* ATCC 27405, we tested transformation efficiency using pNJ020. This plasmid contains 8 GTC**A**T, 7 G**C**W**G**C, 5 G**GC**C, 5 G**A**TC, and no CNC**A**NNNNNN**T**TC sites. Plasmid DNA was methylated by the control strain Top10 *Δdcm*, which only methylates GATC, and pNJ020 methylated with Cthe_0519 and Phi3TI, which methylates G**A**TC, GTC**A**T, G**GC**C (partially), and G**C**W**G**C, where all putative restriction systems should be evaded. The control transformation in which pNJ020 was only methylated with native *E. coli* Dam yielded no colonies, while the methylated plasmid yielded an average of 80 colony forming units (CFU)/microgram of plasmid DNA. Transformation of plasmid pAMG216 was also tested, which contains one Type I motif, and an average of 40 CFU/μg was observed (Table 2.2).

Discussion

Development of genetic systems in non-model microbes is a grand challenge for the study of both fundamental and applied microbiology. Here, we show a rational, systematic process for developing transformation methods in *C. thermocellum* ATCC 27405 by overcoming the native RM systems. This was done by first identifying the methylated motifs using high-throughput sequencing techniques, including the rarely used WGBS technique for identifying m^5C motifs. Methylome analysis

Table 2.2: Transformation efficiency of *C. thermocellum* ATCC 27405 using methylated plasmid DNA.

No transformation was seen using control DNA without *C. thermocellum* methylation. Units are in colony forming units (CFU) per microgram of plasmid DNA

Plasmid	**Experiment**			**Average**
	1	**2**	**3**	**(CFU/μg)**
pNJ020	27	62	151	80
pAMG216	14	20	85	40

was followed by heterologous expression of targeted methyltransferases from the *E. coli* chromosome to mimic the *C. thermocellum* ATCC 27405 methylome. Plasmid DNA was passaged through this *E. coli* strain to methylate it, and functionality of the methylases was confirmed through restriction enzyme digest of the plasmid DNA. Properly methylated plasmid avoids the native RM systems and allowed transformation using commercially available equipment, which opens the door to metabolic engineering and the development of more advanced genetic tools. By enabling reliable transformation, the type strain of *C. thermocellum* can now be studied through genetic manipulation.

The first step to understanding and overcoming the native RM systems in *C. thermocellum* ATCC 27405 and other organisms is identifying the methylated motifs and the cognate methyltransferases. PacBio SMRT sequencing is the most commonly utilized method for microbial methylome analysis and can accurately identify m^6A and m^4C motifs. SMRT sequencing revealed three methylated motifs, and the corresponding RM systems were identified using REBASE. While SMRT sequencing is also able to detect m^5C motifs, it does not always identify these motifs (20). For example, the two m^5C motifs in *C. thermocellum* ATCC 27405 were not discovered in the SMRT methylome analysis. Therefore, WGBS is a vital step to reliably determine the full methylome of a bacterial strain. By using WGBS on an Illumina platform, we were readily able to detect m^5C motifs with less sequencing coverage. Recently, Oxford Nanopore sequencing has also been shown to detect DNA methylation (52). As technologies advance, the simplicity of methylome analysis will likely increase.

Previous studies have expressed methyltransferases to methylate DNA prior to transformation, but functionality of these enzymes beyond examining the impact on transformation efficiency is rarely tested. Here, we tested functionality of the expressed methyltransferases through a restriction enzyme digestion assay in which restriction activity is blocked when the plasmid is methylated. While the Cthe_0519 methyltransferase fully methylated the plasmid DNA, the Phi3TI methyltransferase only partially methylated the DNA. By digesting plasmid DNA isolated from the *E. coli* methylation strain with

an enzyme that overlaps with the methylated motif of interest, functionality can be easily determined by the percentage of DNA cut/uncut. This approach also unambiguously confirmed that Cthe_0519 methylates GTC**A**T.

An alternative approach to evading RM systems is to use plasmids that lack the targeted sequence. We used this approach for the putative Type I RM system (Cthe_1144-1145), where the motif was avoided by using plasmid DNA (pNJ020) that does not contain the motif CNC**A**NNNNNN**T**TC. Thus, isolation of pNJ020 out of the *E. coli* methylation strain results in plasmid DNA that fully mimicked the ATCC 27405 methylome. While this approach can be helpful, it is not always possible to avoid or remove the target motifs. For instance, when the motif is short, such as a four-base recognition sequence, it may not be feasible to remove them all. Some motifs may also happen to be in critical parts of the sequence, such as when they are in the origin of replication or in a sequence being used for homologous recombination. Therefore, expression of methyltransferases in *E. coli* will likely continue to be an attractive approach for developing transformation tools in new organisms.

Interestingly, Cthe_1144-1145 seems to not be part of an active restriction system. It lacks a predicted restriction enzyme, and when transformation of a plasmid containing the corresponding Type I motif (pAMG216) was tested, there was not a substantial difference in transformation efficiency relative to pNJ020, suggesting that this system indeed lacks a restriction enzyme. The two-fold difference in efficiency can likely be explained by the difference in plasmid size (pNJ020 is 3,561 bp, while pAMG216 is 4,700 bp), where the smaller plasmid has a greater number of plasmid molecules per microgram of DNA and may enter the cell more easily due to its smaller physical size.

We have demonstrated reproducible transformation of *C. thermocellum* strain ATCC 27405 and are now poised to improve these methods to increase frequency. Improved methylation by the Phi3TI methyltransferase will likely increase transformation efficiency. Also, for reasons unknown, strain ATCC

27405 does not form a cell pellet during centrifugation as well as strain DSM 1313. Therefore, identifying growth conditions under which ATCC 27405 forms tight cell pellets would make competent cell preparation simpler and presumably increase competent cell concentration, potentially leading to increased efficiencies. However, all approaches for genetic manipulation in *C. thermocellum* to date have relied on use of replicating plasmids, even for gene deletions (35), so the full suite of *C. thermocellum* genetic tools is now available in strain ATCC 27405, even at the current transformation efficiency.

In conclusion, the combination of complete methylome analysis via both PacBio SMRT sequencing and WGBS and validated DNA methyltransferase expression allowed the rational development of transformation methods for *C. thermocellum* ATCC 27405. We anticipate that the approach for obtaining transformation demonstrated in this work may be applied to many other bacterial strains, especially in new, non-model organisms.

References

1. Davis RH. The age of model organisms. Nature Reviews Genetics. 2004;5:69.
2. Yan Q, Fong SS. Challenges and Advances for Genetic Engineering of Non-model Bacteria and Uses in Consolidated Bioprocessing. Frontiers in Microbiology. 2017;8(2060).
3. Zeldes BM, Keller MW, Loder AJ, Straub CT, Adams MW, Kelly RM. Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals. Front Microbiol. 2015;6:1209.
4. Arber W, Linn S. DNA modification and restriction. Annu Rev Biochem. 1969;38:467-500.
5. Murray NE. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol Mol Biol Rev. 2000;64(2):412-34.
6. Lauster R, Trautner TA, Noyer-Weidner M. Cytosine-specific type II DNA methyltransferases. A conserved enzyme core with variable target-recognizing domains. Journal of molecular biology. 1989;206(2):305-12.
7. Pingoud A, Jeltsch A. Recognition and cleavage of DNA by type-II restriction endonucleases. Eur J Biochem. 1997;246(1):1-22.
8. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, et al. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res. 2003;31(7):1805-12.
9. Liu G, Ou HY, Wang T, Li L, Tan H, Zhou X, et al. Cleavage of phosphorothioated DNA and methylated DNA by the type IV restriction endonuclease ScoMcrA. PLoS Genet. 2010;6(12):e1001253.
10. Purdy D, O'Keeffe TA, Elmore M, Herbert M, McLeod A, Bokori-Brown M, et al. Conjugative transfer of clostridial shuttle vectors from Escherichia coli to Clostridium difficile through circumvention of the restriction barrier. Mol Microbiol. 2002;46(2):439-52.
11. Yasui K, Kano Y, Tanaka K, Watanabe K, Shimizu-Kadota M, Yoshikawa H, et al. Improvement of bacterial transformation efficiency using plasmid artificial modification. Nucleic Acids Res. 2009;37(1):e3.
12. Sandoval NR, Venkataramanan KP, Groth TS, Papoutsakis ET. Whole-genome sequence of an evolved Clostridium pasteurianum strain reveals Spo0A deficiency responsible for increased butanol production and superior growth. Biotechnology for Biofuels. 2015;8(1):227.
13. Mermelstein LD, Welker NE, Bennett GN, Papoutsakis ET. Expression of Cloned Homologous Fermentative Genes in Clostridium Acetobutylicum ATCC 824. Bio/Technology. 1992;10(2):190-5.
14. Chung D, Farkas J, Westpheling J. Overcoming restriction as a barrier to DNA transformation in Caldicellulosiruptor species results in efficient marker replacement. Biotechnology for biofuels. 2013;6(1):82.
15. Blow MJ, Clark TA, Daum CG, Deutschbauer AM, Fomenkov A, Fries R, et al. The Epigenomic Landscape of Prokaryotes. PLoS Genet. 2016;12(2):e1005854.
16. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Banerjee O, et al. Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing. Nature biotechnology. 2012;30(12):1232-9.
17. Murray IA, Clark TA, Morgan RD, Boitano M, Anton BP, Luong K, et al. The methylomes of six bacteria. Nucleic Acids Res. 2012;40(22):11450-62.
18. Lee WC, Anton BP, Wang S, Baybayan P, Singh S, Ashby M, et al. The complete methylome of Helicobacter pylori UM032. BMC Genomics. 2015;16(1):424.

19. O'Loughlin JL, Eucker TP, Chavez JD, Samuelson DR, Neal-McKinney J, Gourley CR, et al. Analysis of the Campylobacter jejuni genome by SMRT DNA sequencing identifies restriction-modification motifs. PLoS One. 2015;10(2):e0118533.
20. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-89.
21. Pyne ME, Bruder M, Moo-Young M, Chung DA, Chou CP. Technical guide for genetic advancement of underdeveloped and intractable Clostridium. Biotechnology advances. 2014;32(3):623-41.
22. Clark TA, Lu X, Luong K, Dai Q, Boitano M, Turner SW, et al. Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. BMC Biology. 2013;11(1):4.
23. Yu M, Ji L, Neumann DA, Chung D-h, Groom J, Westpheling J, et al. Base-resolution detection of N 4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite-sequencing. Nucleic Acids Res. 2015;43(21):e148-e.
24. Suzuki H, Takahashi S, Osada H, Yoshida K. Improvement of transformation efficiency by strategic circumvention of restriction barriers in Streptomyces griseus. Journal of microbiology and biotechnology. 2011;21(7):675-8.
25. Johnston CD, Skeete CA, Fomenkov A, Roberts RJ, Rittling SR. Restriction-modification mediated barriers to exogenous DNA uptake and incorporation employed by Prevotella intermedia. PLOS ONE. 2017;12(9):e0185234.
26. Akinosho H, Yee K, Close D, Ragauskas A. The emergence of Clostridium thermocellum as a high utility candidate for consolidated bioprocessing applications. Frontiers in Chemistry. 2014;2(66).
27. Ng TK, Weimer TK, Zeikus JG. Cellulolytic and physiological properties of Clostridium thermocellum. Arch Microbiol. 1977;114(1):1-7.
28. Ng TK, Zeikus JG. Comparison of Extracellular Cellulase Activities of Clostridium thermocellum LQRI and Trichoderma reesei QM9414. Applied and environmental microbiology. 1981;42(2):231-40.
29. Lamed R, Setter E, Bayer EA. Characterization of a cellulose-binding, cellulase-containing complex in Clostridium thermocellum. Journal of bacteriology. 1983;156(2):828-36.
30. Ellis LD, Holwerda EK, Hogsett D, Rogers S, Shao X, Tschaplinski T, et al. Closing the carbon balance for fermentation by *Clostridium thermocellum* (ATCC 27405). Bioresour Technol. 2012;103(1):293-9.
31. Raman B, Pan C, Hurst GB, Rodriguez M, Jr., McKeown CK, Lankford PK, et al. Impact of Pretreated Switchgrass and Biomass Carbohydrates on Clostridium thermocellum ATCC 27405 Cellulosome Composition: A Quantitative Proteomic Analysis. PLOS ONE. 2009;4(4):e5271.
32. Stevenson DM, Weimer PJ. Expression of 17 Genes in Clostridium thermocellum ATCC 27405 during Fermentation of Cellulose or Cellobiose in Continuous Culture. Applied and Environmental Microbiology. 2005;71(8):4672-8.
33. Wilson CM, Rodriguez M, Johnson CM, Martin SL, Chu TM, Wolfinger RD, et al. Global transcriptome analysis of Clostridium thermocellum ATCC 27405 during growth on dilute acid pretreated Populus and switchgrass. Biotechnology for biofuels. 2013;6(1):179.
34. Poudel S, Giannone RJ, Rodriguez M, Raman B, Martin MZ, Engle NL, et al. Integrated omics analyses reveal the details of metabolic adaptation of Clostridium thermocellum to lignocellulose-derived growth inhibitors released during the deconstruction of switchgrass. Biotechnology for Biofuels. 2017;10(1):14.
35. Olson DG, Lynd LR. Transformation of Clostridium thermocellum by electroporation. Methods in enzymology. 2012;510:317-30.

36. Papanek B, Biswas R, Rydzak T, Guss AM. Elimination of metabolic pathways to all traditional fermentation products increases ethanol yields in Clostridium thermocellum. Metab Eng. 2015;32:49-54.
37. Rydzak T, Lynd LR, Guss AM. Elimination of formate production in Clostridium thermocellum. J Ind Microbiol Biotechnol. 2015;42(9):1263-72.
38. Tian L, Papanek B, Olson DG, Rydzak T, Holwerda EK, Zheng T, et al. Simultaneous achievement of high ethanol yield and titer in Clostridium thermocellum. Biotechnology for Biofuels. 2016;9(1):116.
39. Tripathi SA, Olson DG, Argyros DA, Miller BB, Barrett TF, Murphy DM, et al. Development of pyrF-based genetic system for targeted gene deletion in Clostridium thermocellum and creation of a pta mutant. Applied and environmental microbiology. 2010;76(19):6591-9.
40. Lin PP, Mi L, Morioka AH, Yoshino KM, Konishi S, Xu SC, et al. Consolidated bioprocessing of cellulose to isobutanol using Clostridium thermocellum. Metabolic engineering. 2015;31:44-52.
41. Tyurin MV, Desai SG, Lynd LR. Electrotransformation of Clostridium thermocellum. Applied and Environmental Microbiology. 2004;70(2):883-90.
42. Maki ML, Armstrong L, Leung KT, Qin W. Increased expression of β-glucosidase A in Clostridium thermocellum 27405 significantly increases cellulase activity. Bioengineered. 2013;4(1):15-20.
43. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 2015;43(Database issue):D298-9.
44. Klapatch TR, Demain AL, Lynd LR. Restriction endonuclease activity in Clostridium thermocellum and Clostridium thermosaccharolyticum. Applied Microbiology and Biotechnology. 1996;45(1):127-31.
45. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A. 2000;97(12):6640-5.
46. Guss AM, Olson DG, Caiazza NC, Lynd LR. Dcm methylation is detrimental to plasmid transformation in *Clostridium thermocellum*. Biotechnology for biofuels. 2012;5(1):30.
47. Haldimann A, Wanner BL. Conditional-Replication, Integration, Excision, and Retrieval Plasmid-Host Systems for Gene Structure-Function Studies of Bacteria. Journal of Bacteriology. 2001;183(21):6384-93.
48. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature methods. 2010;7(6):461-5.
49. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):pp. 10-2.
50. Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, Urich MA, et al. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome research. 2013;23(10):1663-74.
51. Mermelstein LD, Papoutsakis ET. In vivo methylation in Escherichia coli by the Bacillus subtilis phage phi 3T I methyltransferase to protect plasmids from restriction upon transformation of Clostridium acetobutylicum ATCC 824. Applied and Environmental Microbiology. 1993;59(4):1077-81.
52. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al. Mapping DNA methylation with high-throughput nanopore sequencing. Nature methods. 2017;14(4):411-3.

Chapter III

Dynamic evolution of defense systems targeting foreign DNA in *Streptococcus thermophilus* strains

This chapter is part of a manuscript to be submitted by Lauren A. Riley, E. Anne Hatmaker, Irenee C. Payne, Lexiang Ji, Robert Schmidt, Yunzhou Wei, Clare A. Edwards, Ralph Zhang, Michael P. Terns, Adam M. Guss:

Riley, L. A.*, Hatmaker, E. A*., Zhang, R., Edwards, C. A., Payne, I. C., Schmidt, R., Wei, Y., Terns, M.P., Guss, A.M., (2020) Dynamic Evolution of Defense Systems Targeting Foreign DNA in *Streptococcus thermophilus*. In Preparation

E. Anne Hatmaker performed the Type I restriction-modification alignment. Yunzhou Wei, Clare A. Edwards, Ralph Zhang, and Michael P. Terns identified the CRISPR systems and created the CRISPR interference assay plasmids. Lauren A. Riley and Ireene C. Payne transformed *S. thermophilus*. Lexiang Ji and Robert Schmidt performed Whole Genome Bisulfite Sequencing. Lauren A. Riley performed genomic comparisons of the restriction-modification systems and performed the activity assays. Lauren A. Riley, E. Anne Hatmaker, and Adam M. Guss wrote and revised the manuscript.

Abstract

Microbial cultures are often under threat from foreign DNA, such as bacteriophage, and microbes have evolved defense mechanisms to protect against these threats. Understanding of the evolution and functionality of defense systems is critical to understand community robustness. *Streptococcus thermophilus* is an important member of the community used in fermented dairy production. This species has evolved several defense systems including Restriction-Modification (RM) systems and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) systems. Genomic comparisons and activity assays in two strains, DGCC 7710 and JIM 8232, demonstrate the role of horizontal gene transfer (HGT) on the evolution of Type I systems. HGT was identified between JIM 8232 and its industrial co-culture partner, *Lactococcus lactis,* which lead to the creation of a hybrid Type I RM system, generating a novel targeted motif. Further, the Type I RM systems in the two strains

demonstrate the modularity of this system. The Type III-A CRISPR system was shown to be functional in JIM 8232, even though it is inactive in DGCC 7710. Evasion of RM systems in JIM 8232 allowed genetic transformation in the strain for the first time, enabling future genetic studies on the Type III-A CRISPR system. We add to a growing body of work dealing with microbial defense systems and provide a deeper understanding of the evolution and modularity of these systems within *S. thermophilus* strains.

Introduction

Microorganisms natively exist in complex communities, comprised of many different species, all living in a contiguous environment and interacting with each other (1). This allows the communities to thrive in diverse habitats and enables robustness against disturbances that impact a community's survival (2). Disturbances to microbial communities include the invasion of foreign viruses and other mobile DNA, and these invasions can have an economically devastating impact in commercial bacterial communities, especially within the dairy industry (3). Bacteriophages are a major driving force of community evolution, and bacterial communities have evolved immune systems to protect against an attack. The assimilation of immune systems is often through horizontal gene transfer events (4).

Within the dairy industry, the production of fermented products, such as yogurt and cheese, relies on important microbial communities comprised of lactic acid bacteria, including species *Streptococcus thermophilus* and *Lactococcus lactis* (5). *S. thermophilus* strains within the dairy industry have undergone extensive reductive evolution during which they have lost function of many genes, including virulence genes and genes related to carbohydrate metabolism (5, 6). Co-culture of *S. thermophilus* with other lactic acid bacteria has allowed for adaptation to the milk and yogurt environment through acquisition of lactose fermentation operons and exopolysaccharide (EPS) operons (6). Operons beneficial to the *S. thermophilus* immune system and the dairy environment have been

demonstrated to be acquired through horizontal gene transfer between strains of *S. thermophilus* and their industrial counterparts (5, 7).

The prokaryotic immune system is comprised of several different systems, with clustered regularly interspaced short palindromic repeat (CRISPR) and Restriction-modification (RM) systems as two of most important. These two systems are capable of identifying and degrading foreign genetic material by targeting specific DNA motifs. CRISPR systems confer adaptive immunity which allows for inclusion of new targets, or spacers, into the related CRISPR array as the cell encounters new threats. CRISPR systems are encoded in multigene complexes which include several CRISPR-associated (Cas) genes as well as an "array" which includes repeated short sequences of around 30 bp, interspersed with variable target sequences, usually from phages, which allow the cell to prevent future infection by degrading any sequences containing the target (8). CRISPR systems are typically divided into three types, based on the *cas* genes present, *cas3* in Type I, *cas9* in Type II, and *cas10* in Type III (9). CRISPR has been extensively studied in a few strains of *S. thermophilus*, most notably DGCC 7710, where the functionality of Cas9 was initially discovered, and the activity of several CRISPR systems have been demonstrated (10-12).

RM systems are innate immune systems, detecting the methylation state at particular bases within specific DNA recognition sequences, then cleaving DNA with foreign methylation patterns at these motifs (13). This confers resistance to foreign DNAs such as bacteriophages when they are released from related organisms that have different RM systems. RM systems are classified into four types. Type I systems are composed of a variable DNA specificity subunit which identifies specific DNA motifs (HsdS), a conserved methyltransferase which methylates the DNA (HsdM), and a conserved restriction endonuclease that is responsible for cutting unmethylated DNA (HsdR). The HsdS subunit contains two target recognition domains (TRD), which are DNA-binding domains that each recognize one half of the targeted motif. They are separated by an alpha helix, whose length determines number of

bases between the TRD half-sites, resulting in a pattern like EcoKI – AAC-NNNNNN-GTGC (14). Next, Type II and Type III systems are comprised of a methylation enzyme and a restriction enzyme that target the same sequence. Type IV systems only have a restriction subunit and cut methylated DNA rather than unmethylated DNA (13). RM systems are less studied in *S. thermophilus* (11), relative to CRISPR systems. In *S. thermophilus* DGCC 7710, two RM specificity subunits were detected in a proteomic study when the strain was infected with phage 2972, but no research has presented further evidence of RM functionality thus far (11, 15).

Cellular defense systems can shape interactions and gene flow within microbial communities, in addition to their industrial importance in species such as *S. thermophilus*. In a coevolution experiment within *S. thermophilus* it was found that CRISPR target inclusion drove phage evolution, highlighting the ecological importance of understanding defense mechanisms (16). In this study, we add to the understanding of the evolution of microbial defense systems and provide a deeper understanding of *S. thermophilus*, in particular the two strains at the center of this work, DGCC 7710 and JIM 8232 through analysis of the RM and CRISPR systems both through genomic comparisons and functionality assays.

Materials and Methods

Strains, growth, and culture conditions

S. thermophilus JIM 8232 and DGCC 7710 (17) were grown aerobically in LM17, which is M17 media (BD Difco) supplemented with 50 mL 10% lactose per Liter, at either 37°C or 42°C. Media was supplemented with 5 µg/mL Chloramphenicol (cm) when indicated. *Escherichia coli* strains, Stellar (*dam*- *dcm*-), Top10 (*dam*+ *dcm*+), Top10 Δ*dcm*::*frt* (18), and BW25113 Δ*dam*::*kan* (19), were grown with Luria-Bertani (LB) medium aerobically at 37°C with 15 µg/mL cm as needed.

Genome sequencing and methylome analysis

Genomic DNA was isolated from *S. thermophilus* JIM 8232 and DGCC 7710 and sent to the Joint Genome Institute (JGI, Walnut Creek, CA) for PacBio sequencing. For both *S. thermophilus* JIM 8232 and DGCC 7710, a PacBio SMRTbell™ library (10) was constructed and sequenced on the PacBio RS platform (Pacific Biosciences, Menlo Park, CA, USA). Methylome analysis was performed using the PacBio SMRT Analysis 2.2.0 platform pipeline RS Modification and Motif Analysis after mapping filtered reads to a reference genome using BLASR version 1 (20). Whole Genome Bisulfite sequencing (WGBS) was performed on both strains as previously described (18) To visualize the differences between two *S. thermophilus* strains, JIM 8232 and DGCC 7710 RM systems, EasyFig (21) was used to align single genes.

Plasmids

A pWAR vector with a chloramphenicol resistance marker, pWYZ68 or pNoMotif, was used as a control in all transformations and as the backbone for all cloned vectors. Plasmids containing Type I recognition sequences were constructed using Gibson Assembly (New England Biolabs). To construct the plasmids with both Type I motifs, pMotif-JIM for JIM 8232 and pMotif-DGCC for DGCC 7710, oligonucleotides containing the Type I motifs and homology to pWYZ68 were inserted into the KpnI site of pWYZ68. Plasmids were constructed containing each Type I motif in duplicate with a 100 bp spacer, pLAR126-129. To construct these plasmids gBlocks (Integrated DNA Technology) were integrated into the KpnI site of pWYZ68.

To construct the CRISPR target vectors, transcribed targets were expressed downstream of a P_{pgm} promoter sequence while non-transcribed targets were placed downstream of a rho-independent terminator sequence. See Table 3.1 for target and PAM/Protospacer Flanking Sequence (PFS). Oligos of the target sequence inserts were annealed and phosphorylated using PNK (NEB) prior to insertion. Following inverse PCR linearization of the pWAR vector with Phusion High-Fidelity DNA Polymerase

(NEB), blunt end ligation of the insert was completed with T4 DNA Ligase (NEB). A few of the vectors were created by inverse PCR using primers containing the target sequence as 5' overhangs.

All plasmids were transformed into *E. coli* Stellar *dam- dcm-* (Stellar) and Top10 *dam+ dcm+* (Invitrogen). The control plasmid was also transformed into Top10 Δ*dcm*::*frt* and BW25113 Δ*dam*::*kan* (19). Plasmid DNA was isolated using ZymoPure II plasmid midiprep kit (Zymo Research) and used for *S. thermophilus* transformations.

S. thermophilus electrocompetent cell preparation

The two strains of *S. thermophilus*, JIM 8232 and DGCC 7710, were each inoculated into 10 mL LM17 and incubated overnight at 37°C. 100 mL of LM17 broth was inoculated with 1% of the overnight culture and grown at 42°C until an OD_{600} range of 0.4-0.6. The culture was placed on ice for 15 minutes then centrifuged in a 500 mL bottle at 5000xg at 4°C. Supernatant was decanted and the pellet was resuspended in 2 mL of cold wash solution (10% glycerol + 0.4M sorbitol). The cells were transferred to two 1.5 mL Eppendorf tubes and centrifuged at 15000 RPM for 1 minute at 4°C. Wash step was repeated twice more with 1 mL wash solution. After the last wash the pellet was resuspended in 500 mL wash solution.

Testing plasmid interference and transformation efficiency

Fresh electrocompetent cells were prepared for each strain and transformed in duplicate at least twice, for an n=4 or more. To test efficiency, 25 µL of competent cells were transformed with 200 ng of each plasmid. Transformants were resuspended in 1 mL of cold recuperation solution (LM17 broth + 0.4M sorbitol + 20mM $MgCl_2$ + 2mM $CaCl_2$) and placed at 42°C, non-shaking. After a two-hour recovery period, 1-50 µL of the recovery was plated on LM17 agar with 5 µg/mL chloramphenicol and incubated for 36 hours at 37°C. Colonies were counted from each plate.

Table 3.1: The target sequence used for each CRISPR system in both strains of *S. thermophilus* and how they were determined.

Locus	Target & PAM/Protospacer Flanking Sequence (PFS)
Sth DGCC 7710 CC1 (Type II-A)	5'-TGTTTGACAGCAAATCAAGATTCGAATTGTTGAGAAA-3' The target sequence is taken from spacer1 of the CRISPR locus with the PAM sequence from *in silico* analysis as 5'-NNAGAAW-3'(22-24) .
Sth DGCC 7710 CC2 (Type III-A)	5'-TTTCTAGGAATGGGTAATTATAGCGAGCTAGAAAGCTAATTACATTGCTAACGCTT-3' The target sequence is taken from spacer1 of the CRISPR locus with the PFS from *Streptococcus thermophilus ϕ* 73. Both a transcribed and non-transcribed target were used.
Sth DGCC 7710 CC3 (Type II-A)	5'-AAATTCTAAACGCTAAAGAGGAAGAGGACATGGTG-3' The target sequence is taken from spacer1 of the CRISPR locus with the PAM sequence from *in silico* analysis as 5'-NGGNG-3'(22-24).
Sth DGCC 7710 CC4 (Type I-E)	5'-AGTTGAGAACATACCTATATCAATGGCCTCCCACGCATAAGC-3' The target sequence is taken from the spacer1 sequence of the CRISPR locus. The first nucleotide has been mutated to a 'C' in this case. The 5' tag sequence is derived from the target of spacer4. The PAM of *DGCC* 7710 CRISPR4 was previously determined by *in vitro* analysis to be 'AA'(25-27) .
Sth JIM 8232 CC1 (Type II-A)	5'-ATGAGGTCGAGAAGGAAAAACGATATGCAGATAGAAAAGT-3' The target sequence is taken from spacer1 of the CRISPR locus with the PAM sequence from *in silico* analysis as 5'-NNAGAAW-3'(22-24).
Sth JIM 8232 CC2 (Type III-A)	5'-CATTGCTAACGCTTATTTAGAAGTAGCGTTAGAATCAAGCTCGCTTGGTGATAGTGAAT-3' The target sequence is taken from Spacer1 of CRISPR2 with the PFS sequence derived from the target in *Streptococcus thermophilus ϕ* 5093. Both a transcribed and non-transcribed target were used.
Sth JIM 8232 CC3 (Type II-A)	5'-ACACCCCTCCTCTCTCTCATAGATATCCTTGTGTTCTTTCATCGTTGTAT-3' The target sequence is from a spacer of the CRISPR locus. Because of ambiguous directionality of the locus and no consensus PAM, 10 bp adjacent to the target in *Streptococcus thermophilus ϕ D4276* were added on either side.

Lactococcus lactis methyltransferase expression

The Type I methyltransferase and specificity subunits (*hsdMS*) from plasmid pUC063B from *Lactococcus lactis* strain UC063 was codon optimized and synthesized as a gBlock (Integrated DNA Technologies). The gBlock was cloned using NEBuilder into a replicating vector under control of an arabinose inducible promoter, p_{Bad}, and transformed into *E. coli* GT115. The resulting strain was grown with 1mM arabinose to induce methyltransferase expression and sent for PacBio sequencing and methylome analysis at SNPsaurus (Eugene, OR).

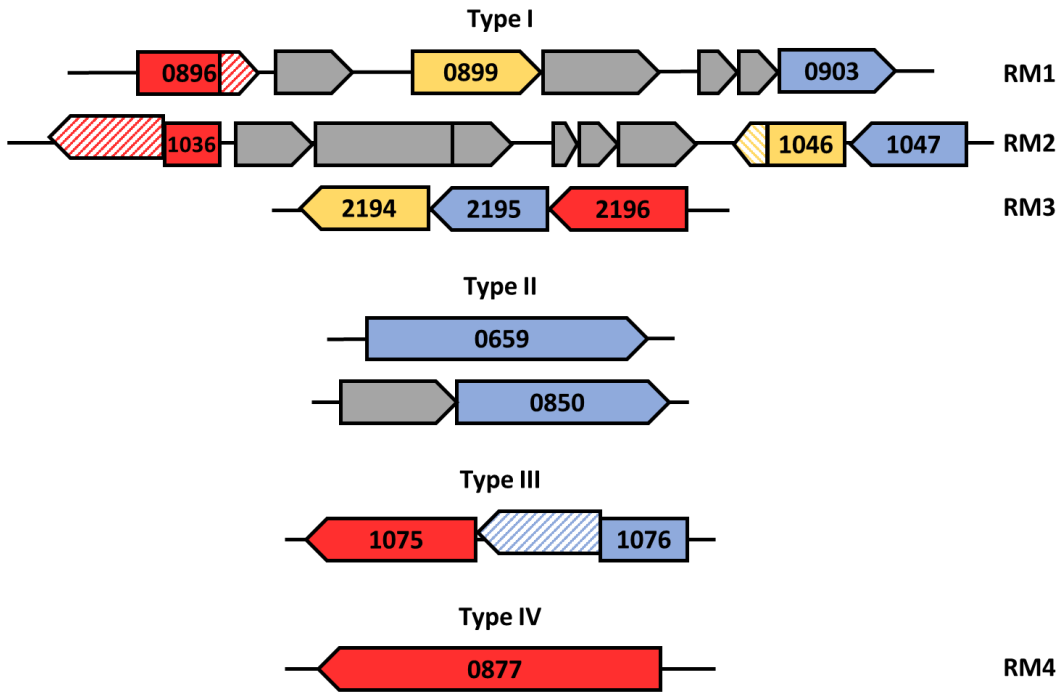
Results

Genomic comparison of restriction systems

The genomes of *Streptococcus thermophilus* JIM 8232 and DGCC 7710 were examined for the presence of putative RM systems using the New England Biolabs Restriction Enzyme Database (REBASE) (28) and genome annotations. *S. thermophilus* JIM 8232 encodes three putative Type I systems: STH8232_0896-0903 (RM system 1, or RM1), STH8232_1036-1047 (RM2), and STH8232_2194-2196 (RM3). Restriction enzyme STH8232_0896 contains a premature stop codon, and restriction enzyme STH8232_1036 contains frameshift mutation (Figure 3.1A). DNA specificity subunit STH8232_1046 also contains a premature stop codon which likely renders RM2 completely inactive. The JIM 8232 genome encodes two putative Type II methyltransferases, STH8232_0659 and STH8232_0850, although neither has a corresponding restriction enzyme predicted. JIM 8232 also encodes a putative Type III system, STH8232_1075-1076, but the methyltransferase gene contains a frame shift mutation. Finally, JIM 8232 is predicted to encode one Type IV restriction system, STH8232_0877 (RM4) (Figure 3.1A).

The second strain, *S. thermophilus* DGCC 7710, also encodes three predicted Type I RM systems: CW339_03630- 03665 (RM1), CW339_04305-04345 (RM2), and CW339_09150-09160 (RM3). Two genes have substitutions resulting in premature stop codons: the restriction enzyme CW399_09160 in RM3

A



B

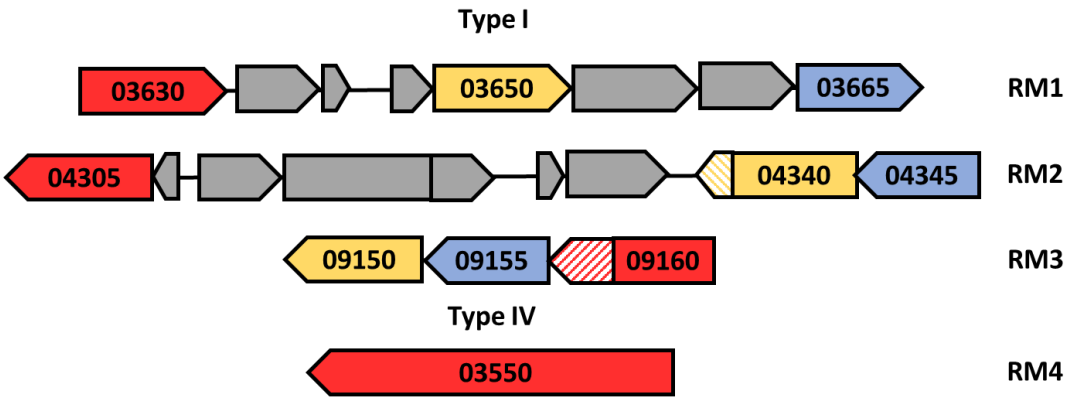


Figure 3.1: Restriction modification systems in *S. thermophilus* as predicted by New England Biolabs REBASE (28).

A) RM systems in JIM 8232 and B) RM systems in DGCC 7710. Red, restriction enzyme; Yellow, DNA specificity subunit; Blue, Methyltransferase. Premature stop codons are indicated with line arrows. Frameshift mutations are indicated with raised line arrows.

and the specificity subunit CW399_04340 in RM2. For both genes, all codons after the stop codon are in frame. DGCC 7710 also encodes a Type IV restriction enzyme, CW339_03550 (RM4) (Figure 3.1B). Functional characterization of the RM systems has not been previously examined for either organism.

The three Type I and the Type IV systems share substantial similarity between the two strains. Therefore, we compared the nucleotide identity for each orthologous pair, including the nucleotides after the premature stop codons and frameshifts. For RM1, the *hsdM* and *hsdR* subunits each share 99% identity between JIM8232 and DGCC 7710, while *hsdS* is less similar at 24% identity. For RM2, all three subunits are 99-100% identical between the strains. In RM3, *hsdM* and *hsdR* share 99% identity between the strains, while the *hsdS* is only 62% identical. Finally, the Type IV system genes are RM4 is 99% identical.

Methylome analysis

To begin functional characterization of these RM systems, we performed complete methylome analysis for both strains using Pacific BioSciences (PacBio) Single-Molecule Real-Time (SMRT) sequencing to detect 6-methyladenine and 4-methylcytosine motifs and Whole Genome Bisulfite Sequencing (WGBS) to detect 5-methylcytosine motifs. PacBio SMRT sequencing revealed two methylated Type I motifs consistent with Type I RM systems in each strain (Table 3.2) while WGBS detected no m^5C motifs in either strain.

Pairwise comparison between JIM 8232 and DGCC 7710 RM systems

Interestingly, the methylome analysis revealed a motif that is partially the same between the two strains, GAR**A**NNNNNNGC**T** from JIM 8232 and GAR**A**NNNNNNN**T**TCA from DGCC 7710 suggesting they may share a TRD targeting the GAR**A** half of the motif. While the HsdS subunits from RM1 are very different from each other, and the HsdS subunits from RM2 are nearly identical, the HsdS subunits from RM3 show an intermediate level of homology (see above). Upon deeper investigation, we found that the

Table 3.2: Methylome analysis of *S. thermophilus* JIM 8232 and DGCC 7710.

Capitalized bases represent 6-methyladenine on the forward (A) or reverse (T) strand. "% modified" indicates the percent of motifs in the genome where methylation was detected.

Strain	Methylated motifs	% Modified
S. thermophilus JIM 8232	GAR**A**NNNNNNGC**T**	99.6
	AGCNNNNNG**T**C	99.8
S. thermophilus DGCC 7710	GAR**A**NNNNNNNN**T**TCA	99.6
	GC**A**NNNNNNC**T**AR	99.0

that the RM3 HsdS C-terminal TRD is nearly identical between the two strains, but that the N-terminal TRD shows very limited similarity (Figure 3.2A, B). The RM3 HsdS subunit from *S. thermophilus* DGCC 7710 had corresponding identical sequences found in other *S. thermophilus* strains such as KLDS.SM, MN-BM-A02, ND07, ASCC1275, and C106.

The RM3 HsdS subunit from *S. thermophilus* JIM 8232, on the other hand, was not found in any other *S. thermophilus* strains. This suggests that the N-terminal TRD from JIM 8232 may have been acquired via HGT. To identify possible sources of this TRD, the non-homologous portion of RM3 HsdS was used as a BLAST query against the non-redundant protein database. Interestingly, an identical TRD was found encoded on a plasmid from a *L. lactis* strain, UC063. A pairwise comparison between the RM systems in *S. thermophilus* JIM 8232 and *L. lactis* strain UC063 revealed 99% nucleotide identity between the *S. thermophilus* system and the *L. lactis* system in the HsdM and HsdR subunits (Figure 3.2B, C). The first half of the DNA specificity subunit is highly similar between the two *S. thermophilus* strains, but the second half is not (Figure 3.2).

Expression of the L. lactis Type I methyltransferase

Based on the similarity to JIM 8232, we hypothesized that the motif recognized by the *L. lactis* sp. UC063 HsdS would be GC**T** for one TRD and an unknown sequence for the other. To test this, we expressed the *L. lactis* UC063 methyltransferase and DNA specificity subunit (*hsdMS*) in *E. coli* and performed PacBio SMRT analysis. Methylome analysis revealed a core motif with a TRD targeting **A**GC, or the complementary GC**T** (Table 3.3). REBASE (29) also predicts the activity of this system to be **A**GCNNNNNCC**T**.

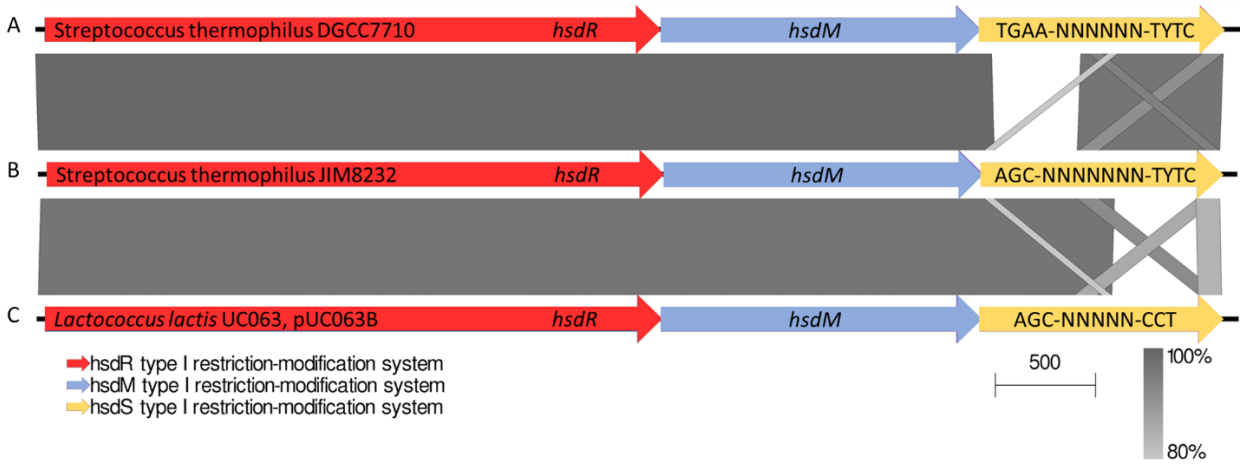


Figure 3.2: Pairwise comparison of homologous Type I RM systems from *S. thermophilus* strains and an *L. lactis* plasmid.

Table 3.3: Methylome analysis of *E. coli* expressing an *L. lactis* UC063 Type I methyltransferase.

Motif	% Modified
G**A**TC	99.9
AGCNNNNNCC**T**	99.7

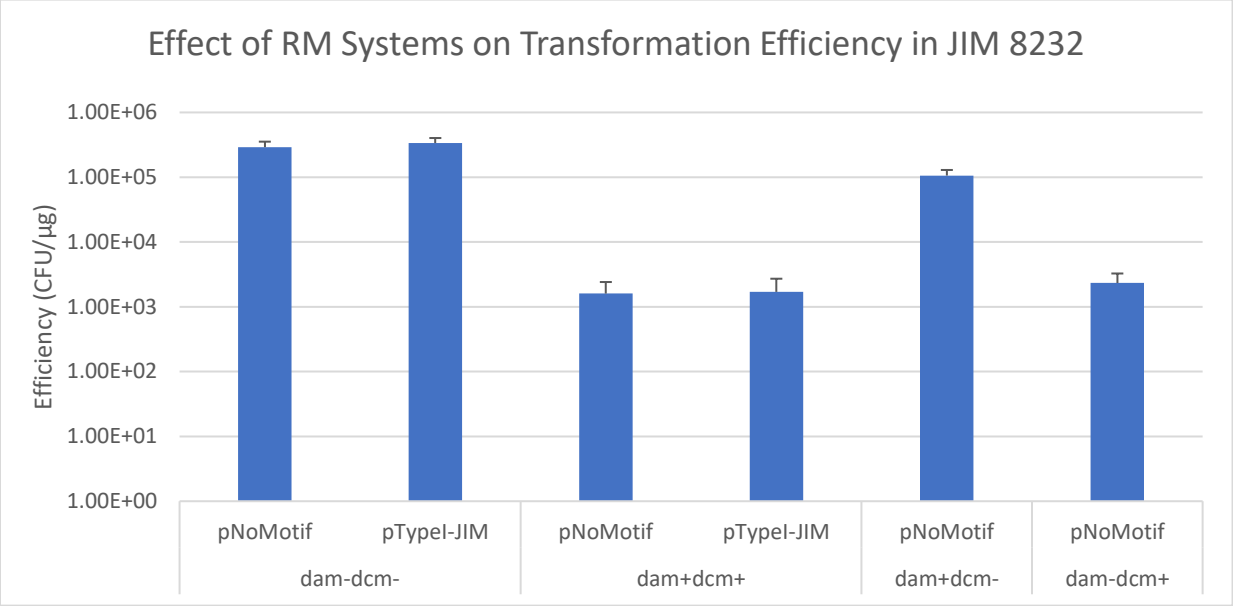
Functionality of R-M systems

The functionality of the RM systems was explored by testing the impact of each methylated motif on transformation efficiency in DGCC 7710 and JIM 8232. First, the Type I activity was tested through the comparison of transformation efficiencies (Colony forming units/ microgram DNA) using a control vector and a mutated vector containing additional Type I motifs. The control, pNoMotif, does not contain any motifs for JIM 8232 and natively contains one GC**A**NNNNNNC**T**AR recognition sequence for DGCC 7710. Plasmids were then created, pTypeI-DGCC and pTypeI-JIM, containing extra Type I motifs. Plasmid pTypeI-JIM carries one motif of each system in JIM 8232 and pTypeI-DGCC carries one extra GC**A**NNNNNNC**T**AR motif and two GAR**A**NNNNNNN**T**TCA motifs. JIM 8232 showed no difference in transformation efficiency between the control and pTypeI-JIM (Figure 3.3A). DGCC 7710, however, showed a 30-fold decrease in transformation efficiency with pTypeI-DGCC over the control (Figure 3.3B).

To determine whether one or both Type I RM system are contributing to the decrease in transformation efficiency in DGCC 7710, plasmids were created that each contain only one of the motifs. For each motif, a DNA fragment was inserted into the control vector containing two copies of the motif, separated by 100 random bases. Only the vector containing the GC**A**NNNNNNC**T**AR motif (pTypeI-GC**A**) in DGCC 7710 demonstrated decreased transformation efficiency, with transformation efficiency similar to the two-motif mutated vector (pTypeI-DGCC) (Figure 3.3B). Decrease of transformation efficiency when GC**A**NNNNNNC**T**AR is present is indicative of the RM1 in DGCC 7710 (Figure 3.1B).

Both *S. thermophilus* genomes encode putative Type IV restriction systems, which target methylated DNA. Therefore, to determine the impact of native *E. coli* Dam and Dcm methylation on the Type IV RM systems and transformation efficiency, transformation with two plasmid methylation states was compared, *dam+ dcm+* and *dam- dcm-*. JIM 8232 showed a decrease in transformation efficiency with methylated pNoMotif over the control, unmethylated pNoMotif (Figure 3.3A). To determine if Dam

A



B

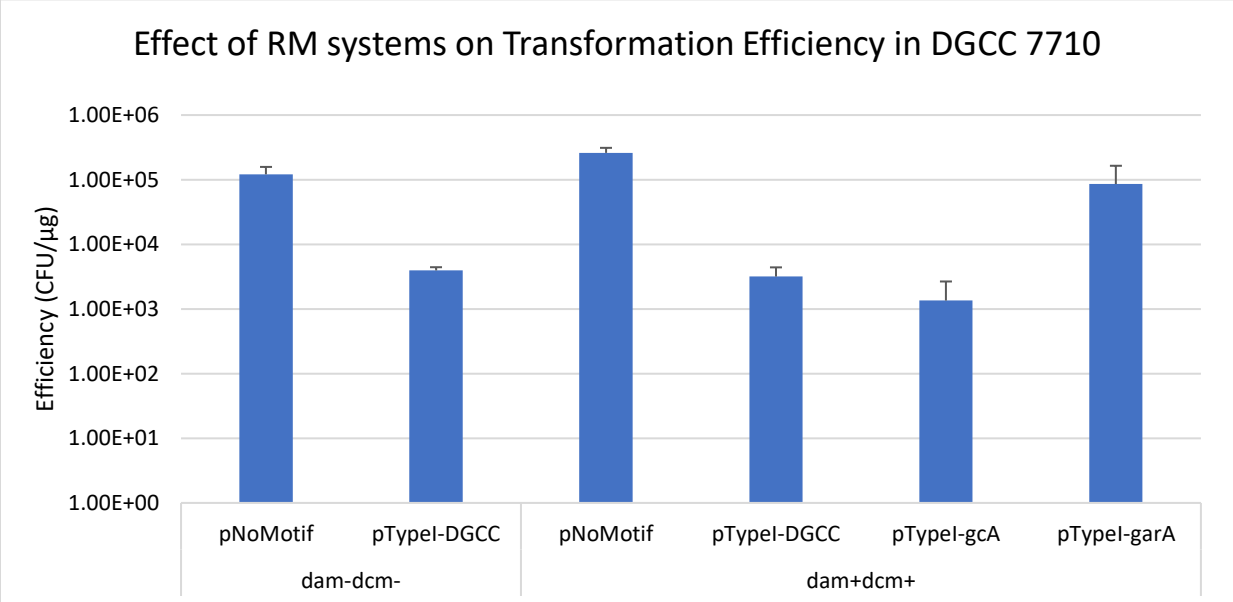


Figure 3.3: The effect of RM systems on transformation efficiency in *S. thermophilus*

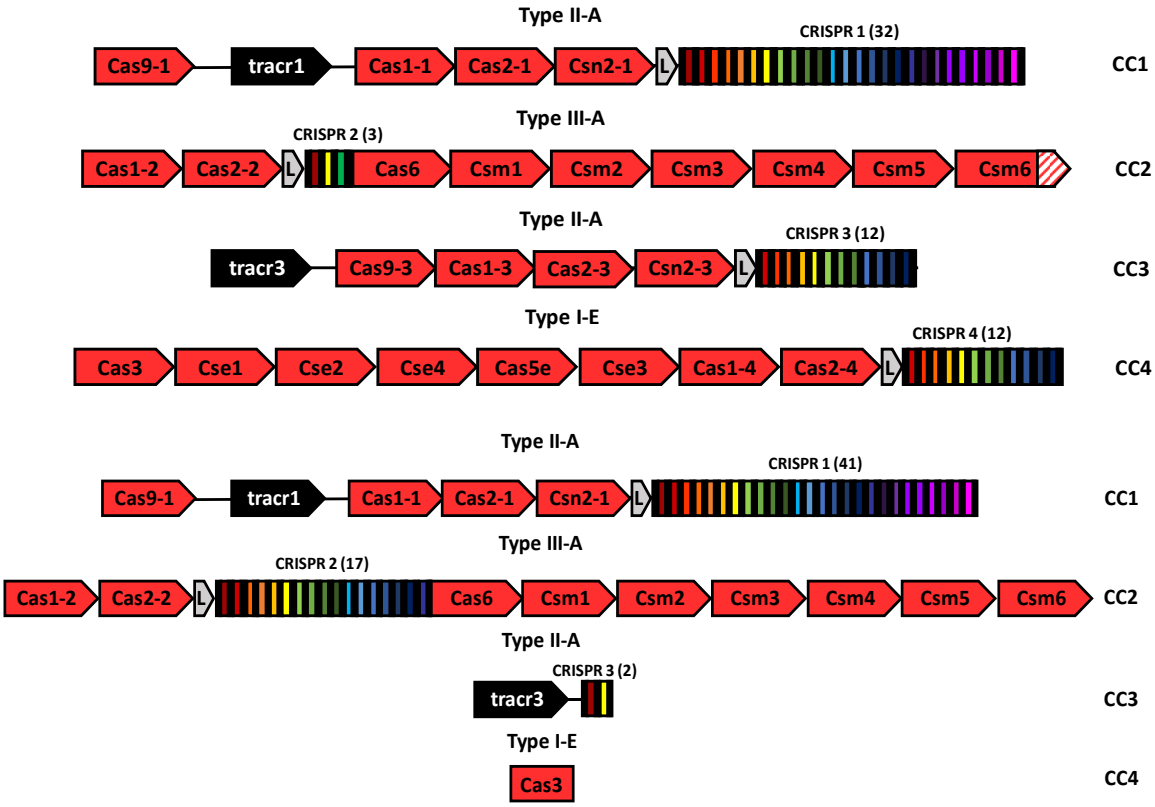
A) JIM 8232 B) DGCC 7710. A plasmid with no additional Type I motifs (pNoMotif) and plasmids with added motifs targeting each Type I RM system were compared. Each plasmid was either isolated out of an *E. coli dam+ dcm+* or *dam- dcm-* strain. Values represent the average and standard deviation of at least 4 replicates.

or Dcm methylation was responsible for the decrease in transformation efficiency, Dam and Dcm methylation were examined separately. The plasmid isolated out of *dam- dcm+* demonstrated decreased transformation efficiency, but *dam+ dcm-* DNA did not (Figure 3.3A). In *S. thermophilus* DGCC 7710, no difference in transformation efficiency was observed with methylated or unmethylated pNoMotif plasmid DNA (Figure 3.3B).

Genomic analysis of CRISPR/Cas systems

Next, the CRISPR systems were compared between the two strains. DGCC 7710 was previously shown to encode four CRISPR-Cas systems (CC1-CC4), including two Type II-A systems (CC1 and CC3), a Type III-A system (CC2), and a Type I-E system (CC4) (8). Each of the systems have a CRISPR array and encodes a cluster of *cas* genes (Figure 3.4A), but the *csm6* gene of CC2 is not expressed because of a premature stop codon (15). The crRNAs from each of the four loci are independently processed into type-specific crRNA species (30), the two Type IIA systems have previously been shown to be active in both adaptation and target interference (10, 22, 31-33), and CC4 has been shown interference active *in vitro* (25, 26). JIM 8232, in comparison, also has four systems: an intact Type II-A systems (CC1) and a Type III-A system (CC2), as well as an apparently degraded Type II-A CRISPR-array without any *cas* encoding gene adjacent to it and a single truncated cas3 encoding gene from a Type I-E system (Figure 3.4B) (34). The CRISPR/Cas systems in JIM 8232 have not previously been tested for functionality. The systems in each strain are highly similar, with the *cas* genes in the Type II-A, or CC1, systems being 99% identical. These two systems also have identical repeats, but the spacers in the CRISPR arrays are not identical in content or number. The type III-A *cas* genes are also highly similar with the *Cas1-6* and the *csm3-6* genes 99% identical and the *csm1* and *csm2* genes 95% identical. The lone truncated *cas3* in JIM 8232 is identical to DGCC 7710 for the first half and the second half is missing from JIM 8232.

A



B

Figure 3.4: CRISPR systems within *Streptococcus thermophilus* strains.

(A) CRISPR systems within *S. thermophilus* DGCC 7710. (B) CRISPR systems within *S. thermophilus* JIM 8232. Not to scale.

Functionality of CRISPR/Cas Systems

The CRISPR/Cas systems in both strains of *S. thermophilus* were tested for activity using a target interference assay to compare the activity to JIM 8232. Vectors containing target sequences for each system followed by the corresponding PAM were constructed. Transcribed targets were also constructed for the Type III-A system (CC2) in each strain by adding a promoter (P-CC2) (Table 3.1). Each vector was transformed into its corresponding host and transformation efficiencies (CFU/ µg DNA) were determined. Plasmids for JIM 8232 CC1 and the transcribed version of target sequence of CC2 showed target interference, while non-transcribed CC2 (NP-CC) and CC3 showed no target interference (Figure 3.5A). In DGCC 7710, CC1, CC3, and CC4 showed target interference while CC2 did not (Figure 3.5B), consistent with previous reports.

Discussion

Microbial communities are ever evolving their immune systems to protect against hostile DNA. Evolution of these systems and protection against invasion are critical to the survival of a community. Horizontal gene transfer (HGT) plays a large role in the spread and evolution of immune systems within a community. This process is important not only within natural environments, but also within commercial ones such as lactic acid bacterial communities in the dairy industry for the production of fermented milk products. While *S. thermophilus* DGCC 7710 CRISPR/Cas defense systems have been extensively examined (11, 23), other defense systems, including RM systems, have been largely understudied in all strains of *S. thermophilus*. Through genome comparisons and experimental validation, we demonstrate the heredity and functionality of RM systems and show evidence for gene flow between *S. thermophilus* and its dairy co-culture counterpart, *Lactococcus lactis.* This work expands our understanding of the evolution of immune systems in *S. thermophilus* and in microbial communities.

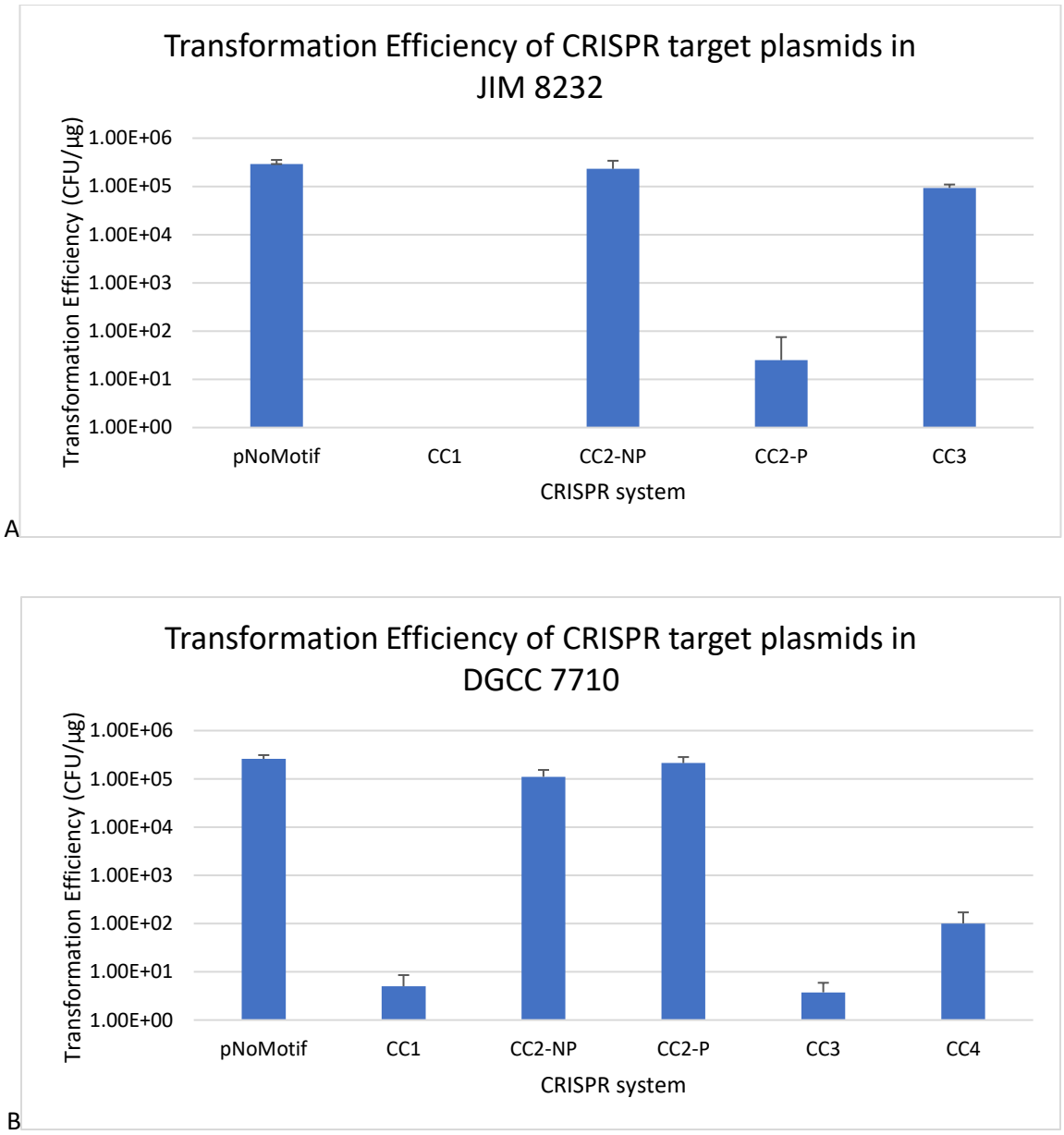


Figure 3.5: Activity of CRISPR systems in *S. thermophilus*

A) JIM 8232 B) DGCC 7710. Plasmids containing target and PAM or PFS for each CRISPR system were transformed into *S. thermophilus* and the transformation efficiencies are reported. For CC2 a non-transcribed (no promoter; CC2-NP) and transcribed (promoter; CC2-P) target was used.

Immune systems are frequently spread through HGT, occasionally leading to a vast array of systems, targeting dissimilar motifs within strains of the same species. *S. thermophilus* strains JIM 8232 and DGCC 7710 superficially contain highly similar putative RM and CRISPR systems, but upon closer investigation, they actually share few details in common. A specific example of HGT leading to the evolution of RM systems is demonstrated in the RM3 Type I systems. Methylome data revealed one TRD is identical between the strains, GAR**A**. Comparison of *S. thermophilus* DNA specificity subunits and those from other Type I systems revealed partial identity between Sth8232_2194-2196 and CW339_09150-09165, and partial identity between Sth8232_2194-2196 and the *L. lactis* sp. UC063 *hsdS*. The half of the DNA specificity subunit that is similar between the two strains of *S. thermophilus* recognizes the GAR**A** motif, while the half similar between JIM 8232 and *L. lactis* recognizes the GC**T** motif. It is common within the dairy industry communities for *S. thermophilus* and *L. lactis* to grow together, which presumably leads to HGT between the strains. This hybrid *hsdS* suggests the evolution of a new targeted motif and subunit within dairy industry co-cultures. This new heterogeneity presumably helps protect against invading bacteriophage that would destroy a monoculture. Thus, the RM systems in *S. thermophilus* are rampantly evolving to create new targeted motifs furthering the arms race between bacteria and hostile DNAs.

The *hsdS* in both the RM1 and the RM3 systems are dissimilar to each other and help to demonstrate the modularity of Type I RM systems. The hybrid RM3 enzyme described above demonstrates the modularity of HsdS subunit itself, where the TRDs can be swapped to create new combinations of targeted motifs and the alpha helix length can change, altering the number of N's between the targeted motifs. The two RM1 systems demonstrate the modularity of the whole system, where HsdS subunits can be swapped between strains that use the same HsdRM subunits. Here, the *hsdS* in DGCC 7710 is found in many other strains of *S. thermophilus,* while the *hsdS* in JIM 8232 is seemingly unique, with its origin presumably coming from a currently unsequenced organism. Thus, the

HsdS subunit is able to be replaced with a different subunit while maintaining the other parts to create a new RM system. This principle is useful when considering the idea of designer Type I RM systems. Complementation of a HsdS of interest with well characterized HsdM and HsdR subunits could be used to create a new functional RM system. The HsdS could be an already existing protein or it could be designed. A combination of different target recognition domains could be coupled with a predetermined alpha helix domain. The alpha helix is highly conserved and determines the number of bases between recognition sites. Designer specificity subunits and RM systems opens up the possibility of targeting DNA methylation using these enzymes.

Another example of evolved variability between the strains is the Type IV restriction systems, which are 99% identical to each other, but only the JIM 8232 Type IV targets the C**C**W**G**G motif methylated by *E. coli* Dcm. Three amino acids differ between the two orthologs, and it is currently unclear whether these changes result in inactivation or altered specificity of the enzyme. Alternatively, gene expression levels could be different, resulting in differential targeting of Dcm-methylated DNA. As RM systems evolve and base pair changes occur, the specificity and activity of the systems can change, leading to new recognition motifs and systems able to protect against new invading DNAs.

The four CRISPR/Cas systems in *S. thermophilus* DGCC 7710 have been extensively studied, and CC1, CC3, and CC4 have been previously shown active in interference assays. On the other hand, CRISPR in JIM 8232 has not been well studied due to the lack of genetic tools available. This strain has four CRISPR/Cas systems (CC1-4) that are highly similar to DGCC 7710, though CC3 and CC4 are highly degraded, missing most of the system. The target interference assay determined that CC1 and CC2 systems are active in JIM8232 while CC3 is not active due to a degraded CRISPR/Cas system. Due to the lack of a CRISPR array in CC4, it was not tested for functionality.

The CC2 system is only functional in JIM 8232, which is likely due to the early termination of the *csm6* gene in DGCC 7710. JIM 8232 has a complete set of CC2 CRISPR genes, including a functional *cms6* gene. Testing of both promoted and non-promoted CC2 targets in JIM8232 demonstrates activity of Type III-a systems as RNases, only the transcribed CC2 had decreased transformation efficiency (35). Previously, the Type III-A system was not studied in *S. thermophilus* DGCC 7710 due to the inactivity of the system in this strain but transformation of *S. thermophilus* JIM 8232 enables studies of a highly similar system.

Immune systems in bacteria are ever evolving to protect against invading DNAs, and this work provides insight into that evolution. The two strains of *S. thermophilus* at the center of this work illustrate defense system evolution and diversification, furthering the arms race. Other members of the microbial community can serve as sources of new DNA via HGT, allowing for the acquisition of new systems and even the creation of novel recognition sequences via hybridization between different specificity subunits. As new invading DNAs are introduced to a community the creation of this diversity is likely critical to community stability

References

1. Konopka A. What is microbial community ecology? The ISME Journal. 2009;3(11):1223-30.
2. Shade A, Peter H, Allison S, Baho D, Berga M, Buergmann H, et al. Fundamentals of Microbial Community Resistance and Resilience. Frontiers in Microbiology. 2012;3(417).
3. Garneau JE, Moineau S. Bacteriophages of lactic acid bacteria and their impact on milk fermentations. Microb Cell Fact. 2011;10 Suppl 1(Suppl 1):S20-S.
4. van Houte S, Buckling A, Westra ER. Evolutionary Ecology of Prokaryotic Immune Mechanisms. Microbiology and Molecular Biology Reviews. 2016;80(3):745-63.
5. Bolotin A, Quinquis B, Renault P, Sorokin A, Ehrlich SD, Kulakauskas S, et al. Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus. Nature Biotechnology. 2004;22:1554.
6. Alexandraki V, Kazou M, Blom J, Pot B, Tsakalidou E, Papadimitriou K. The complete genome sequence of the yogurt isolate Streptococcus thermophilus ACA-DC 2. Standards in Genomic Sciences. 2017;12:18.
7. Bellanger X, Roberts AP, Morel C, Choulet F, Pavlovic G, Mullany P, et al. Conjugative Transfer of the Integrative Conjugative Elements ICESt1 and ICESt3 from <em>Streptococcus thermophilus</em&gt. Journal of Bacteriology. 2009;191(8):2764.
8. Horvath P, Barrangou R. CRISPR/Cas, the Immune System of Bacteria and Archaea. Science. 2010;327(5962):167-70.
9. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, Horvath P, et al. Evolution and classification of the CRISPR-Cas systems. Nature reviews Microbiology. 2011;9(6):467-77.
10. Wei Y, Chesne MT, Terns RM, Terns MP. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic Acids Research. 2015;43(3):1749-58.
11. Dupuis M-È, Villion M, Magadán AH, Moineau S. CRISPR-Cas and restriction–modification systems are compatible and increase phage resistance. Nature Communications. 2013;4:2087.
12. Horvath P, Romero DA, Coûté-Monvoisin A-C, Richards M, Deveau H, Moineau S, et al. Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. Journal of Bacteriology. 2008;190(4):1401-12.
13. W Arber a, Linn S. DNA Modification and Restriction. Annual Review of Biochemistry. 1969;38(1):467-500.
14. Murray NE. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol Mol Biol Rev. 2000;64(2):412-34.
15. Young JC, Dill BD, Pan C, Hettich RL, Banfield JF, Shah M, et al. Phage-induced expression of CRISPR-associated proteins is revealed by shotgun proteomics in Streptococcus thermophilus. PLoS One. 2012;7(5):e38077.
16. Paez-Espino D, Sharon I, Morovic W, Stahl B, Thomas BC, Barrangou R, et al. CRISPR Immunity Drives Rapid Phage Genome Evolution in <span class="named-content genus-species" id="named-content-1">Streptococcus thermophilus</span&gt. mBio. 2015;6(2):e00262-15.
17. Hatmaker EA, Riley LA, O'Dell KB, Papanek B, Graveley BR, Garrett SC, et al. Complete Genome Sequence of Industrial Dairy Strain Streptococcus thermophilus DGCC 7710. Genome Announcements. 2018;6(6):e01587-17.

18. Riley LA, Ji L, Schmitz RJ, Westpheling J, Guss AM. Rational development of transformation in Clostridium thermocellum ATCC 27405 via complete methylome analysis and evasion of native restriction-modification systems. J Ind Microbiol Biotechnol. 2019.
19. Datsenko KA, Wanner BL. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. Proc Natl Acad Sci U S A. 2000;97(12):6640-5.
20. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13(1):238.
21. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27(7):1009-10.
22. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, et al. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J Bacteriol. 2008;190(4):1390-400.
23. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, et al. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J Bacteriol. 2008;190(4):1401-12.
24. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology. 2009;155(Pt 3):733-40.
25. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, Siksnys V. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. The EMBO journal. 2011;30(7):1335-42.
26. Sinkunas T, Gasiunas G, Waghmare SP, Dickman MJ, Barrangou R, Horvath P, et al. In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. The EMBO journal. 2013;32(3):385-94.
27. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun. 2012;3:945.
28. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 2015;43(Database issue):D298-9.
29. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Research. 2015;43(D1):D298-D9.
30. Carte J, Christopher RT, Smith JT, Olson S, Barrangou R, Moineau S, et al. The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in Streptococcus thermophilus. Mol Microbiol. 2014;93(1):98-112.
31. Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010;468(7320):67-71.
32. Magadan AH, Dupuis ME, Villion M, Moineau S. Cleavage of phage DNA by the Streptococcus thermophilus CRISPR3-Cas system. PLoS One. 2012;7(7):e40913.
33. Wei Y, Terns RM, Terns MP. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. Genes Dev. 2015;29(4):356-61.
34. Delorme C, Bartholini C, Luraschi M, Pons N, Loux V, Almeida M, et al. Complete genome sequence of the pigmented Streptococcus thermophilus strain JIM8232. J Bacteriol. 2011;193(19):5581-2.
35. Tamulaitis G, Venclovas Č, Siksnys V. Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. Trends in Microbiology. 2017;25(1):49-61.

Chapter IV

Simple and rapid site-specific integration of multiple heterologous DNAs in *E. coli*

This chapter is part of a manuscript to be submitted by Lauren A. Riley, Irenee C. Payne, Melissa Tumen-Velasquez, and Adam M. Guss:

Riley, L. A., Payne, I. C., Tumen-Velasquez, M., Guss, A. M., (2020) Site-specific recombinases for integration of heterologous DNA in *E. coli.* In Preparation

Lauren A. Riley constructed and tested the serine phage integration system. Melissa Tumen-Velasquez and Irenee C. Payne constructed the *C. clariflavum* 4-2a *E. coli* methylation strain. Lauren A. Riley and Irenee C. Payne transformed *C. clariflavum* 4-2a. Lauren A. Riley and Adam M. Guss wrote and revised the manuscript.

Abstract

Escherichia coli is the most studied and well understood microorganism, but even it can be limited by genetic tools, including the ability to rapidly integrate multiple DNA constructs efficiently into the chromosome. Site-specific, serine recombinases from bacteriophage can be useful tools, catalyzing a single, unidirectional recombination event between two specific DNA sequences, *attB* and *attP,* without requiring host proteins for functionality. Using these recombinases, we have developed a system to integrate stably up to twelve genetic constructs into in the *E. coli* chromosome sequentially. A cassette of *attB* sites was inserted into the chromosome and the corresponding recombinases were cloned onto temperature sensitive plasmids to mediate recombination between a non-replicating plasmid containing an *attP* site and the corresponding *attB* site on the chromosome. Using this approach, the efficiency of DNA insertion into the *E. coli* chromosome was approximately 10^7 cfu/µg DNA for six of the recombinases. The method was further streamlined via co-transformation of the recombinase plasmid and integrating plasmid, reaching efficiencies of about 10^5 cfu/µg DNA or higher for six recombinases. The integrating plasmid further contains ΦC31 attachment sites flanking the antibiotic gene, which can then be removed by transient expression of the ΦC31 recombinase, allowing reuse of antibiotic resistance genes for gene stacking. Using this system, eight DNA methyltransferases

from *Clostridium clariflavum* 4-2a were expressed from the *E. coli* chromosome to allow evasion of the *C. clariflavum* restriction systems and enable the first transformation of this cellulose-degrading bacterium as an illustrative example. This system expands the ability to simply insert DNA into the *E. coli* chromosome and provides a road map for multiple gene integration in other non-model organisms.

Introduction

Escherichia coli is one of the most important model organisms for bioengineering and it is arguably the most well-understood and studied (1, 2). A large genetic toolbox has been developed for *E. coli*, which includes methods for insertion of heterologous DNA into the chromosome. These tools include homologous recombination for vector integration, lambda Red recombinase-based recombineering and its newer CRISPR/Cas variants for gene knockouts and insertions (3-5), and site-specific recombinases for heterologous DNA integration into the chromosome (6, 7). Despite the genetic tools already available, research in *E. coli* can still be hindered by the speed and simplicity of genome engineering tools available. Homologous recombination tools to insert plasmid DNA into the chromosome typically take several steps and selections. Due to the labor and time involved, this technique is not often used in *E. coli*. The lambda Red recombinase system and CRISPR/Cas systems greatly improves upon homologous recombination in *E. coli* where markerless mutants can efficiently be created but it still has limits on the size of DNA that can be successfully integrated into the chromosome, with the efficiency greatly reduced above 3 kb (3).

Site-specific recombinase systems such as the conditional-replication, integration, and modular plasmid (CRIM) system use site-specific DNA recombinases from native *E. coli* phages, like λ, to integrate plasmids into specific spots in the chromosome. However, this system does not provide a way to remove the plasmid backbone and the selectable markers (8), and it is limited to five insertion sites. Site-specific recombinases enable a single recombination event between two specific DNA sequences, which enable

DNA integration and excision events to occur (9). They are classified into two families based on the catalytic amino acid in the active site, tyrosine- or serine-recombinases. The tyrosine family includes the commonly used Cre and Flp recombinases and the λ integrase from the *E. coli* bacteriophage λ. Cre and Flp recognize identical recombination sites (*loxP* and *frt*, respectively), and therefore recombination is reversible (10). The integrase from phage λ catalyzes a single crossover event at two non-identical sites, *attB* (attachment site from the Bacterium) and *attP* (attachment site from the Phage), to create two new sites, *attL* and *attR* (Left and Right) (11). The *att* sites from this type of recombinase are about 350-450 bp in size with a common 15 bp core where recombination occurs (11, 12). The λ integrase also requires accessory protein IHF (integration host factor) for functionality, and therefore only functions in *E. coli* and very close relatives (12).

The second family of recombinases includes serine phage recombinases, such as the archetypical ΦC31 (7, 9, 11). These serine recombinases offer benefits over recombinases like Cre, Flp, and λ integrase. They function similarly to the λ integrase by enabling unidirectional recombination at attachment sites, *attB* and *attP*, to create two new sites, *attL* and *attR*. However, the *att* sites from the serine family are typically smaller than the tyrosine family, ranging in size from 35-100 bp (11). The unidirectionality enables a stable integration event that won't spontaneously excise from the genome. Furthermore, unlike λ integrase, this group does not require any host-specific factors. Therefore, large serine recombinases from phylogenetically diverse sources are able to function in a wide variety of heterologous hosts, and they have been developed as genetic tools in diverse members of the Bacteria, Archaea, and Eukaryota (6, 7, 13-15). They have also been used in *E. coli* as synthetic biology "parts" (16, 17). Beyond the ΦC31 recombinase, there are several well characterized serine recombinases that have been shown to function outside of their native host (7). These characteristics allow serine-phage recombinases to be useful in genome engineering.

Some applications require more than five successive insertions of heterologous DNA into the *E. coli* chromosome, exceeding the capacity of the CRIM system. One such application is the expression of DNA methyltransferases for the transformation of new bacteria. Restriction-modification (RM) systems have been demonstrated to be an important barrier to transformation of bacteria, where foreign DNA is recognized and destroyed (18-21). RM systems are comprised of a restriction enzyme that targets and degrades non-methylated DNA at a specific DNA motif, while the corresponding methyltransferase protects the host genome by methylating DNA at the same motif (22). Bacteria often encode more than one RM system, and to successfully transform an organism these RM systems need to be overcome through proper methylation of DNA prior to transformation. An example organism is *Clostridium clariflavum* strain 4-2a which is of interest for consolidated bioprocessing and biofuel research due its ability to degrade lignocellulosic biomass via the production of an enzyme complex, called the cellulosome (23). This strain has not yet been transformed, which is likely due to the eight predicted systems. Evasion of the *C. clariflavum* RM systems will require the expression of more than five DNA methyltransferases from the *E. coli* chromosome. Therefore, we targeted expression of *C. clariflavum* methyltransferases as an example of the utility of this system.

We sought to expand upon the CRIM system by leveraging large serine recombinases to develop a system for inserting heterologous DNA into the *E. coli* chromosome. By further designing the system to allow simple removal of the selectable marker, we have developed a system that allows integration of up to twelve constructs without the accumulation of repetitive DNA and selectable markers. Using this system, we expressed eight methyltransferases encoded in the *C. clariflavum* 4-2a genome to methylate DNA and enable transformation for the first time in this strain.

Materials and Methods

Strains and Culture Conditions

E. coli Top 10 Δdcm::frt was used for all replicating vectors and integration of the *attB* cassette (21). PIR dependent plasmids, containing the oriR6K origin of replication, were propagated in GT115 (Invivogen). All *E. coli* strains were grown in Luria-Bertani (LB) broth with antibiotics, as necessary, for maintenance of vectors and for selection of integrants. Carbenicillin was used at 50 µg/mL, kanamycin was used at 50 µg/mL for replicating vectors and 30 µg/mL for single copy integrants, and spectinomycin and streptomycin together at 100 µg/mL for replicating vectors and 50 µg/mL for single copy integrants. All strains were grown at 37℃, unless the strain contains a temperature sensitive recombinase helper plasmid in which case it is grown at 30℃. *Clostridium clariflavum* 4-2a was grown in a Coy anaerobic chamber in CTUFD (21) at 50℃. Media was supplemented with 5 ug/mL thiamphenicol when needed. All strains created are listed (Table 4.1)

Construction integration helper plasmids

The first set of recombinase helper plasmids were constructed by GenScript, where the λ integrase was replaced in pINT-TS (8) with the synthesized serine recombinase that were previously described (7). Recombinases were codon optimized for expression in *E. coli*.

The second set of recombinases were cloned under the *Clostridium thermocellum* DSM 1313 enolase promoter (24) with a modified RBS to be consensus, AGGAGGA. First, ΦC31 was synthesized with the enolase promoter into a pUC replicating vector. The promoter and recombinase were then polymerase chain reaction (PCR) amplified using Phusion High Fidelity master mix (Thermo Scientific) and inserted using Gibson Assembly (New England Biolabs – NEB) into pINT-TS, replacing the native promoter and λ recombinase (pLAR047). Using pLAR047 as a backbone, the remaining serine recombinases were cloned and inserted, replacing ΦC31, using Gibson Assembly.

Table 4.1: Strains utilized for the serine phage integration system.
Strain collection number, genotype, and the strain description

Strain #	Genotype	Description
AG583	Top10 Δ*dcm*::*frt*	
AG2004	GT115 + pLAR080	pAH144+*attB* cassette (*aad*+)
AG2005	AG 583 HK::pLAR080	*attB* cassette in the chromosome of an *araC*- *dcm*- strain
AG3525	WM3188 Δ*dcm::frt* HK::pLAR080	*attB* cassette in the chromosome of an *araC*+ *dcm*- strain
AG4277	BW25113 Δ*mcrA::frt* Δ*mcrC-mrr::frt* Δ*dcm::frt* Δ*dam::frt* HK::pLAR080	*attB* cassette in the chromosome of an *araC*+ *dam*- *dcm*- strain
AG1434	GT115 + pGSs009	Non-replicating integration plasmid with p_{T5lac} and lacZα
AG3409	GT115 + pLAR067	Non-replicating integration plasmid with p_{Bad} and lacZα
AG3273	GT115 + pLAR031	Non-replicating *attP*-cassette plasmid
AG5425	GT115+ pLAR140	pGSs009 with BxB1 *attP* site
AG1460	AG583 + pGSs037	pINT-TS ΦC31
AG1461	AG583 + pGSs038	pINT-TS RV
AG1463	AG583 + pGSs040	pINT-TS TG1
AG1464	AG583 + pGSs041	pINT-TS BxB1
AG1484	AG583 + pGSs042	pINT-TS A118
AG1485	AG583 + pGSs043	pINT-TS ΦC1
AG1486	AG583 + pGSs044	pINT-TS ΦK38
AG1487	AG583 + pGSs045	pINT-TS MR11
AG1488	AG583 + pGSs046	pINT-TS R4
AG1489	AG583 + pGSs047	pINT-TS SPβc
AG1490	AG583 + pGSs048	pINT-TS TP901-1
AG1492	AG583 + pGSs050	pINT-TS Φ370
AG1495	AG583 + pGSs053	pINT-TS BL3
AG1496	AG583 + pGSs054	pINT-TS ΦBT1

Table 4.1: Continued

Strain #	Genotype	Description
AG1524	AG583 + pGSs082	pINT-TS Wβ
AG3463	AG583 + pLAR074	P_{Ct_eno} -BxB1
AG3512	AG583 + pLAR052	P_{Ct_eno} -Φ370
AG3513	AG583 + pLAR053	P_{Ct_eno} -A118
AG3515	AG583 + pLAR055	P_{Ct_eno} - ΦC1
AG3516	AG583 + pLAR056	P_{Ct_eno} -ΦK38
AG3517	AG583 + pLAR057	P_{Ct_eno} -MR11
AG3518	AG583 + pLAR058	P_{Ct_eno} -R4
AG3520	AG583 + pLAR060	P_{Ct_eno} -TG1
AG3521	AG583 + pLAR061	P_{Ct_eno} -TP901-1
AG3522	AG583 + pLAR062	P_{Ct_eno} -BL3
AG3524	AG583 + pLAR047	P_{Ct_eno} -ΦC31
AG 5477	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::pLAR080 R4::CloclaDRAFT_2133 phik38:: CloclaDRAFT _2368-2369 BXB1:: CloclaDRAFT _1830-1831 TP901:: CloclaDRAFT _1058 TG1:: CloclaDRAFT _1868-1869 BL3:: CloclaDRAFT _1961 phi370:: CloclaDRAFT _1994 A118:: CloclaDRAFT _1058	*C. clariflavum* 4-2a methylation strain
AG 1933	***Clostridium clariflavum* 4-2a**	

Construction and integration of attB cassette

The poly-*attB* cassette was synthesized by GenScript with all attachment sites, previously described (7), except ΦC31. The *attB* cassette was PCR amplified and inserted into the multiple cloning site of pAH144 (8) using Gibson Assembly (pLAR080). The *attB* cassette was then integrated into the HK phage site (8) of *E. coli* Top 10 Δ*dcm::frt,* BW25113 Δ*mcrA::frt* Δ*mcrC-mrr::frt* Δ*dcm::frt*, and BW25113 Δ*mcrA::frt* Δ*mcrC-mrr::frt* Δ*dcm::frt* Δ*dam::frt*, creating AG 2005, 3525, and 4277 respectively.

Construction of Integration Suicide Vector and attP Cassette

The base suicide integration plasmid, pGSs009, was synthesized and constructed by GenScript. It was constructed from pAH55 (8) and a synthesized insert, containing ΦC31 *att* sites flanking the T5lac promoter and lacZα. The promoter in pGSs009 was replaced with P_{BAD} from pLA2 (8) to create pLAR067. For recombinase efficiency experiments, a synthesized 13 *attP* cassette, excluding Spβ, was first inserted into the BamHI and NdeI sites of pGS009, using Gibson Assembly. Next, oligonucleotides with the Spβ *attP* and homology to pGSs009 were inserted into the EcoO1091 site, creating pLAR031.

Testing Integration Efficiencies in Two-Steps

To test integration efficiency of the first set of recombinase helper plasmids, each plasmid was transformed into the poly-*attB* cassette strain, AG 2005. Electrocompetent cells were made in duplicate of each strain and 200 ng of pLAR031 was electroporated into 25 microliters of each competent cell batch, in duplicate. Transformants were resuspended in 250 microliters of SOC and recovered at 37℃ for 1 hour and 42℃ for 30 minutes. The recovery was plated on LB and kanamycin plates and incubated at 37℃ overnight. Colonies were picked into 5mL LB with kanamycin and PCR screened after growth.

Testing Integration Efficiencies in One-Step

To test integration efficiency of the more highly expressed recombinase helper plasmids, electrocompetent cells were made of the poly-*attB* cassette strain in duplicate. Both the helper plasmid and pLAR031 were co-transformed in duplicate, using 200 ng of each plasmid, into 25 microliters of

electrocompetent cells. Transformants were resuspended in 250 microliters of S.O.C. and recovered at 37℃ for 1 hour. The recovery was plated on LB and kanamycin plates and placed at 37℃ overnight.

Removal of Plasmid Backbone

To remove the plasmid backbone, electrocompetent cells were made from the confirmed integration strain and transformed with the ΦC31 helper plasmid. Cells were recovered at 30℃ for 1 hour and plated on LB and carbenicillin plates. Colonies were picked into LB and grown at 37℃ for 8 hours. 5 microliters of the outgrowth were streaked on LB plates and grown overnight. Colonies were then patched on LB, LB and carbenicillin, and LB and kanamycin plates to screen for the knockout of the kanamycin resistance marker and loss of helper plasmids.

Methyltransferase functionality and expression

Methylome analysis was performed on *C. clariflavum* 4-2a using Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing (25)and Whole Genome Bisulfite Sequencing (WGBS) as previously described (21).

Methyltransferases (CloclaDRAFT _2368_2369, CloclaDRAFT _1830_1831, CloclaDRAFT _1994, CloclaDRAFT _1058, CloclaDRAFT _1051, CloclaDRAFT _1868_1869, CloclaDRAFT _2133, and CloclaDRAFT _1961) were codon optimized and synthesized by Genscript with an *attP* site. Each methyltransferase was then cloned into pLAR067, under control of pBad, using NEB builder. Each were integrated into AG4277 as described above yielding strain AG5645. Plasmid DNA of interest was transformed into AG5645 and cultures were grown with 1 mM arabinose to induce methyltransferase expression. DNA was isolated using ZymoPURE II plasmid midiprep kit (Zymo Research).

Transformation of C. clariflavum 4-2a

Two 5 mL cultures were inoculated with *C. clariflavum* 4-2a and grown overnight. The next day, two 200 mL cultures were inoculated with 1% inoculum and grown until an optical density (O.D.) of 0.9. Once the cultures were grown, they were centrifuged in 50 mL conical tubes at room temperature and

6,000xg for 15 minutes. The supernatant was decanted and 25 mL of electroporation buffer (250 mM sucrose, 10% glycerol) was added to the tube without disrupting the cell pellet. Cells were centrifuged again and washed twice more. After the last spin, the cell pellets were resuspended in ~100 µL electroporation buffer and transferred to a microcentrifuge tube.

Using fresh electrocompetent cells, 20 µL of cells were transformed with 1 µg of pMTL83151 (26). Cells were electroporated in a 1 mM cuvette with a square wave using a Bio-Rad Gene Pulser Xcell Electroporation System set at at 1200 V with a 1.5 mS pulse. Cells were then resuspended in 1 mL CTFUD medium and incubated for 3 hours at 50°C. After recovery, transformants were plated within CTFUD with 1.5% agar and 5 µg/mL thiamphenicol and incubated for 4 days at 50°C. Colonies from the plates were picked into CTFUD and screened by PCR for the presence of the *cat* gene.

Results

Design of the integration system

A heterologous DNA integration system using serine recombinases was created by first integrating a cassette of *attB* sites into the *E. coli* chromosome at the HK022 phage site, using the CRIM system. The cassette contains 14 *attB* sites: Φ370, ΦBT1, R4, BxB1, TP901-1, RV, Spβc, TG1, ΦC1, MR11, ΦK38, A118, Wβ, and BL3. A non-replicating integration vector was then constructed that contains the kanamycin resistance marker and the conditional oriR6K origin of replication, which can only replicate in an *E. coli* strain that expressed the *pir* gene. This *oriR6K-kanR* cassette was flanked by ΦC31 *attB* and *attP* sites, which will facilitate marker removal (see below). Initially, the plasmid was constructed with an inducible promoter driving lacZα outside of the ΦC31 sites, creating pGSs009 and pLAR067. This plasmid has NdeI and BamHI sites flanking lacZα to enable easy cloning into the plasmid. A similar plasmid was constructed, containing the corresponding *attP* sites to the poly-*attB* cassette, replacing lacZα (Figure 4.1A), creating pLAR031. Each of the corresponding large serine recombinases were cloned into a helper temperature sensitive plasmid, pInt-Ts (8), replacing the λ integrase. These genes are

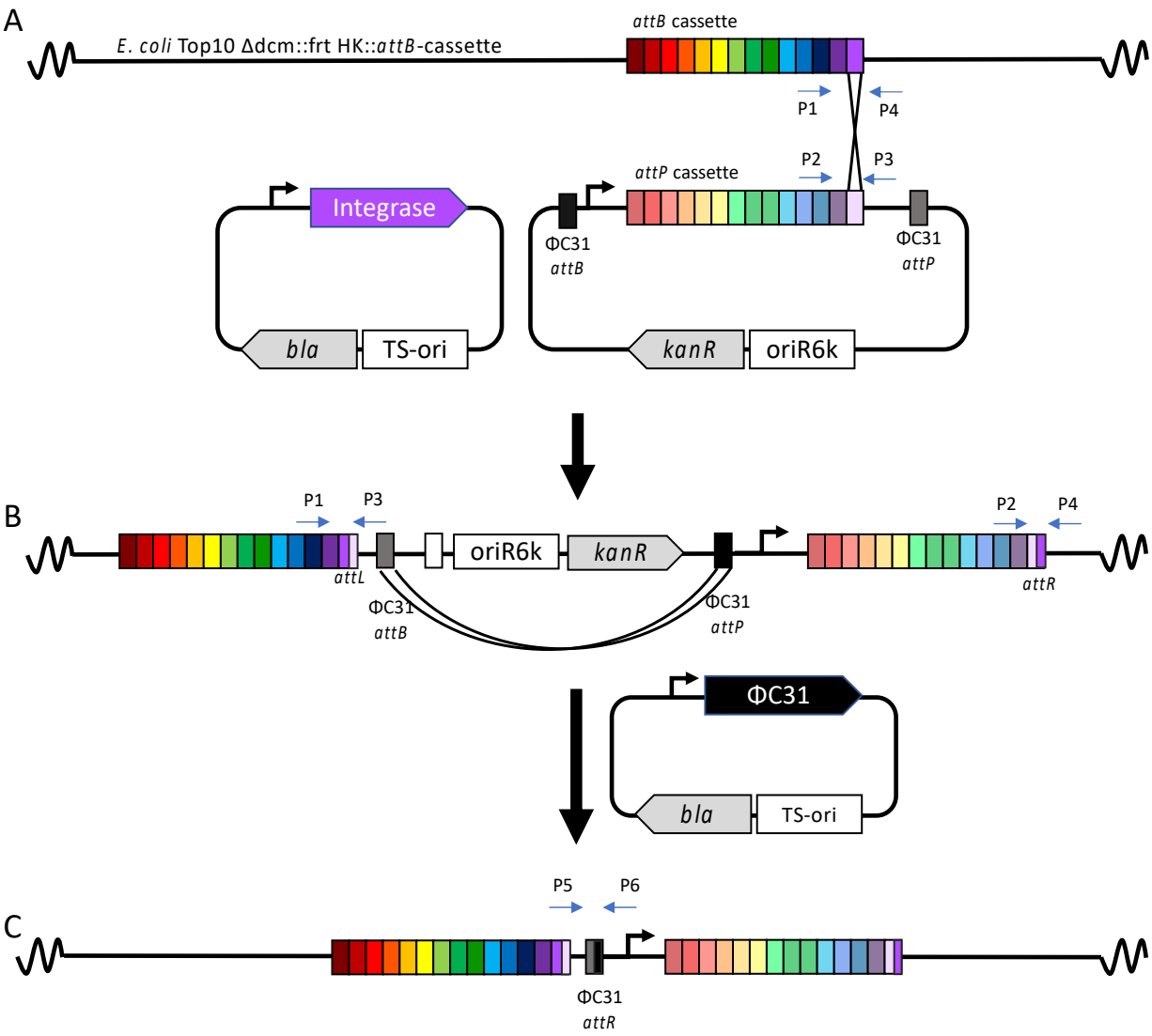


Figure 4.1: Site-specific recombinase system for DNA integration into the *E. coli* chromosome. A) 14 *attB* sites were added to chromosome at the HK phage site and the 14 corresponding serine recombinases were cloned onto temperature sensitive plasmids to mediate integration between a plasmid containing an *attP* site and the corresponding *attB* site on the chromosome. B) The kanamycin resistance marker and plasmid origin of replication can then be removed by introducing the ΦC31 recombinase on a temperature sensitive plasmid.

under the control of the λ pR promoter and cI857, a heat-inducible promoter system. This enables the recombinase to mediate integration and, after raising the temperature, facilitates the loss of the plasmid.

DNA integration occurs between an *attP* site from the non-replicating plasmid and the corresponding *attB* on the *E. coli* chromosome, which is mediated by the corresponding recombinase helper plasmid. The recombinase catalyzes a single recombination event between the *att* sites to integrate the entire non-replicating plasmid into the chromosome. Two new attachment sites are formed, *attL* and *attR,* flanking the newly integrated DNA. (Figure 4.1A). Next, the *oriR6K-kanR* cassette can be removed by introducing the ΦC31 helper plasmid. This results in recombination between the ΦC31 *att* sites, leaving behind a single ΦC31 *attR* site (Figure 4.1B).

Testing Integration Efficiency of the Recombinase helper Vectors

To test integration efficiency of the recombinases, the temperature sensitive helper plasmids were first transformed into an *E. coli* strain that contains the poly-*attB* cassette. A non-replicating plasmid was constructed with a poly-*attP* cassette, replacing lacZα in pGSs009. Then, competent cells were made of each recombinase and the poly-*attP* cassette plasmid was transformed in. Colonies were counted and colony forming units per microgram of DNA (CFU/μg) counts were reported (Figure 4.2). Integration was observed with twelve of the recombinases, of which six (A118, TG1, BL3, Wβ, TP901, and Spβ) had an efficiency of approximately 10^7 CFU/μg of DNA. Another four showed recombination frequencies in the range of 10^5 – 10^6 CFU/μg of DNA. ΦC1 enabled the lowest recombination frequency of an average of 200 CFU/μg of DNA, while two recombinases, ΦBT and RV, did not enable any integration.

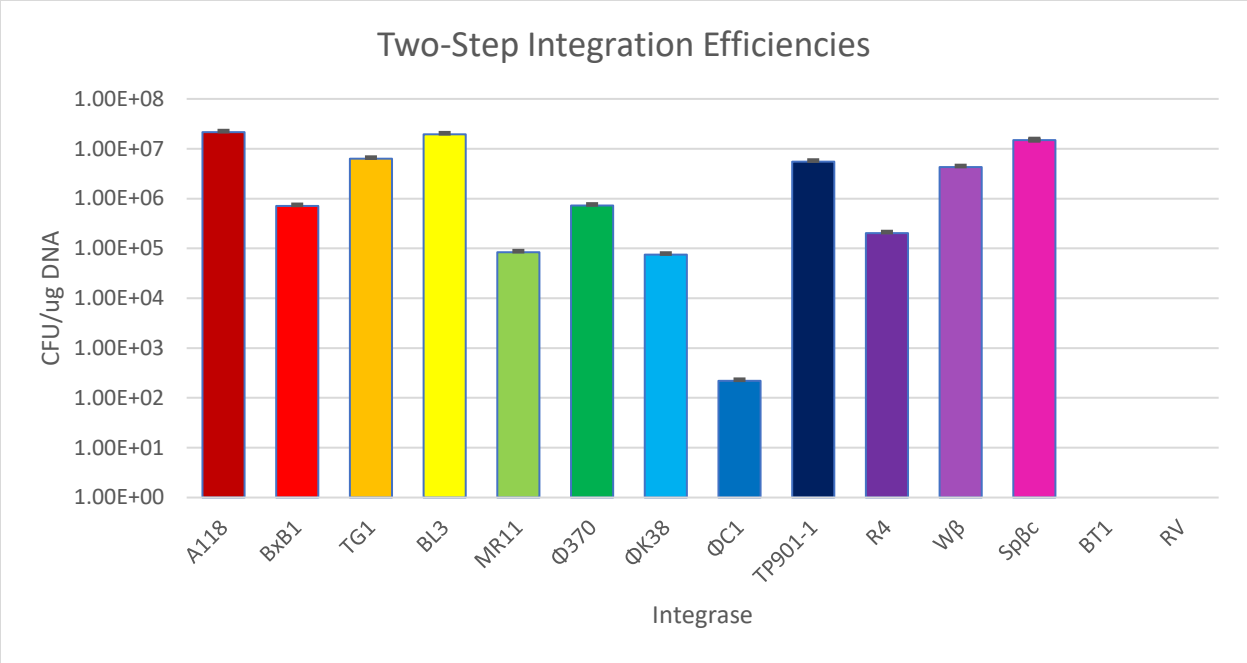


Figure 4.2: Integration efficiencies of each recombinase
Numbers reported in colony forming units per µg (CFU/µg) of suicide *attP* vector.

Integration Confirmation

Serine recombinases have been shown to integrate into naturally occurring secondary *attB* sites, called pseudo-*attB* sites, that have similarity to the native sequences (27). To determine if the non-replicating plasmid integrated into the intended loci, rather than a pseudo-*att* site, colonies from the resulting transformations were screened using a four-primer multiplexed PCR screen, with primers flanking the two *att* sites (P1, P2, P3, P4 indicated in Figure 4.1A), similar to the screen used for the CRIM system (8). The four-primer system gives distinct band sizes to distinguish between the *attB*, *attP*, *attL*, and *attR* sites to evaluate if integration occurred at the intended locus. One band forms for the suicide integration vector (P1 and P2) and one band forms for the parent strain (P3 and P4). After integration, two bands, of different size from the parents, form (P1 and P4, P2 and P3). If the poly-*attP* cassette plasmid integrated into a pseudo-*attB* site one band, indicative of the parent, will form. Ten colonies from each transformation were screened using this four-primer system to examine integration into the correct locus. All colonies showed the correct integration bands for all 10 colonies, except MR11 where 7 of 10 integrated correctly and ΦC1 where 4 of 10 integrated correctly.

Streamlined integration system via co-transformation

To determine if the process could be further streamlined to accelerate strain construction, we explored whether co-transformation of the recombinase helper plasmid and the integrating plasmid could allow for DNA insertion into the chromosome. In this scenario, the recombinase would be again be transiently expressed, but this time without selecting for the helper plasmid, and the resulting recombinase protein would recombine the *attP* plasmid into the chromosome. If successful, this would eliminate one round of competent cell preparation, transformation, and plasmid curing. When using the helper plasmids described above in combination with integrating poly-*attP* cassette plasmid at a non-permissive temperature, no kanamycin-resistant colonies were observed. Because expression of the recombinases could have been too low to enable DNA insertion, the recombinase helper plasmids were

reconstructed to use the *Clostridium thermocellum* enolase promoter (P_{Ct_eno}), which has been shown to be highly active in *E. coli* (24) and is presumably constitutively expressed. Therefore, transient expression of the recombinase should be sufficient to enable integration using the *Ct_eno* promoter. Of the twelve recombinases shown to be functional in Figure 4.2, ten were successfully cloned with the P_{Ct_eno} promoter, including all except Spβc and Wβ.

The new recombinase plasmids were co-transformed with the poly-*attP* cassette vector into the *E. coli* poly-*attB* cassette strain and recovered at a temperature too high to allow replication of the helper plasmid. The integration efficiencies of the more highly expressed recombinases were tested by transforming both the helper plasmid and the non-replicating *attP*-cassette plasmid simultaneously into the poly-*attB* cassette *E. coli* strain. All 10 enabled integration in the one-step integration process (Figure 4.3). Five of the recombinases had an insertional efficiency of 10^5 CFU/μg of DNA or higher: BxB1, TG1, BL3, Φ370, and R4. ΦC1 and TP901 recombinases also enabled DNA integration in a single step but had low efficiencies of less than 100 CFU/μg of DNA.

oriR6K-kanR Backbone Removal

Some applications will require the removal of the antibiotic resistance gene. Therefore, we tested the ΦC31 recombinase-based marker-removal system described above using a strain with the *attP*-cassette plasmid integrated into the R4 *attB* locus (Figure 4.1B,C). After integration into R4 was confirmed, the temperature-sensitive plasmid encoding the ΦC31 recombinase was transformed into the strain. ΦC31 enables recombination to remove the section of DNA between the ΦC31 *att* sites, removing the kanamycin resistance marker and oriR6K origin of replication, but leaving the genetic "cargo" on the chromosome. Once the ΦC31 recombinase was into transformed into the strain, colonies were picked and incubated at a 37°C to the cure the strain of the helper plasmid. Colonies were then screened for successful recombination between the ΦC31 *att* sites and loss of the helper plasmid by patching on both antibiotics used in the process. Colonies that did not grow on either antibiotic

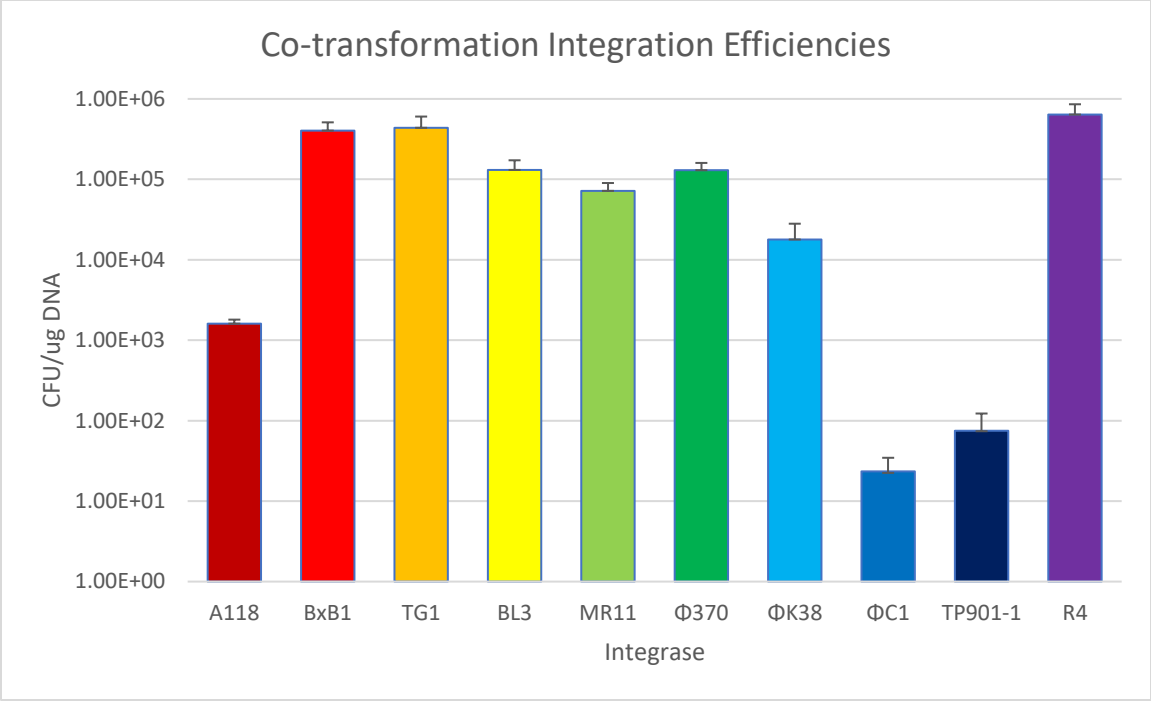


Figure 4.3: Integration efficiencies of each second-generation recombinase
Numbers reported in colony forming units per µg of suicide *attP* vector (CFU/µg of DNA).

were then PCR screened for removal of the backbone, using primers P5 and P6 (indicated on figure 4.1B) that flank the ΦC31 sites. After plating at 37 ˚C, 12 colonies were patched onto plates containing kanamycin and plates containing carbenicillin. Of these, 10 were sensitive to both antibiotics, and 100% of the sensitive strains were confirmed to have lost the *oriR6K-kanR* cassette by PCR.

Identification and expression of methyltransferases for C. clariflavum 4-2a

As a demonstration of the utility of this system, we targeted expression of *C. clariflavum* DNA methyltransferases to methylate plasmids prior to transformation. Methyltransferases in *C. clariflavum* 4-2a were identified using the NEB Restriction Enzyme Database (REBASE) and genome annotations (28). Eight methyltransferase gene clusters were identified: CloclaDRAFT_2368_2369, CloclaDRAFT _1830_1831, CloclaDRAFT _1994, CloclaDRAFT _1058, CloclaDRAFT 1051, CloclaDRAFT _1868_1869, CloclaDRAFT _2133, and CloclaDRAFT _1961. To determine which methyltransferases are functional, methylome analysis was performed using PacBio SMRT analysis to identify motifs containing 6-methyladenine (m^6A) and 4-methylcytosine (m^4C) and Whole Genome Bisulfite Sequencing (WGBS) to identify 5-methylcytosine (m^5C). Methylome analysis revealed eight m^6A motifs using SMRT analysis and no motifs with WGBS (Table 4.2).

To mimic the methylome of *C. clariflavum* to evade restriction, each identified methyltransferase was expressed from an arabinose-inducible promoter and inserted into the *E. coli* chromosome using the recombination system described above. The methyltransferases were cloned into the non-replicating integration vector with a single *attP* site. Each plasmid was then integrated into a *dam- dcm-* strain so that native *E. coli* methylation is not recognized by *C. clariflavum*. The resulting strain is AG5477 (Table 4.1) and methylome analysis was performed to determine the functionality of the strain. Six motifs were identified, missing ATGC**A**T and CAGA**A**G (Table 4.3).

Table 4.2: Methylome analysis of *C. clariflavum* 4-2a.
Methylated motifs are indicated with the methylated base capitalized, if T is methylated, the methylation is on the A of the complimentary strand. The type of methylation is indicated. %Mod indicates the percentage of motifs modified vs total motifs identified in the genome.

Motif	Type	% Mod
G**A**TC	m^6A	99.6
GAC**A**T	m^6A	96.4
TGAA**A**G	m^6A	99.7
CAGA**A**G	m^6A	99.7
GCG**A**TD	m^6A	96.3
CN**A**YNNNNC**T**C	m^6A	98
ATGC**A**T	m^6A	97.7
G**A**GNNNNNNNR**T**C	m^6A	99.6

Table 4.3: Methylome analysis of *C. clariflavum* 4-2a *E. coli* methylation strain.
Methylated motifs are indicated with the methylated base capitalized, if T is methylated, the methylation is on the A of the complimentary strand. The type of methylation is indicated. %Mod indicates the percentage of motifs modified vs total motifs identified in the genome.

Motif	Type	% Mod
G**A**TC	m^6A	99.9
GCG**A**TD	m^6A	99.9
GAC**A**T	m^6A	99.9
G**A**GNNNNNNNR**T**C	m^6A	99.7
CN**A**YNNNNC**T**C	m^6A	74.3
TGAA**A**G	m^6A	50.5

Transformation of C. clariflavum 4-2a

Even though methylation in *E. coli* was incomplete, we attempted to transform *C. clariflavum*. To best evade the RM systems in *C. clariflavum* and enable transformation, plasmid pMTL83151 was isolated out of the *E. coli* methylation strain AG5477. As a control, unmethylated plasmid was isolated out of Top10 *Δdcm::frt* to determine the impact of DNA methylation on transformation. Two batches of *C. clariflavum* competent cells were transformed with plasmid from each methylation state, in duplicate. Unmethylated plasmid never yield colonies, while the methylated plasmid yielded 9±3 colony forming units per microgram of DNA (CFU/μg).

Discussion

This DNA integration system expands the *E. coli* genetic toolbox to provide a method to simply, quickly, sequentially, and stably integrate up to twelve heterologous DNA constructs into the chromosome of *E. coli*. Integration can occur in a single co-transformation of two plasmids, including one that is common for all insertions into that site and one that can be customized to carry the desired cargo. This site-specific recombination step is quicker than strain construction using classic techniques like homologous recombination and does not suffer from spontaneous resistant mutants than can hinder *sacB* and other counter-selection systems. Our system enables stable integration in two days, while tools such as homologous recombination often requires several steps and at least five days. Another characteristic of the system is the ability of the recombinases to recombine DNA efficiently and reliably. They function without a significant decrease in efficiency as the size of the insert increases. Constructs containing multiple methyltransferase for *Fibrobacter succinogenes* with sizes of upwards of 7kb have been integrated and no decrease in recombination efficiency was seen (Personal communication, IC Payne). Size-based limitations on DNA insertion should be primarily dependent on the diminishing efficiency of DNA entry into the cell as the DNA size increases. This overcomes some of the size-associated challenges with other genetic tools, like use of the lambda Red recombinase system

(3) for heterologous DNA insertion. The ability to rapidly integrate constructs, at single copy, will help to increase the speed of engineering in *E. coli*.

Initially the first set of recombinase helper plasmids were constructed that allowed for integration in a two-step process. Using this approach, twelve of the recombinases were functional and two were not. There is no indication why BT and RV did not function as expected. To decrease the number of steps in this process, the ability to transform both plasmids in one event was desirable. The second set of helper plasmids enabled a one-step integration system, though two of the recombinases were not successfully cloned. We speculate that high expression of the Wβ and Spβ recombinases was toxic for *E. coli*. However, the availability of the first set of plasmids still allows for the use of the Wβ and Spβ recombinases. Of the recombinases that are functional in the one-step process, R4, BxB1, and TG1 are highly efficient and reliable, so these three insertion sites should be targeted first when using this integration system. The ΦC1 and MR11 recombinases are the least reliable, integrating into pseudo-*att* sites and with lower efficiencies; therefore, these should be the last two sites to be considered. There is a decrease of integration efficiency with the strong promoter helper plasmids, but the speed of the one-step process makes them more useful than the first set when inserting several constructs. The two-step approach will be most useful when high integration efficiency is needed, such as for library creation.

The DNA integration system described here contains many useful features. Like the CRIM system (8), it uses the conditional oriR6K origin of replication, which is only able to replicate in *E. coli* strains engineered to express the *pir* gene. This enables the integration vectors to act as a suicide vector in most *E. coli* strains, including the one containing the poly-*attB* cassette. Next, the vector backbone can be removed after DNA integration, because the oriR6K origin and the antibiotic resistance marker are flanked by *att* sites for ΦC31. Removal of the backbone allows for reuse of the antibiotic resistance marker and reduces repetitive DNA, diminishing the likelihood of homologous recombination events when stacking multiple genes into the poly-*attB* cassette. Another feature of this system is the

unidirectional nature of the serine recombinases. Integration creates new attachment sites, *attL* and *attR*, and these new *att* sites are not recognized by the recombinases. This is especially beneficial when integrating multiple constructs in close proximity on the genome. Unlike the Flp-*frt* system, where the resulting scars are still a substrate for the Frt recombinase, ΦC31 does not act on the remaining ΦC31 *attR* sites after the marker is removed. Because these resulting scars cannot recombine with each other, there are no off-target effects if two scars are near each other.

C. clariflavum 4-2a is an organism of interest for biofuels research but had not yet been genetically transformed, in part due to the large number of RM systems within the strain. Therefore, we used this integration system to rapidly insert eight methyltransferases from *C. clariflavum* into the *E. coli* chromosome. The resulting *E. coli* methylation strain was subsequently used methylate plasmid DNA and successfully transform *C. clariflavum* for the first time. Transformation efficiency in this strain is still low, and further optimization will likely increase the efficiency. Two of the motifs found in *C. clariflavum* were not found in the *E. coli* methylation strain, and two other motifs were not modified even close to 100% in the genome, which likely impacts the transformation efficiency. *C. clariflavum* is a thermophilic organism, the activity of the methyltransferases at *E. coli* growth temperatures could decrease functionality. This could explain why the TGAA**A**G motif was only modified in 50% of the motifs in the genome. Utilization of mesophilic methyltransferases that target the same motif or engineered versions of the thermophilic enzymes to increase activity at mesophilic temperatures could help to overcome this problem. Even with the incomplete methylation a low transformation efficiency was enabled which allows for further genetic manipulation of *C. clariflavum* for fundamental studies and metabolic engineering.

References

1. Rosano GL, Ceccarelli EA. Recombinant protein expression in Escherichia coli: advances and challenges. Frontiers in Microbiology. 2014;5:172.
2. Chen X, Zhou L, Tian K, Kumar A, Singh S, Prior BA, et al. Metabolic engineering of Escherichia coli: A sustainable industrial platform for bio-based chemical production. Biotechnology Advances. 2013;31(8):1200-23.
3. Pyne ME, Moo-Young M, Chung DA, Chou CP. Coupling the CRISPR/Cas9 System with Lambda Red Recombineering Enables Simplified Chromosomal Gene Replacement in Escherichia coli. Appl Environ Microbiol. 2015;81(15):5103-14.
4. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular Systems Biology. 2006;2(1).
5. Ao X, Yao Y, Li T, Yang T-T, Dong X, Zheng Z-T, et al. A Multiplex Genome Editing Method for Escherichia coli Based on CRISPR-Cas12a. Frontiers in microbiology. 2018;9:2307-.
6. Elmore JR, Furches A, Wolff GN, Gorday K, Guss AM. Development of a high efficiency integration system and promoter library for rapid modification of Pseudomonas putida KT2440. Metabolic Engineering Communications. 2017;5:1-8.
7. Brown WRA, Lee NCO, Xu Z, Smith MCM. Serine recombinases as tools for genome engineering. Methods. 2011;53(4):372-9.
8. Haldimann A, Wanner BL. Conditional-Replication, Integration, Excision, and Retrieval Plasmid-Host Systems for Gene Structure-Function Studies of Bacteria. Journal of Bacteriology. 2001;183(21):6384-93.
9. Kilby NJ, Snaith MR, Murray JAH. Site-specific recombinases: tools for genome engineering. Trends in Genetics. 1993;9(12):413-21.
10. Sauer B. Inducible Gene Targeting in Mice Using the Cre/loxSystem. Methods. 1998;14(4):381-92.
11. Groth AC, Calos MP. Phage integrases: biology and applications. Journal of molecular biology. 2004;335(3):667-78.
12. Landy A. The lambda Integrase Site-specific Recombination Pathway. Microbiol Spectr. 2015;3(2):27.
13. Thyagarajan B, Olivares EC, Hollis RP, Ginsburg DS, Calos MP. Site-Specific Genomic Integration in Mammalian Cells Mediated by Phage ϕC31 Integrase. Molecular and Cellular Biology. 2001;21(12):3926-34.
14. Xu Z, Brown WRA. Comparison and optimization of ten phage encoded serine integrases for genome engineering in Saccharomyces cerevisiae. BMC Biotechnology. 2016;16(1):13.
15. Guss AM, Rother M, Zhang JK, Kulkarni G, Metcalf WW. New methods for tightly regulated gene expression and highly efficient chromosomal integration of cloned genes for Methanosarcina species. Archaea (Vancouver, BC). 2008;2(3):193-203.
16. Siuti P, Yazbek J, Lu TK. Synthetic circuits integrating logic and memory in living cells. Nat Biotechnol. 2013;31(5):448-52.
17. Hsiao V, Hori Y, Rothemund PW, Murray RM. A population-based temporal logic gate for timing and recording chemical events. Mol Syst Biol. 2016;12(5):869.
18. Mermelstein LD, Welker NE, Bennett GN, Papoutsakis ET. Expression of Cloned Homologous Fermentative Genes in Clostridium Acetobutylicum ATCC 824. Bio/Technology. 1992;10(2):190-5.

19. Purdy D, O'Keeffe TA, Elmore M, Herbert M, McLeod A, Bokori-Brown M, et al. Conjugative transfer of clostridial shuttle vectors from Escherichia coli to Clostridium difficile through circumvention of the restriction barrier. Mol Microbiol. 2002;46(2):439-52.
20. Yasui K, Kano Y, Tanaka K, Watanabe K, Shimizu-Kadota M, Yoshikawa H, et al. Improvement of bacterial transformation efficiency using plasmid artificial modification. Nucleic Acids Res. 2009;37(1):e3.
21. Riley LA, Ji L, Schmitz RJ, Westpheling J, Guss AM. Rational development of transformation in Clostridium thermocellum ATCC 27405 via complete methylome analysis and evasion of native restriction-modification systems. J Ind Microbiol Biotechnol. 2019.
22. Tock MR, Dryden DT. The biology of restriction and anti-restriction. Curr Opin Microbiol. 2005;8(4):466-72.
23. Izquierdo JA, Pattathil S, Guseva A, Hahn MG, Lynd LR. Comparative analysis of the ability of Clostridium clariflavum strains and Clostridium thermocellum to utilize hemicellulose and unpretreated plant material. Biotechnology for biofuels. 2014;7(1):136-.
24. Olson DG, Maloney M, Lanahan AA, Hon S, Hauser LJ, Lynd LR. Identifying promoters for gene expression in Clostridium thermocellum. Metabolic Engineering Communications. 2015;2:23-9.
25. Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-89.
26. Heap JT, Pennington OJ, Cartman ST, Minton NP. A modular system for Clostridium shuttle plasmids. Journal of microbiological methods. 2009;78(1):79-85.
27. Combes P, Till R, Bee S, Smith MC. The streptomyces genome contains multiple pseudo-attB sites for the (phi)C31-encoded site-specific recombination system. J Bacteriol. 2002;184(20):5746-52.
28. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Research. 2015;43(D1):D298-D9.

Chapter V

Developing *Megasphaera elsdenii* into a Platform Organism for Chemical Production from Lactate

This chapter is part of a manuscript to be submitted by Lauren A. Riley, Janet Westpheling, and Adam M. Guss:

Riley, L. A., Westpheling, J., Guss, A. M., (2020) Developing *Megasphaera elsdenii* into a Platform Organism for Chemical Production from Lactate. In Preparation.

The experimental work was performed by Lauren A. Riley. The manuscript was written and revised by Lauren A. Riley, Janet Westpheling, and Adam M. Guss

Abstract

Non-model bacteria will likely be essential for the future of renewable fuel and chemical bioprocessing, but a lack of methods for genetic transformation greatly hinder efforts to develop new microbial chassis organisms. *Megasphaera elsdenii i*s of particular interest due to its ability to efficiently produce medium chain carboxylic acids anaerobically from lactate and glucose feedstocks. To develop *M. elsdenii* into a new platform organism for chemical production, transformation of two strains, ATCC 25940 and NCIMB 702410, was enabled by overcoming their native Restriction-Modification (RM) systems. Methylome analysis was performed on both strains to identify active RM systems and the corresponding methyltransferases were expressed in *E. coli*. Next, the ATCC 25940 strain was engineered for n-butanol production through the heterologous expression of *Clostridium acetobutylicum's adhE2* after proper methylation. Using lactate as the carbon source, 5.33 mM of butanol was produced, representing 7.4% of the detected fermentation products. The *adhE2* was then integrated into the *upp* locus of the ATCC 25940 genome, enabling the use of *upp* as counter-selectable marker in future genetic modifications. This work demonstrates both transformation and metabolic engineering of *M. elsdenii* for the first time, enabling genetic engineering and novel bioprocess development possibilities for then production of fuels and chemicals in this strain.

Introduction

Petroleum based fuels and chemicals are commonly used in a wide variety of products, from gasoline to plastics to fertilizer. Due to increasing concerns with fossil fuels, including greenhouse gas emissions and finite resources, alternative routes for fuel and chemical production are being explored (1). Microorganisms offer a promising alternative, where sustainable feedstocks can be converted into a wide variety of chemicals of interest for bioprocessing. Metabolic engineering of microorganisms is an important area of research for this process where beneficial pathways can be engineered to enable increased carbon funneling to the product of interest. Currently though, only a few model microorganisms are the focus for most genetic engineering for fuel and chemical production. This is due in large part to the ease of genetic modification and the base of knowledge associated with model organisms such as *E. coli*. Non-model organisms are of interest though due to their complex physiological traits that are difficult to transfer to model organisms, but the lack of genetic tools presents a major barrier to industrial application. The development of new synthetic biology chassis organisms could enable novel approaches to bioprocessing.

Non-model bacteria often have complex phenotypes of commercial interest which range from tolerance to bioprocessing conditions, like growth at low pH, to the ability to grow on non-traditional substrates. Significant research has gone into a wide variety of feedstocks, from model feedstocks such as glucose to real world substrates such as lignocellulosic biomass and syngas (2, 3). One feedstock, lactate, holds potential but is currently underutilized. Lactate is routinely produced by lactic acid bacteria at high concentrations and yields and it is able to be produced from a wide variety of feedstocks, including those that don't compete with food sources, such as lignocellulosic hydrolysates (4). Currently, a major end use of lactate is in the food industry, pharmaceuticals (5) and the production of biodegradable polymer polylactic-acid (PLA) (6). Lactate can be biologically upgraded though by microorganisms and turned into useful products such as medium chain length carboxylic acids, which

can act as a platform for the synthesis of a wide variety of commercially important chemicals. While carboxylic acids are produced by many species of bacteria as fermentation products, only a few microorganisms have been demonstrated to produce medium chain length carboxylic acids from lactate, including mixed culture reactor systems (7), *Ruminococcaceae* bacterium CPB6 (8), and *Megasphaera elsdenii*. None of these biocatalysts have been genetically modified, which is an important step to developing a platform organism for bioproduction.

M. elsdenii belongs to a group of ruminal and intestinal lactate- and glucose-fermenting bacteria called Negativicutes. *M. elsdenii* produces carboxylic acids (C3-C8) as fermentation products when growing on lactate and glucose, including formation of propionic, butyric, valeric, hexanoic, and in some cases octanoic acids as major fermentation products (9). Propionic acid is produced via the acrylate pathway (10) and the longer chain acids are likely produced via a chain elongation pathway. Butyryl-CoA is converted to butyrate by a Coenzyme A- transferase, which utilizes acetate as a Coenzyme A acceptor (Figure 5.1) (11). In the absence of genetic tools, research on *M. elsdenii* to date has primarily focused on fermentation studies to understand lactate and sugar catabolism in *M. elsdenii* under batch and steady-state fermentation conditions. Under batch conditions, *M. elsdenii* converts lactate to acetate, propionate, valerate, and butyrate (12). Sugar conversion has also been examined and recently an acid titer of > 5 g/L was achieved from pure glucose and from lignocellulosic sugars in serum bottles. Using pertractive, fed-batch fermentation, organic acid titers reached 57 g/L (13), demonstrating that *M. elsdenii* natively has high flux through this chain elongation pathway, making it a promising host for the high-level production of industrially relevant products derived from this pathway if genetic tools can be developed.

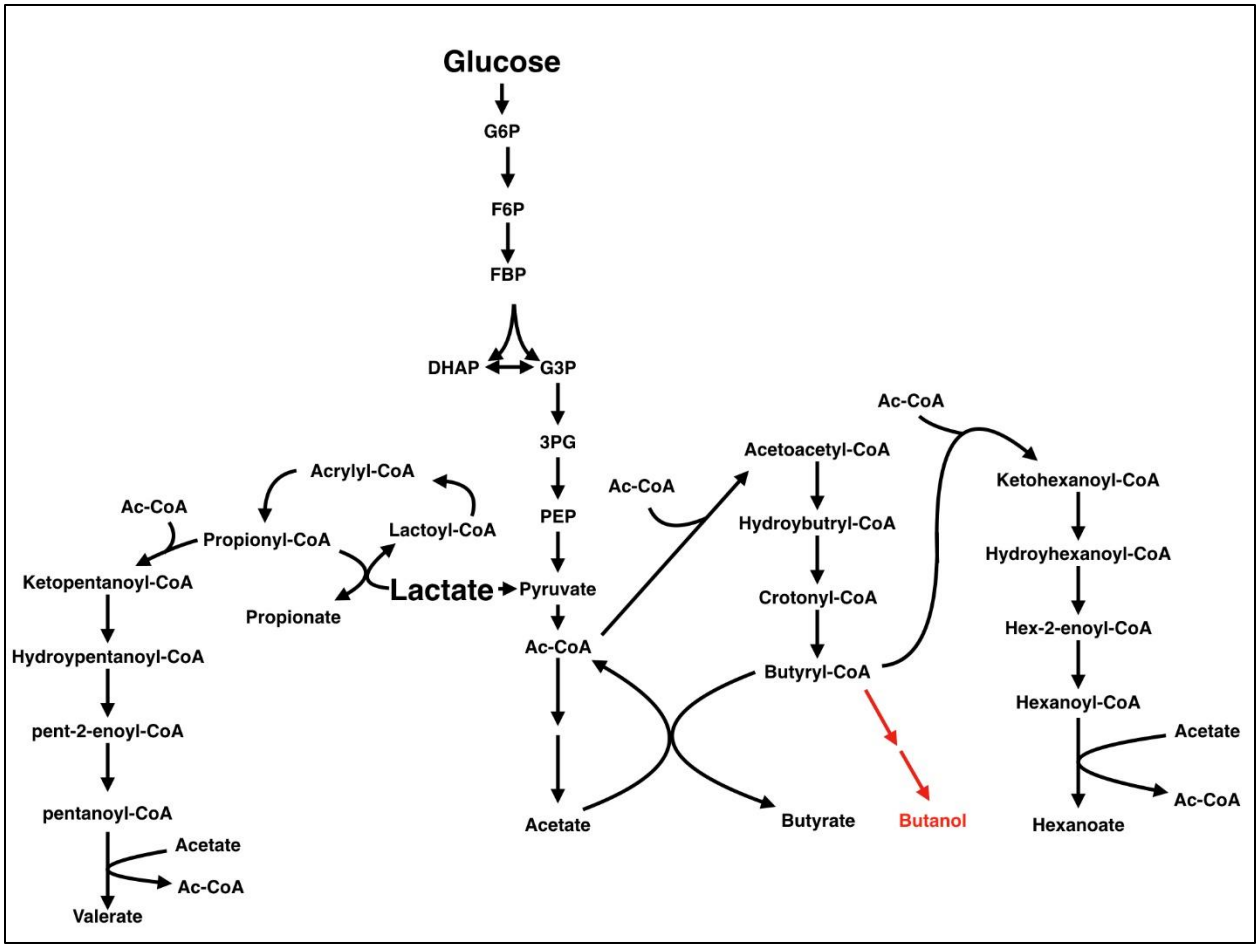


Figure 5.1: Central metabolism for *M. elsdenii* from lactate and glucose.

Propionate is produced via the acrylate pathway. The other acids are produced via chain elongation from both lactate and glucose. The planned addition of n-butanol production is shown by the red pathway.

Recent advances in genetic tool development for non-model organisms allows for rational development of transformation methods. To enable transformation, the native host defense systems, in particular restriction-modification (RM) systems, need to be evaded. RM systems act as an immune system within bacteria to degrade DNA that is methylated differently than the host (14). Typically, RM systems are comprised of a restriction enzyme that targets and cuts a specific unmethylated DNA motif while the corresponding methyltransferase methylates at the same motif in the hosts genome to protect against the restriction enzyme. There are four types of RM systems. Type I RM systems are comprised of a restriction enzyme, DNA specificity subunit, and methyltransferase that target two recognition motifs, separated by a series of random bases, like EcoKI – A**A**C-NNNNNN-G**T**GC (15). Type II and III are comprised of restriction enzyme and methyltransferase and target shorter motifs, like EcoRI-GA**A**TTC Finally, Type IV systems are comprised of a restriction enzyme that targets motifs that are methylated differently than the host (16). The importance of overcoming these systems to enable transformation has been demonstrated in many different species (17-20). Here we evade the RM systems in two strains of *M. elsdenii*, ATCC 25940 and NCIMB 702410 to enable transformation and the development of genetic tools. These tools are leveraged for metabolic engineering of *M. elsdenii* for the production of n-butanol from lactate.

Materials and Methods

Growth and Media Conditions

M. elsdenii ATCC 25940 and NCIMB 702410 (21) were grown in Reinforced Clostridial Media (RCM) either with (BD Difco) or without (HiMedia) 0.5 g/L agar as indicated. A modified RCM was also used for fermentations, containing per L: either 5 g glucose or 5 g sodium D-Lactate (Sigma-Aldrich), 10 g peptone, 10 g beef extract, 3 g yeast extract, 5 g sodium chloride, 0.5 g cysteine HCl, 3 g sodium acetate. Each type of RCM was supplemented with 0.5 ml/L 0.2% (w/v) resazurin as a redox indicator. Growth

occurred at 37°C in a Coy vinyl anaerobic chamber with 85% N_2, 10% CO_2, and 5% H_2 mixed gas atmosphere. Media was supplemented with 5 μg/mL thiamphenicol for plasmid selection.

Methylome Analysis

M. elsdenii genomic DNA was prepared using the Qiagen genomic tip kit according the manufacturer's instructions. Genomic DNA was sent to the Joint Genome Institute (JGI) for sequencing using Single Molecule Real-Time (SMRT) sequencing on the Pacific Biosciences (PacBio) platform (21). Methylated motifs were determined using the SMRT Analysis software and Expression Analysis (22). For Whole Genome Bisulfite Sequencing (WGBS) methylC-seq libraries were formed and sequenced via Illumina, methylated motifs were determined as previously described (17).

Methyltransferase Expression in E. coli

Each of the methyltransferases were cloned into pLAR067 (Chapter 3), under control of p_{Bad}, using Gibson assembly. For *M. elsdenii* ATCC 25940 each of the Type I methyltransferases and DNA specificity subunits were codon optimized and synthesized by Genscript, while the NCIMB 702410 Type I system was PCR amplified from the genome. Mels_0050-51 was inserted into pLAR067 with an R4 *attP* site following it, *mels_1615-16* was inserted without an *attP* site, and C6Y28_02465-02475 was inserted with an BxB1 *attP* site. Mels_1615-1616 was integrated into the *E. coli* WM3118 Δ*dcm::ft* chromosome using the serine phage integration system (Chapter 3) and mels_0050-51 was integrated next into the λ *attB* using the conditional-replication, integration, and modular (CRIM) system (23). C6Y28_02465-02475 was integrated into *E. coli* BW25113 Δ*dcm::frt* Δ*dam::frt* chromosome using the serine phage integration system(Chapter 3).

Plasmids of interest for transformation of *M. elsdenii* include the pMTL8000 series, pMTL85141, pMTL84151, pMTL83151, and pMTL82151 (24) and plasmids containing pNW33N (17), pBC1, and pVJL1. To construct the pBC1-cat (25) and pVJL1-cat (26) plasmids, each origin was synthesized by JGI into

pNJ022 (27) and the pNW33N origin was replaced with *cat* using Gibson assembly. Each plasmid was transformed into each of the *E. coli* methylation strains and cultures of each plasmid were grown with 1mM arabinose for methyltransferase induction. Each of the plasmids were prepped using ZymoPure II plasmid midiprep kit.

Transformation of M. elsdenii

Overnight cultures of *M. elsdenii* were grown in 5 mL RCM (BD Difco). Next, 50 µL was subcultured into 500 mL RCM (HiMedia), in duplicate, and grown overnight at 37°C to stationary phase. All competent cell preparation was done at room temperature. To make competent cells, the culture was centrifuged at 5,000 x g in 500 mL centrifuge bottles for 10 minutes. Supernatant was decanted and cells were washed by resuspending the cell pellet in 250 mL electroporation buffer (250 mM sucrose, 10% glucose). Cells were centrifuged and washed twice more. After the last wash, the electroporation buffer was completely removed. Cells were resuspended in enough electroporation buffer to enable pipetting (~500 uL) and transferred to a microcentrifuge tube. 20 µL of fresh electrocompetent cells were transformed with 1 µg of DNA, in duplicate. A 1 mM cuvette was used and shocked with a square wave at 1200 v and 1.5 ms using a Bio-Rad GenePulser. After electroporation, cells were recovered in 1 mL RCM (BD Difco) and incubated for 3 hours. Cells were then plated in molten RCM + 1.5% agar, and once the agar solidified, they were incubated for 2-3 days at 37°C in sealed boxes in an anaerobic chamber. Colonies were screened by PCR to verify the presence of the plasmid backbone.

Construction of Butanol Expression Vectors

Butanol expression vectors were constructed using Gibson assembly according to the manufacturer's instructions. The *Clostridium acetobutylicum adhE2* was inserted into the MCS of pMTL85141 under control of three different promoters: pLAR131, *E. coli* promoter P_{tac}; pLAR132, P_{tac} with the *M. elsdenii* 16s rRNA subunit UP element; and pLAR133, the 300 bp upstream of mels_747

(P_{Mels_747}). Finally, for pLAR135, *adhE2* was inserted with a Ribosomal Binding Site (RBS) plus 7 bases following *cat* in pMTL85141. Each plasmid was transformed into the appropriate *E. coli* methylation strains, methyltransferase expression was induced with 1 mM arabinose, and plasmid was isolated and transformed into each strain of *M. elsdenii*.

Integration of adhE2 into the chromosome

A plasmid to integrate *adhE2* under control of the P_{mels_747} into the *upp* locus was constructed by Gibson assembly. 1 kb homology arms upstream and downstream of *upp* were cloned flanking *adhE2* into the MCS of pMTL85141. A plasmid for deletion of *upp* without an insert was also created without *adhE2* between the homolog arms. All *E. coli* growth for this cloning was performed at room temperature to minimize plasmid toxicity and enable correct construction of the homology arms. The plasmids were transformed into *M. elsdenii* ATCC 25940 and colonies were subsequently picked into RCM (BD Difco) + 5 µg/mL TM. The liquid cultures were passaged into RCM (BD Difco). Next, the culture was serial diluted in X-compartmentalized petri plates with RCM + 1.5% agar + 20 µg/mL 5-fluorouricil. The plates were incubated overnight, and colonies were streaked on RCM plates. Single colonies were picked into RCM (BD Difco) and PCR screened for insertion into *upp*.

Fermentation and Product Quantification

M. elsdenii ATCC 25940 and NCIMB 702410 wild type, each strain containing the butanol expression vectors, *M. elsdenii* ATCC 25940 Δ*upp, M. elsdenii* ATCC 25940 Δ*upp::adhE2* were grown in 5 mL RCM (HiMedia) with thiamphenicol, if necessary, overnight. 50 µL of each strain was added to Balch tubes containing 10 mL of modified RCM with lactate and glucose. Each strain was cultured in quadruplicate for 96 hours. Samples were taken at 24-hour intervals and fermentation products were quantified using High Performance Liquid Chromatography (HPLC). Lactate, glucose, acetate, butyrate, valerate, propionate, and butanol were quantified on Agilent 1260 infinity series with the Aminex-HPX-

87H column (Bio-Rad). The HPLC was run with 5 mM sulfuric acid mobile phase at 0.6 mL/min with the column at 65°C. The volume loading was 20 µL.

Results

Methylome Analysis of M. elsdenii ATCC 25940 and NCIMB 702410

To enable transformation and overcome the native RM systems in both strains of *M. elsdenii,* the potential RM systems were first identified using New England Biolabs' Restriction Enzyme Database, REBASE (28). *M. elsdenii* ATCC 25940 putatively encodes two Type I systems (Mels_0050-0052, Mels_1615-1617), two Type II (Mels_0974-0976, Mels_1947), and two Type IV systems (mels_1373, mels_2001-2002). *M. elsdenii* NCIMB 702410 is predicted to encode one Type I system (C6Y28_02465-02475), two Type II systems (C6Y28_01935-01940, C6Y28_04345), and two Type IV systems (C6Y28_00750, C6Y28_09510-09515). Methylome analysis was then performed to determine which of the methylation subunits are active in each strain. Methylated motifs were identified using Single Molecule Real-Time sequencing (SMRT) on the PacBio platform (22) and Whole Genome Bisulfite Sequencing (WGBS) using Illumina (29). SMRT analysis revealed two Type I motifs and G**A**TC in *M. elsdenii* ATCC 25940 (where the bolded **A** represents 6-methyladenine) and one Type I motif in NCIMB 702410. WGBS did not reveal any m^5C motifs in either strain (Table 5.1). According to the NEB Restriction Enzyme Database (REBASE)(28) the Type II, Mels_0974-0976, is predicted to act on G**A**TC, suggesting that it is the active methyltransferase while the other Type II systems are not active.

Expression of Methyltransferases in E. coli

To mimic the *M. elsdenii* methylome, the corresponding methyltransferases were expressed from the *E. coli* chromosome. Expression of methyltransferases in *E. coli* enables proper methylation of plasmids prior to transformation in *M. elsdenii*. For ATCC 25940 both Type I methyltransferases and DNA specificity subunits (Mels_0050-0051 and mels_1615-1616) were integrated into the *E. coli*

Table 5.1: Methylome analysis and transformation efficiencies of *M. elsdenii* ATCC 25940 and NCIMB 70241.

The methylated base is bold in each motif. Transformation efficiencies are reported in colony forming units per microgram of DNA (CFU/µg) for each plasmid successfully transformed. Plasmids were isolated out of the best methylation strain, no colonies formed using wild type *E. coli*

Strain	**Methylated motifs**	**Transformation Efficiency**		
		pMTL85141	pBC1-cat	pVJL1-cat
ATCC 25940	G**A**TC G**A**GNNNNNNGA**T** C**A**GNNNNNN**T**RTC	8350±911	2020±813	3620±1236
NCIMB 702410	CG**A**NNNNNNNN**T**RTC	1360±283	170±33	343±148

chromosome. An *E. coli dam*+ *dcm*- strain was utilized so that Dam methylates G**A**TC and Dcm methylation won't interfere with the Type IV restriction enzymes in the strain. For NCIMB 702410, the single Type I methyltransferase and DNA specificity subunits were integrated into an *E. coli dam- dcm*- strain because neither of the *E. coli* motifs were seen in the methylome data for this strain. To validate functional expression of the methyltransferase, methylome analysis via SMRT sequencing was performed on each of the *E. coli* methylation strains. Each strain correctly mimics their corresponding *M. elsdenii* strain, proving all the methyltransferases were functionally expressed (Table 5.2).

Antibiotic Minimum Inhibitory Concentrations

To determine which antibiotic resistance genes could be used for selection, the minimum inhibitory concentration (MIC) of nine antibiotics were determined. Each strain of *M. elsdenii* was grown with 8 different concentrations of antibiotic, ranging from 1000-0.46 μg/mL, on a plate reader (Table 5.3). The numbers reported are the highest MIC of three runs. Thiamphenicol and tetracycline required the lowest concentration of antibiotic for growth inhibition, at 4.1 μg/mL and 1.4 μg/mL respectively. Based on these results, thiamphenicol was used for selection of transformants.

Successful Transformation of M. elsdenii ATCC 25940 and NCIMB 702410

To enable transformation for the first time in this species, multiple plasmids with different origins of replication were tested including pBP1, pCB102, pCD6, and pIM13 from the pMTL8000 series (24) and others, pNW33N (17), pBC1 (25), and pVJL1 (26). The pMTL8000 series, pNW33N, and pBC1 are known *Clostridium* and *Bacillus* spp. origins of replication while pVJL1 comes from *Veillonella* spp. which is the phylogenetically closest genus to *Megasphaera* that is genetically tractable. Each of these plasmids were isolated out of wild type *E. coli* (*dam+dcm*-) and the best methylation strain for each strain of *M. elsdenii*, and transformation was performed using electroporation. No transformation was

Table 5.2: Methylome analysis of the *E. coli* strains expressing *M. elsdenii* methyltransferases.

The methylated base is bold in each motif.

E. coli Strain	Methylome
ATCC 25940	G**A**TC G**A**GNNNNNNGA**T** C**A**GNNNNNN**T**RTC
NCIMB 702410	CG**A**NNNNNNNN**T**RTC

Table 5.3: Antibiotic MIC for *M. elsdenii*

A) ATCC 25940 and B) NCIMB 702410. Numbers reported in µg/mL. The highest MIC of three runs is reported.

Antibiotic	*M. elsdenii* ATCC 27540	*M. elsdenii* NCIMB 702410
Gentamycin	110	37
Chloramphenicol	37	12
Thiamphenicol	4.1	4.1
Apramycin	110	110
Kanamycin	110	330
Carbenicillin	110	110
Streptomycin and Spectinomycin	110	110
Tetracycline	1.4	1.4
Erythromycin	110	37

observed with any plasmid isolated out of wild type *E. coli*. The pMTL85141 (pIM13 origin), pVJL1-cat, and pBC1-cat plasmids were successfully transformed into each strain of *M. elsdenii* when properly methylated to mimic the methylome. The highest transformation efficiency used pMTL85141, with 8350 colony forming units per microgram of DNA (CFU/µg) for ATCC 25940 and 1360 CFU/µg for NCIMB 702410 (Table 5.1). Transformation was less efficient but still reproducible with pBC1-cat and pVJL1-cat.

Engineering M. elsdenii ATCC 25940 for n-butanol production

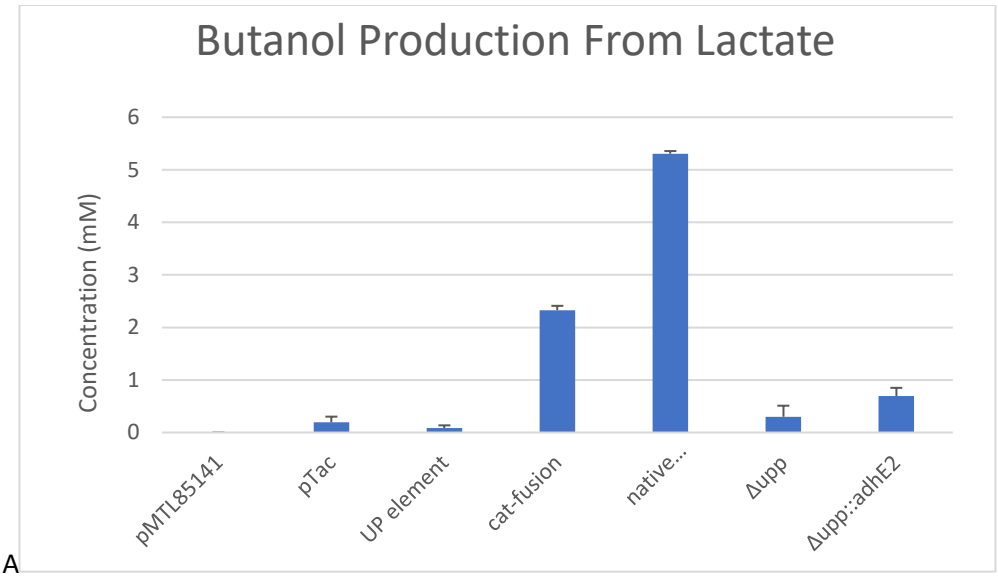
Using the newly developed genetic tools, we next applied these tools for metabolic engineering in *M. elsdenii*. *M. elsdenii* has naturally high flux to butyryl-CoA for butyrate biosynthesis, so n-butanol production was selected as the first target. To enable butanol production, four plasmids were constructed with *adhE2* from *Clostridium acetobutylicum*. This bifunctional aldehyde/alcohol dehydrogenase, *adhE2*, has been demonstrated to have a high affinity for converting butyryl-CoA to butyraldehyde to butanol while other characterized *adhE* preferentially convert acetyl-CoA to ethanol (30, 31).

There are no characterized promoters in *M. elsdenii;* therefore, to enable production of high levels of AdhE2, four different expression plasmids were tested. The first expression plasmid used P_{tac} to drive expression which is a natively highly expressed *E. coli* promoter. Next, a putative UP-element from the *M. elsdenii* ATCC 25940 16S rRNA gene was placed upstream of P_{tac}. A native promoter from *M. elsdenii* (P_{Mels_747}) was also used. Mels_747 encodes an acyl-CoA dehydrogenase and is a likely a part of central carbon metabolism. Because central metabolism genes are often highly expressed, the 300 base pair region upstream of Mels_747 was used to express *adhE2.* Finally, expression of *cat* in pMTL85141 is sufficient for thiamphenicol resistance in *M. elsdenii,* so *adhE2* was placed downstream of *cat* to utilize the same promoter.

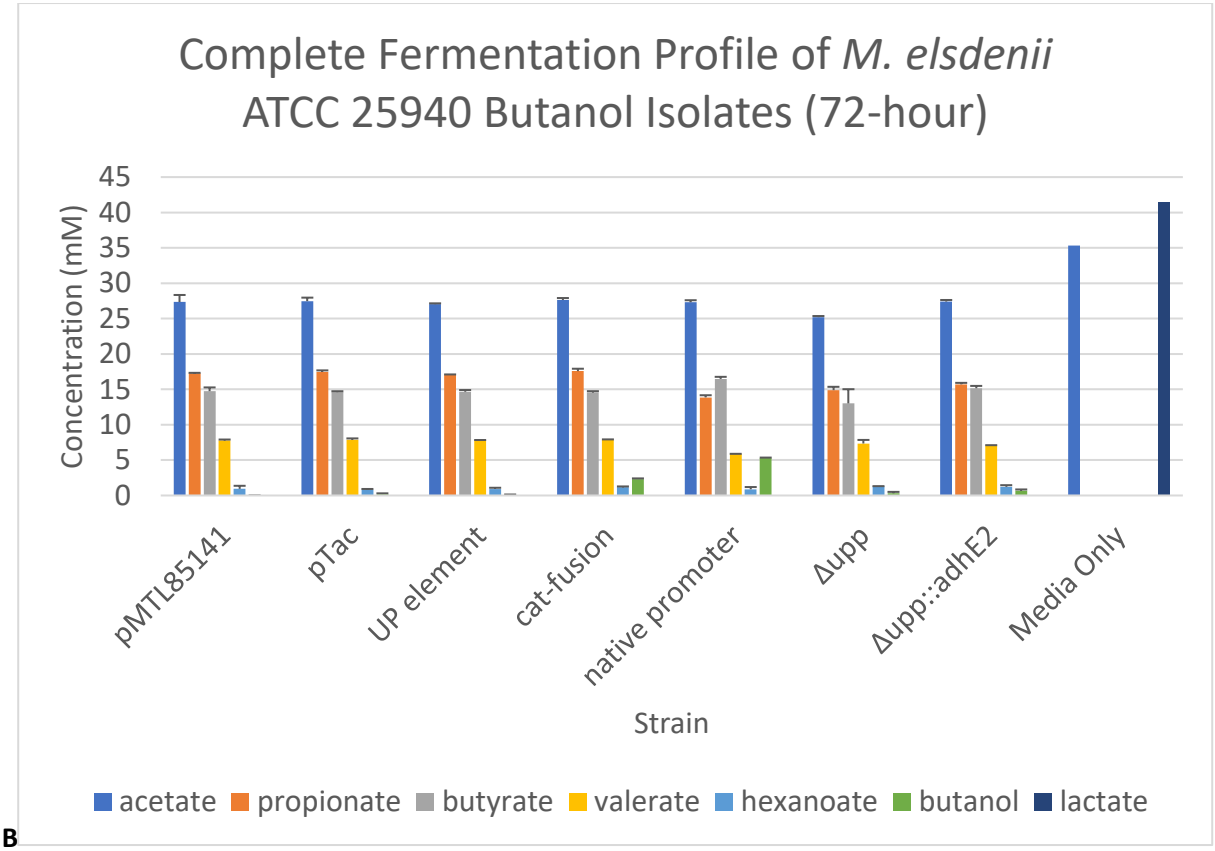
Each plasmid was transformed into each strain of *M. elsdenii* and fermentations were performed using two growth conditions - medium with 5g/L glucose (27.8 mM) and with 5 g/L sodium lactate (44.6 mM) as substrate and thiamphenicol. *M. elsdenii* was grown in each condition for 96 hours and samples were taken at 24-hour intervals for quantification of the carbon substrates and fermentation products. Peak butanol production was seen in the 72-hour samples and butanol concentration decreased at 96 hours, therefore the fermentation profile at 72 hours is reported (Figure 5.2). No butanol was detected in the NCIMB 702410 cultures using either substrate, and no butanol was observed when ATCC 25940 was grown on glucose (data not shown). Wild type *M. elsdenii* ATCC 25940 containing a control plasmid without *adhE2* also showed no butanol production after 4 days. However, the *M. elsdenii* ATCC 25940 strain containing P_{tac}*::adhE2* and the modified P_{tac} containing the *M. elsdenii* UP-element produced about 0.2 mM butanol when grown on lactate. About 2.3 mM butanol was produced using a transcriptional fusion with *cat*. The highest butanol concentration, 5.33 mM, was demonstrated using P_{Mels_747} (Figure 5.2A). The fermentation profile for all strains except the P_{Mels_747} promoter are similar. In the P_{Mels_747}*::adhE2* strain where butanol was produced, butyrate increased by about 2 mM, while propionate was decreased by about 5 mM and valerate decreased by 2 mM. In all cases, the lactate was completely consumed in the 24-hour samples. The growth medium also contains 36 mM sodium acetate, which can serve as a co-substrate. Therefore, the residual acetate is the amount that remained unconsumed (Figure 5.2B).

Integration of adhE2 into the Chromosome of M. elsdenii ATCC 25940

Commercial processes will ultimately require metabolic pathways to be encoded on the chromosome rather than from plasmids. Therefore, we next developed an approach to integrate P_{Mels_747}::*adhE2* into the chromosome. The gene encoding uracil phosphoribosyltransferase (Mels_2191), *upp*, was targeted as the integration site. This gene has been used as a counter selectable marker in a variety of bacteria, where 5-fluororuricil is toxic when *upp* is present but is non-toxic in the absence of



A



B

Figure 5.2: Fermentation profile of *M. elsdenii* ATCC 25940 butanol producing strains from lactate.

A) Butanol production B) Complete fermentation profile.

upp (32, 33). Using a homologous recombination approach, an *adhE2* insertion vector was constructed with 1 kb homology arms upstream and downstream of *upp* flanking *adhE2*, in the multiple cloning site of a replicating vector, pMTL85141. The resulting vector was transformed into *M. elsdenii* ATCC 25940*,* colonies were picked into medium supplemented with thiamphenicol, transferred to medium lacking thiamphenicol, and finally plated on 5-fluorouricil to select for colonies where *adhE2* was integrated into the chromosome in place of *upp*. Using this approach, *adhE2* was successfully inserted into the *upp* locus and the thiamphenicol resistance gene was removed, resulting in strain *Δupp::adhE2*. A Δ*upp* strain lacking *adhE2* was also created in a similar manner to use as a control in fermentations. Fermentation was performed with the control strain and Δ*upp*::*adhE2* using 5 g/L sodium lactate as the growth substrate. A small amount of butanol was detected in the Δ*upp* control strain, while about 0.7 mM butanol was produced in the Δ*upp*::*adhE2* (Figure 5.2A). Both strains had decreased propionate in comparison to the pMTL85141 control (Figure 5.2B).

Discussion

The development of new non-model microbes as synthetic biology chassis organisms is important to the future of biofuel and chemical production due to their native traits of interest that enable the use of new feedstocks, process optimization, and product formation. One such organism, *M. elsdenii*, is of particular interest due to its ability to anaerobically catabolize lactate for efficient medium chain carboxylic acid formation. Previous research on *M. elsdenii* has focused on characterization of fermentation profiles and metabolic modeling due to the lack of genetic tools. By establishing genetic tools for gene integration and deletion for the first time in this genus, we have enabled rational strain engineering. As proof of concept, we demonstrated metabolic engineering for butanol production. While the anaerobic lactate fermentation and efficient chain elongation pathways from *M. elsdenii* were leveraged for alcohol production in this study, many other possible products can be derived from the

carboxylic acid production pathway, and transformation in this organism opens up the metabolic engineering potential.

When *M. elsdenii* was engineered for butanol production from lactate and acetate, propionate and valerate production decreased in comparison to the empty vector control. *M. elsdenii* produces propionate as a way to balance redox reactions, where acetate and butyrate synthesis results in excess NADH production, which is consumed via the acrylate pathway for the production of propionate. Butanol production consumes more NADH than butyrate production, which decreases the need for NAD-recycling via propionate production. Valerate likely comes from the condensation of acetyl-CoA with propionyl-CoA, and with less propionyl-CoA available, valerate concentrations also decrease as expected. Interestingly, while butanol was able to be produced from lactate in *M. elsdenii* ATCC 25940, no butanol was produced using glucose or NCIMB 702410. Applying systems biology "omics" tools will be important to help determine the differences between the two strains as well as the differences between the two substrates that account for the different phenotypes. The information gained could help identify further engineering targets to enable butanol production under these conditions.

Different bioprocessing conditions have the potential to increase butanol production or allow for novel co-culture conditions. Fermentation studies in *M. elsdenii* suggest that in batch fermentations, like the ones performed here, propionate is a major fermentation product. The high concentration of lactate in the medium drives forward the propionyl-CoA transferase reaction in the acrylate pathway. When *M. elsdenii* is grown on lactate in a chemostat under carbon-limited conditions, acetate and butyrate are the major fermentation products (34). Therefore, to increase butanol production in *M. elsdenii*, mimicking these conditions could be beneficial to increase carbon flux and NADH availability for butanol and medium chain carboxylic acids. *M. elsdenii's* capability to efficiently produce medium chain carboxylic acids from lactate also opens up novel bioprocessing options. Many strains of bacteria are capable of producing lactate from renewable feedstocks at high yields. Co-culture of *M. elsdenii* with

lactic acid producing bacteria could allow for efficient production of butanol and medium chain carboxylic acids from a complex feedstock like lignocellulosic biomass. With genetic tools now available in *M. elsdenii*, co-cultures can be engineered to explore this possibility.

Production of butanol using a replicative plasmid allowed for substantial amounts of butanol to be synthesized. Integration of *adhE2* into the chromosome enables butanol production without the need for antibiotic, but butanol production significantly decreased in this strain. This is likely due to a decrease in AdhE2 protein abundance when the multicopy *adhE2* plasmid is replaced by single copy on the chromosome. However, integration of *adhE2* into the chromosome enables further genetic modifications in the strain, which will be critical for further strain optimization. It enables the use of *upp* as a counter-selectable marker for future gene deletions. Initial future efforts should focus on increasing expression of *adhE2* and deleting competing pathways to increase for carbon and NADH flux to butanol. Finally, development of additional genetic tools will allow for further pathway optimization to increase the carbon and NADH availability for butanol production.

References

1. Naik SN, Goud VV, Rout PK, Dalai AK. Production of first and second generation biofuels: A comprehensive review. Renewable and Sustainable Energy Reviews. 2010;14(2):578-97.
2. Devarapalli M, Atiyeh HK. A review of conversion processes for bioethanol production with a focus on syngas fermentation. Biofuel Research Journal. 2015;2(3):268-80.
3. Olson DG, McBride JE, Shaw AJ, Lynd LR. Recent progress in consolidated bioprocessing. Curr Opin Biotechnol. 2012;23(3):396-405.
4. Gao C, Ma C, Xu P. Biotechnological routes based on lactic acid production from biomass. Biotechnology advances. 2011;29(6):930-9.
5. Juturu V, Wu JC. Microbial production of lactic acid: the latest development. Critical Reviews in Biotechnology. 2016;36(6):967-77.
6. Jamshidian M, Tehrany EA, Imran M, Jacquot M, Desobry S. Poly-Lactic Acid: Production, Applications, Nanocomposites, and Release Studies. Comprehensive Reviews in Food Science and Food Safety. 2010;9(5):552-71.
7. Andersen SJ, Candry P, Basadre T, Khor WC, Roume H, Hernandez-Sanabria E, et al. Electrolytic extraction drives volatile fatty acid chain elongation through lactic acid and replaces chemical pH control in thin stillage fermentation. Biotechnology for biofuels. 2015;8(1):221.
8. Zhu X, Zhou Y, Wang Y, Wu T, Li X, Li D, et al. Production of high-concentration n-caproic acid from lactate through fermentation using a newly isolated Ruminococcaceae bacterium CPB6. Biotechnology for Biofuels. 2017;10(1):102.
9. Jeon BS, Choi O, Um Y, Sang B-I. Production of medium-chain carboxylic acids by Megasphaera sp. MH with supplemental electron acceptors. Biotechnology for biofuels. 2016;9:129-.
10. Prabhu R, Altman E, Eiteman MA. Lactate and Acrylate Metabolism by Megasphaera elsdenii under Batch and Steady-State Conditions. Applied and Environmental Microbiology. 2012;78(24):8564-70.
11. Lee N-R, Lee CH, Lee D-Y, Park J-B. Genome-Scale Metabolic Network Reconstruction and In Silico Analysis of Hexanoic acid Producing Megasphaera elsdenii. Microorganisms. 2020;8(4):539.
12. Marounek M, Fliegrova K, Bartos S. Metabolism and some characteristics of ruminal strains of Megasphaera elsdenii. Applied and environmental microbiology. 1989;55(6):1570-3.
13. Nelson RS, Peterson DJ, Karp EM, Beckham GT, Salvachúa D. Mixed carboxylic acid production by Megasphaera elsdenii from glucose and lignocellulosic hydrolysate. Fermentation. 2017;3(1):10.
14. Arber W, Linn S. DNA modification and restriction. Annu Rev Biochem. 1969;38:467-500.
15. Murray NE. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol Mol Biol Rev. 2000;64(2):412-34.
16. W Arber a, Linn S. DNA Modification and Restriction. Annual Review of Biochemistry. 1969;38(1):467-500.
17. Riley LA, Ji L, Schmitz RJ, Westpheling J, Guss AM. Rational development of transformation in Clostridium thermocellum ATCC 27405 via complete methylome analysis and evasion of native restriction-modification systems. J Ind Microbiol Biotechnol. 2019.
18. Chung D, Farkas J, Westpheling J. Overcoming restriction as a barrier to DNA transformation in Caldicellulosiruptor species results in efficient marker replacement. Biotechnology for biofuels. 2013;6(1):82.
19. Yasui K, Kano Y, Tanaka K, Watanabe K, Shimizu-Kadota M, Yoshikawa H, et al. Improvement of bacterial transformation efficiency using plasmid artificial modification. Nucleic Acids Res. 2009;37(1):e3.

20. Mermelstein LD, Welker NE, Bennett GN, Papoutsakis ET. Expression of Cloned Homologous Fermentative Genes in Clostridium Acetobutylicum ATCC 824. Bio/Technology. 1992;10(2):190-5.
21. Hatmaker EA, Klingeman DM, O'Dell KB, Riley LA, Papanek B, Guss AM. Complete Genome Sequences of Two Megasphaera elsdenii Strains, NCIMB 702410 and ATCC 25940. Microbiology Resource Announcements. 2019;8(3):e01430-18.
22. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nature methods. 2010;7(6):461-5.
23. Haldimann A, Wanner BL. Conditional-Replication, Integration, Excision, and Retrieval Plasmid-Host Systems for Gene Structure-Function Studies of Bacteria. Journal of Bacteriology. 2001;183(21):6384-93.
24. Heap JT, Pennington OJ, Cartman ST, Minton NP. A modular system for Clostridium shuttle plasmids. Journal of microbiological methods. 2009;78(1):79-85.
25. De Rossi E, Milano A, Brigidi P, Bini F, Riccardi G. Structural organization of pBC1, a cryptic plasmid from Bacillus coagulans. Journal of bacteriology. 1992;174(2):638-42.
26. Liu J, Xie Z, Merritt J, Qi F. Establishment of a tractable genetic transformation system in Veillonella spp. Applied and environmental microbiology. 2012;78(9):3488-91.
27. Rydzak T, Garcia D, Stevenson D, Sladek M, Klingeman D, Holwerda E, et al. Deletion of Type I glutamine synthetase deregulates nitrogen metabolism and increases ethanol production in Clostridium thermocellum. Metabolic Engineering. 2017;41.
28. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Research. 2015;43(D1):D298-D9.
29. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nature protocols. 2015;10(3):475-83.
30. Atsumi S, Cann AF, Connor MR, Shen CR, Smith KM, Brynildsen MP, et al. Metabolic engineering of Escherichia coli for 1-butanol production. Metabolic engineering. 2008;10(6):305-11.
31. Fontaine L, Meynial-Salles I, Girbal L, Yang X, Croux C, Soucaille P. Molecular characterization and transcriptional analysis of adhE2, the gene encoding the NADH-dependent aldehyde/alcohol dehydrogenase responsible for butanol production in alcohologenic cultures of Clostridium acetobutylicum ATCC 824. Journal of bacteriology. 2002;184(3):821-30.
32. Solem C, Defoor E, Jensen PR, Martinussen J. Plasmid pCS1966, a New Selection/Counterselection Tool for Lactic Acid Bacterium Strain Construction Based on the oroP Gene, Encoding an Orotate Transporter from <em>Lactococcus lactis</em&gt. Applied and Environmental Microbiology. 2008;74(15):4772.
33. Shi T, Wang G, Wang Z, Fu J, Chen T, Zhao X. Establishment of a Markerless Mutation Delivery System in Bacillus subtilis Stimulated by a Double-Strand Break in the Chromosome. PLOS ONE. 2013;8(11):e81370.
34. Prabhu R, Altman E, Eiteman MA. Lactate and Acrylate Metabolism by Megasphaera elsdenii under Batch and Steady-State Conditions. Applied and Environmental Microbiology. 2012;78(24):8564-70.

Chapter VI

Conclusion

Non-model bacteria will be critical to the future of renewable fuels and chemicals. The native complex traits they encode can help to streamline bioprocessing, enable the use of new feedstocks, and enable the production of new chemicals. Non-model organisms are not often used though due to the lack of basic knowledge and the genetic tools available. The first barrier to metabolic engineering and genetic tool development in non-model bacteria is the ability to rapidly enable transformation at high efficiencies. Historically, transformation in a new organism is done on a case-by-case basis and can take years for a researcher to find success. Therefore, to bring on a new era of engineering in non-model bacteria this process needs to be streamlined. The work outlined in this dissertation addresses this barrier by developing a pipeline for the rapid transformation of new bacterial strains. Using *C. thermocellum* ATCC 27405 as a model, a pipeline to enable transformation was described (1). The major transformation barrier is addressed: overcoming the native immune systems within bacteria, primarily the Restriction-Modification (RM) systems. To do this, the complete methylome is determined, which identifies the potentially active restriction modification systems in a strain. Next, the corresponding methyltransferases are expressed from the *E. coli* chromosome. Methyltransferase expression allows for *E. coli* to methylate the DNA of interest to mimic the host's methylome so that after transformation, the host recognizes the heterologous DNA as it's own. Using this pipeline and overcoming the RM systems, each chapter in this dissertation demonstrates transformation for the first time or increase transformation efficiency in non-model bacteria.

The work described focuses on six different bacterial strains, but the first steps of the pipeline have also been employed to a set of photogenically diverse organisms, demonstrating the importance across the bacterial tree of life. To enable the first step of the pipeline, methylome analysis has been performed on 27 different strains of bacteria (Table A.1). The number of active RM systems within each strain ranges from zero, in *Pyrococcus furiosus* COM1, to 9, in *Fibrobacter succinogenes* S85. As the number of RM systems increase, the likelihood of successful transformation without methylation

decreases. An example of this is two strains of *C. thermocellum*, where DSM1313 is genetically tractable with two active systems, even without proper DNA methylation (2). The type strain ATCC 27405, on the other hand, had no reliable genetic tools due to five active RM systems. Another trend included in this data is the vast difference in motifs in highly similar strains, like the two strains of *C. thermocellum*. Therefore, when developing transformation for a new isolate, each strain's methylome needs to be considered, even in closely related strains like *C. thermocellum* ATCC 27405 and DSM 1313.

Bacterial immune systems can be variable within strains of the same species, and often they are acquired through horizontal gene transfer within the bacterial communities they inhabit (3). The work with *S. thermophilus* strains JIM 8232 and DGCC 7710 dives deeper into the function and heredity of immune systems. Two of the most well-known immune systems, RM and CRISPR/Cas, were compared at the genomic and functional levels. The RM systems in the two strains demonstrate how horizontal gene transfer with co-culture counterparts, like *L. lactis*, can create new RM systems and targeted motifs. A Type I RM system (RM3) is present in both strains of *S. thermophilus*, where the *hsdM* and *hsdR* subunits are nearly identical. However, half of the *hsdS* in strain JIM 8232 is identical to strain DGCC 7710 and half identical to a strain of *L. lactis*, creating a hybrid subunit with novel specificity. This demonstrates the rampant evolution of RM systems in *S. thermophilus* that creates new targeted motifs, furthering the arms race between bacteria and hostile DNAs. By understanding the RM systems in both strains, transformation was demonstrated for the first time in strain JIM 8232, and transformation was increased in strain DGCC 7710. By enabling transformation in strain JIM 8232, functionality of the CRISPR systems was able to be tested for the first time, furthering the knowledge base of an important immune system and genetic tool.

The second step in the pipeline is expression of methyltransferases in *E. coli*. Transformation of *C. thermocellum* ATCC 27405 acts as a blueprint for this. Methyltransferases in this work were expressed from the *E. coli* chromosome using the CRIM system (4), which harnesses native *E. coli* phage

recombinases. While this system proved useful, it is limited to five insertion sites and the antibiotic resistance markers are not easily removable. To streamline the process of methyltransferase expression, a new DNA integration system using serine phage recombinases in *E. coli* was developed (5). The developed system allows for rapid integration of 12 constructs into the chromosome with the ability to recycle the antibiotic resistance marker. This system is valuable when creating *E. coli* methylation strains for organisms like *C. clariflavum* 4-2a that have 8 RM systems that need to be expressed. The *E. coli* methylation strain created with this system for *C. clariflavum* 4-2a enabled transformation for the first time in any *C. clariflavum* strain. Using the serine phage integration system, *E. coli* methylation strains were created that target many of the organisms where methylome analysis was performed (Table A.2). These *E. coli* methylation strains proved valuable for overcoming the RM systems in both this dissertation and in collaborative work on other organisms of interest. Creation of these strains using the integration system demonstrates the ability to rapidly expresses methyltransferases stably from the *E. coli* chromosome. This system allows further increase in throughput of the transformation pipeline.

Using the DNA methylation pipeline described, transformation was enabled in several strains of bacteria (Table A.3). Transformation of these non-model bacterial strains demonstrates the importance of overcoming RM systems. In strains such as *Megasphaera elsdenii*, the RM systems are an absolute barrier to transformation. No colonies were obtained using unmethylated DNA and thousands of colonies were obtained using DNA that fully mimics the *M. elsdenii* methylome. While the utilization of fully methylated DNA is ideal for overcoming RM systems, partially methylated DNA can enable low transformation efficiencies. Using DNA that was close to 100% methylated, transformation was enabled in *C. thermocellum* ATCC 25940. In *C. clariflavum* 4-2a, using DNA that is methylated at less than 75% of the sites, low levels of transformation were enabled. Low levels of transformation can act as a starting point for optimization of transformation and the development of basic genetic tools.

Once transformation is enabled in non-model bacteria, more advanced genetic tools can be developed and metabolic engineering can occur. *M. elsdenii* is an example of a non-model organism that holds bioprocessing potential but had no genetic tools available (6, 7). This organism is capable of fermenting lactate and glucose to efficiently produce medium chain length carboxy-acids, in particular butyrate and hexanoate, with the highest titers near 60 g/L (8). This organism was taken through the transformation pipeline, enabling efficient transformation for the first time in this species. Using the transformation techniques, *M. elsdenii* was then engineered for butanol production through the expression of *adhE2* (9). Utilizing lactate as the carbon source, butanol production was demonstrated. Gene deletion and insertion tools were then developed, allowing for future genetic modifications to increase yield and optimize flux. This work begins to establish *M. elsdenii* as a platform organism for the bioproduction of fuels and chemicals. Intermediates in the pathway leading to medium chain carboxy acids can be used for the production of a variety of other chemicals and fuels. *M. elsdenii* was initially engineered for alcohol production, and it can now be utilized for many other purposes using the genetic tools developed here. This work also demonstrates the blueprint for rapid transformation and metabolic engineering of any non-model bacterium of interest for bioprocessing.

Often development of a new non-model microorganism is done on a case by case basis and it can take years to develop transformation and basic genetic tools. To overcome this problem, a rapid and systematic pipeline for enabling transformation in new strains of bacteria is outlined in this dissertation. Methylome analysis was performed on several non-model bacterial strains (Table A.1), the methyltransferases were expressed in *E. coli* (Table A.2), and transformation efficiency was increased or enabled for the first time (Table A.3). This work begins to open up the possibilities of non-model organisms for bioprocessing. To further this work, the development of well characterized and phylogenically diverse genetic parts will be essential. The work with thermophiles, *C. thermocellum* and *C. clariflavum*, demonstrates the need for well characterized methyltransferases that function at

mesophilic temperatures. *M. elsdenii* demonstrates the need for well characterized promoters across the tree of life, where commonly used *E. coli* promoters don't function well in the Negativicutes. As more genetic parts are identified and characterized, it will speed up genetic engineering in a new non-model bacterial strain.

We are entering an era of both rational and systematic design of tools for the metabolic engineering of important non-model bacteria. Transformation methods that overcome native RM systems via targeted DNA methylation and the utilization of libraries of genetic parts has enabled the manipulation of numerous new hosts to exploit their native complex phenotypes. Development of a genetic "toolbox" for rapid analysis of new non-model organisms and their close relatives with the same desirable metabolic capabilities is critical as non-model model organisms begin to be utilized. As more bacteria become genetically tractable and more tools are established in these hosts, advanced genome engineering tools like CRISPR-Cas, biosensors, and phage recombinases will further accelerate metabolic engineering efforts. The ability to fine tune metabolic pathways in organisms that have beneficial, complex phenotypes will enable engineering for increased titer, rate, and yield for a range of important products to build a biobased, sustainable future.

References

1. Riley LA, Ji L, Schmitz RJ, Westpheling J, Guss AM. Rational development of transformation in Clostridium thermocellum ATCC 27405 via complete methylome analysis and evasion of native restriction-modification systems. J Ind Microbiol Biotechnol. 2019.
2. Tyurin MV, Desai SG, Lynd LR. Electrotransformation of Clostridium thermocellum. Applied and Environmental Microbiology. 2004;70(2):883-90.
3. Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. Nature reviews Microbiology. 2020;18(2):113-9.
4. Haldimann A, Wanner BL. Conditional-Replication, Integration, Excision, and Retrieval Plasmid-Host Systems for Gene Structure-Function Studies of Bacteria. Journal of Bacteriology. 2001;183(21):6384-93.
5. Brown WRA, Lee NCO, Xu Z, Smith MCM. Serine recombinases as tools for genome engineering. Methods. 2011;53(4):372-9.
6. Jeon BS, Choi O, Um Y, Sang BI. Production of medium-chain carboxylic acids by Megasphaera sp MH with supplemental electron acceptors. Biotechnology for biofuels. 2016;9.
7. Prabhu R, Altman E, Eiteman MA. Lactate and Acrylate Metabolism by Megasphaera elsdenii under Batch and Steady-State Conditions. Applied and Environmental Microbiology. 2012;78(24):8564-70.
8. Nelson RS, Peterson DJ, Karp EM, Beckham GT, Salvachua DR. Mixed Carboxylic Acid Production by *Megasphaera elsdeni*i from Glucose and Lignocellulosic Hydrolysate. Fermentation. 2017:Medium: ED; Size: Article No. 10.
9. Fontaine L, Meynial-Salles I, Girbal L, Yang X, Croux C, Soucaille P. Molecular characterization and transcriptional analysis of adhE2, the gene encoding the NADH-dependent aldehyde/alcohol dehydrogenase responsible for butanol production in alcohologenic cultures of Clostridium acetobutylicum ATCC 824. Journal of bacteriology. 2002;184(3):821-30.

Appendix

Table A.1: Methylome analysis in strains of non-model bacteria. Methylated motifs were determined by PacBio SMRT analysis and Whole Genome Bisulfite Sequencing. # of Mtases indicates the number of RM systems encoded in the genome.

Organism	# of MTases	Methylated Motifs
Acidothermus cellulolyticus 11b	2	RGATCY, GGCCTG
Bacillus coagulans DSM 2314	3	CAGNNNNNNRTCG, CRTANNNNNNDGT, CYAANNNNNNNTCC
Clostridium acetobutylicum ATCC 824	5	CAAAAAR, CTGAAG, GASTC, GCNGC
Clostridium beijerinckii NCIMB 8205	1	CNTAYNNNNNNCTTC
Clostridium beijerinckii DSM 6423	2	CAGNNNNNTAC, CCCAT
Clostridium carboxidivorans P7	7	AGAAGC, CRAAAAR, GATAAT, CAYNNNNNNCTGC, GCANNNNNNNTNNCG, CGANNNNNNNNTGG
Clostridium clariflavum 4-2a	8	GATC, GACAT, TGAAAG, CAGAAG, ATGCAT, GCGATD, CNAYNNNNCTC, GAGNNNNNNNRTC
Clostridium clariflavum DSM 19372	14	GATC, GACAT, TGAAAG, CAGAAG, ATGCAT, GACNNNNNNNTTC, YGGCCR
Clostridium kluyveri DSM 555	5	GATC, CCAAG, GWTAAT, CAAAAAR, CCGG
Clostridium ljungdahlii DSM 13528	7	CAGNNNNNGAT, GAATAYC, GATAAT, CAAAAAR
Clostridium tyrobutyricum ATCC 25755	9	GATAAT, CAGNNNNNNTAYC, CAAAAAR, CGTCC
Clostridium phytofermentans ISDg	2	GATC, CTGAAG
Clostridium thermocellum ATCC 27405	8	GATC, CNCANNNNNNTTC, GTCAT, GCWGC, GGCC
Clostridium thermocellum DSM 1313	5	GATC, CGACC
Fibrobacter succinogenes S85	9	GCAGG, CACAT, YACAGC, CTGCAGH, GAANNNNNTGCC, GACGTC, GWCAG, GCTGAG, GGWCC
Geobacillus thermodenitrificans DSM465	3	GCCAT
Geobacillus thermoglucosidasius DSM 2542	5	GATC, ACCCA, CCAAAT, GRNGAAGY
Geobacillus toebii DSM 14590	7	GARAATT, GAAAATT
Glycomyces sp. MMG10089	2	TCGANNNNNNNCTC, CAGCAT
Megasphaera elsdenii ATCC 25940	3	GATC, GAGNNNNNNGAT, CAGNNNNNNNTRTC
Megasphaera elsdenii NCIMB 702410	1	CGANNNNNNNNTRTC
Moorella thermoacetica ATCC 39073	6	SATC, AACCA, GAWTC, CTCCG, GGGCCC, GAGNNNNNGAT
Pyrococcus furiosus COM1	2	None
Rhodococcus jostii RHA1	2	GGANNNNNNNNNTGCG, GCTCGAGNR
Streptococcus thermophilus JIM 8232	9	AGCNNNNNGTC, AGCNNNNNNTYTC
Streptococcus thermophilus DGCC 7710	3	GCANNNNNNCTAR, GARANNNNNNNTTCA
Thermoanaerbacterium sp. RBIITD	5	GATC, CGNAAT, CCANNNNNNNNTGC, CAGNNNNNNTTC

Table A.2: Genotype of *E. coli* methylation strains

Organism	*E. coli* Genotype
Acidothermus cellulolyticus 11b	WM3188 Δ*dcm*::*frt* HK::*attB*-cassette R4::Acel_0814
Actinobacillus succinogenes 130Z	WM3188 Δ*dcm*::*frt* HK::*attB*-cassette R4::asp39116ORFCRMP
Amycolatopsis ATCC39116	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette Bxb1::3854-3865 R4::0052-0053 TG1::3889-3890
Bacillus coagulans DSM 2314	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette Bxb1::3854-3865 R4::0052-0053 TG1::3889-3890
Clostridium beijerinckii NCIMB 8205	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette MR11::cbei_3160-3161
Clostridium beijerinckii DSM 6423	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette MR11::cibe_3322 BL3::cibe_3327-3328
Clostridium carboxidivorans P7	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette Bxb::ccar_3570-3575 TG1::ccarb_11960-11965 A118::ccarb_24310-24315 MR11::ccarb_7120 R4::ccarb_8705 ccarb_17831
Clostridium clariflavum 4-2a	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette R4::Ccla_2133 phik38::Ccla_2368-2369 BXB1::Ccla_1830-1831 TP901::Ccla_1058 TG1::Ccla_1868-1869 BL3::Ccla_1961 phi370::Ccla_1994 A118::Ccla_1058
Clostridium kluyveri DSM 555	WM3188 Δ*dcm*::*frt* HK::*attB*-cassette TG1::Ckl_2332 phiC1::Ckl_2671-2672 BL3::Ckl_3239 MR11::Ckl_2694 wβ::Ckl_3294
Clostridium tyrobutyricum ATCC 25755	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette MR11::Ctk_27600-10
Clostridium phytofermentans ISDg	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette BL3::cphy_2924 MR11::cphy_267 λ::cphy_267
Clostridium thermocellum ATCC 27405	Top10 Δ*dcm*::*frt* λ::Cthe_0519 HK::phi3T1
Clostridium thermocellum DSM 1313	WM3188 Δ*dcm*::*frt* HK::*attB*-cassette R4::M. RFlT3
Fibrobacter succinogenes S85	Stellar HK::*attB*-cassette TP901::Fisuc_1827 Fisuc_2500 Fisuc_0263 R4::Fisuc_2878 Fisuc_0930 Fisuc_0069 phifc1::Fisuc1346 TG1::Fisuc1298
Geobacillus thermodenitrificans DSM 465	WM3188 Δ*dcm*::*frt* HK::*attB*-cassette R4::gth465ORFAMP
Geobacillus thermoglucosidasius DSM 2542	WM3188 Δ*dcm*::*frt* HK::*attB*-cassette R4::AOT13_9970 MR11::AOT13_05680 MR11::AOT13_12675
Glycomyces sp. MMG10089	Stellar HK::*attB*-cassette BxB1::Glycomyces T2MT A118::Glycomyces T1MT

Table A.2 Continued

Organism	*E. coli* Genotype
Megasphaera elsdenii NCIMB 702410	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette BxB1::C6Y28_02465-02475
Rhodococcus jostii RHA1	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette MR11::RHA1_ro11151-52
Rhodopseudomonas palustris CGA009	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette BL3::pRpa4 MR11::Rpa3376
Thermoanaerobacter tengcongensis MB4T	BW25113 Δ*mcrA*::*frt* Δ*mcrC-mrr*::*frt* Δ*dam*::*frt* Δ*dcm*::*frt* HK::*attB*-cassette BL3::Tte_1546-1547

Table A.3: Transformation in non-model bacteria. Each strain was transformed using DNA isolated out of wild type *E. coli* and the best methylation strain available. Numbers are reported in colony forming units per microgram of DNA (CFU/μg).

	CFU/μg DNA with plasmid isolated from:	
Organism	**Wild type *E. coli***	**Best methylating strain**
C. thermocellum DSM1313	4	18380
C. thermocellum ATCC27405	0	80
C. clariflavum strain 4-2A	0	6
S. thermophilus DGCC7710	1930	261000
S. thermophilus JIM8232	43	106000
G. thermoglucosidasius DSM 2542	1500	9110

Vita

Lauren Riley received her B.S. in biology from Rensselaer Polytechnic Institute in Troy, New York. She first came to Oak Ridge National Laboratory with the summer SULI program where she worked to enhance biofuel production in *E. coli*. After graduation she came back to ORNL as a post-bachelors research associate working to develop genetic tools in *Clostridium thermocellum*. She is now graduating with her PhD in Energy Science and Engineering from the University of Tennessee. Her current research is focused on developing genetic tools to allow genetic modifications of non-model bacteria by overcoming their native immune systems. She is used these new tools to engineer *Megasphaera elsdenii* for biofuel production.