THE DEMOCRATIZATION OF GENOME ENGINEERING: TOOLS AND TECHNIQUES FOR MANIPULATING MICROORGANISMS WITH CRISPR-CAS

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Brian Joseph Mendoza

May 2020

Copyright © 2020 by Brian J. Mendoza
All rights reserved.

ACKNOWLEDGEMENTS

This work was funded by the Department of Energy's BioEnergy Science Center (DOE-BESC), Defense Advanced Research Projects Agency (DARPA) Young Faculty Award, and subsequent DARPA Director's Fellowship. Fellow lab members Dr. Xianliang Zheng (Postdoctoral Fellow) and Jared Clements (undergraduate mentee) were integral to the experimental execution of Chapter 5 and Katherine Krouse (undergraduate mentee) for experimental execution of Chapter 6.

ABSTRACT

The discovery of the Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) family of genome sequences in bacteria and archaea has led to expansive application of the system as a means of interrogating and modifying genetic material across all kingdoms of life. This dissertation has developed algorithms and accompanying software, collectively named CASPER (CRISPR Associated Software for Pathway Engineering and Research), to identify promising sequences for effectively utilizing CRISPR associated (Cas) proteins in any species or community/metagenome of interest. The on- and off-target activity scoring algorithms improve upon previous work, first by utilizing an evolutionary algorithm and then by employing protein simulation to identify the contribution of sequence stability and mismatches to overall activity. Two algorithms, multitargeting and population analyses, employ searches across multiple genomes for deploying both degenerate and specific guides with immediate application in microbiome manipulation and CRISPR-Cas enabled rapid sequence detection assays. Finally, the CASPERpam algorithm was written to rapidly identify potential protospacer adjacent motifs (PAM) that are required for initial binding by Cas proteins/complexes. CASPERpam successfully re-identified 12 experimental PAM sequences and further identified putative PAM sequences for another 1,037 species.

The application of CRISPR-Cas as an antimicrobial is of immediate interest due to rising prevalence of antibiotic resistance and the associated costs of development. Using the aforementioned algorithms for sequence design, this dissertation identified key barriers and opportunities for development of novel CRISPR-Cas antimicrobials that are both highly specific and highly tunable. Specifically, the identification of kinetic limitations to the transient introduction and expression of CRISPR-Cas machinery is a key mechanism of cell persistence. Studies in both model yeast (*Saccharomyces cerevisiae*) and multiple species of bacteria (*Escherichia coli*, *Bacillus subtilis*, and *Staphylococcus aureus*) highlight the benefits of co-targeting defensive genes with high turnover rate Cas enzymes as a means of mitigating such kinetic barriers. These design principles are inherently generalizable to multiple pathogenic species, and benefit from the inherent specificity of CRISPR-Cas sequences to lay the foundation for a powerful new class of antimicrobials.

TABLE OF CONTENTS

Chapter One Introduction 1
General Overview 1
The Rise of CRISPR-Cas 1
The Race Begins 1
CRISPR-Cas Mechanisms and Biology 2
Predicting CRISPR-Cas Behavior 3
Orthologs and applications of their functionalities 4
Engineering new types of Cas 5
CRISPR-Cas as Antimicrobials 6
Targeting Bacteria with CRISPR-Cas 6
Bacterial Responses to CRISPR-Cas 7
Chapter Two Identification and quantification of CRISPR-Cas targets 9
Introduction 9
Methods 11
On-target activity formulation for CRISPR gRNA design 11
Off-target formulation for gRNA design 13
Multitargeting analysis 14
Multipopulation analysis 14
Results and Discussion 15
Development of CASPER algorithm for enhanced on-target activity prediction 15
Development of CASPER algorithm for enhanced off-target activity prediction 16
Development multitargeting analysis 17
Development multipopulation analysis 19
Conclusion 20
Chapter Three A crystal structure energy calculation ensemble for predicting off-target activity of single-effector Cas nucleases 21
Introduction 21
Methods 23
Structure preparation 23
Rosetta algorithms 23
Ensemble generation 24
Statistical analysis for correlation to experimental data 24
Results and Discussion 25
On-target activity is not predicted by heterotriplex energetics 25
Truncation scoring approximates R-loop formation energetics 25
Correlation of ΔΔG proxy values to experimental off-target activity 26
Cas12 dataset highlights genomic loci contribution to off-target activity 27
Conclusions 28
Chapter Four In silico processing of the complete CRISPR-Cas spacer space for identification of PAM sequences 30
Introduction 30

Methods 31
Spacer collection for PAM analysis 31
Genome collection for PAM analysis 32
PAM detection and analysis 32
Computation 33
Results and Discussion 33
Predictive PAM sequences match experimental evidence 33
Quantification of the quality of in silico PAM prediction 34
PAM prediction highlights diversity and patterns within and between Cas types 36
Conclusion 38
Chapter Five Development of CRISPR-Cas antifungals 39
Introduction 39
Materials and Methods 40
Strains and Plasmids 40
Media and culturing conditions 41
Flow cytometry 41
Library generation and sequencing 41
Results and Discussion 42
Screening target combinations in S. cerevisiae 42
Characterization of high-performing CRISPR-Cas targets 42
High cell-loading results in more pronounced community escape 43
Targeted amplicon sequencing reveals escape genotypes 44
Library depletion assay for selecting CRISPR-Cas essential gene targets 45
Characterization of promising targets with tetracycline inducible construct 45
Conclusions 46
Chapter Six Identification of target and cas nuclease combinations for antimicrobials 47
Introduction 47
Materials and Methods 48
Strains and Constructs 48
Media and culture conditions 49
Agar plate screening assay with M13 phage 49
Flow cytometry measurements 49
Library screens in S. aureus 50
Results and Discussion 50
Bacillus subtilis is prone to escape when targeted with spCas9 50
Screening E. coli targets with endogenous CRISPR 51
Phagemid delivery and co-targeting SOS response with CRISPR-Cas 51
Library screen to identify common targets and promising alternative endonucleases in the pathogen Staphylococcus aureus 52
Conclusions 53
Chapter Seven Conclusions and recommendations for future work 54
Expanding population engineering computational workflow 54

Increased modeling accuracy with Rosetta parameterization 54
Expanding principles to pathogenic fungi and validating mechanisms of delivery 55
Controlling the microbiome through selective species 56
References 57
Appendix 70
Vita 102

LIST OF TABLES

Table 1: PDBs used in Rosetta study. 82
Table 2: Top species by confidence score. 90
Table 3 Log2 fold changes for most depleted sgRNA sequences. 101

LIST OF FIGURES

Figure 1 CRISPR-Cas adaptive immunity. 71
Figure 2 CRISPR-Cas types. 72
Figure 3 Correlation of CASPER algorithms. 73
Figure 4 PAM motif appearances. 74
Figure 5 Multitargeting analysis. 75
Figure 6 Multitargeting cellular processes. 76
Figure 7 Population targeting. 77
Figure 8 Algorithm schematic. 78
Figure 9 PAM density scoring table. 79
Figure 10 Rosetta computational workflow. 80
Figure 11 On target activity lack of correlation. 81
Figure 12 Truncation schematic. 82
Figure 13 RosCas correlation statistics. 83
Figure 14 Ensemble sensitivity. 84
Figure 15 lbCas12 correlation and statistics. 85
Figure 16 Schematic of CASPERpam workflow. 86
Figure 17 CASPERpam algorithm for identifying PAM sequences. 87
Figure 18 Predicted PAM for validated species. 88
Figure 19 Predicted PAM sequences and categorization. 89
Figure 20 Schematic of co-targeting and agar serial dilution heat maps. 91
Figure 21 Detailed characterization of 18 strains. 92
Figure 22 Effect of inoculation concentration on killing efficiency. 93
Figure 23 Targeted amplicon sequencing results. 94
Figure 24 Single plasmid library and characterization. 95
Figure 25 Characterization of library-identified sgRNAs. 96
Figure 26 Growth characterization of *B. subtilis*. 97
Figure 27 Characterization of saCas9 and ecoCASCADE. 98
Figure 28 Library depletion assays in *S. aureus.* 100

CHAPTER ONE
INTRODUCTION

General Overview

The purpose of this thesis is to lay out a framework for marrying principles of engineering with the rise of CRISPR-Cas as a means of manipulating a variety of organisms in diverse environments[1]. This combinatorial space is poorly suited to examination with traditional pedagogies of biological research. While many fundamental principles and governing dynamics remain elusive in each field, taking an engineering approach to solve some of the more complex problems can not only serve as a useful interdisciplinary lens, but also streamline the deployment of impactful applications to tackle problems ranging from disease epidemics to biodiversity loss and agricultural stress from climate change. It is the hope of the author and his collaborating researchers that the tools and techniques stemming from this research will aid in fulfilling more promises and evading pitfalls facing the deployment of CRISPR-Cas as it moves into its next frontier.

The Rise of CRISPR-Cas

The Race Begins

While the existence of the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) array and its associated (Cas) open reading frames (ORFs) had been known since the 1990s[2], their biochemical nature was a mystery until 2009 when Francis Mojica inferred its functionality as an adaptive immune system by mapping the sequences in the array (the spacers) to known sequences from bacteriophages[2,3]. Mojica *et al.* in 2009 also inferred that these sequences were specifically acquired, as suggested by the presence of integrase-like proteins nearby[3]. Further identification of nuclease domains in the nearby ORFs suggested that the proteins were being honed to viral sequences for destruction[4]. Shortly following the discovery of DNA as a target molecule, the full characterization of the homing complex was elucidated and its biochemistry quickly began to be applied to other systems by heterologously expressing the necessary machinery[5-7]. The final piece of the puzzle for potential widespread application was the innovation of the fusing of the two necessary RNA fragments, the crRNA (CRISPR RNA) and the tracrRNA (trans-activating RNA) into a single RNA that could be easily expressed from a single RNA Polymerase III promoter[8]. The potential impact of a system that could program an enzyme complex to a specific site in the genome with the use of a simple 20-30 nucleotide RNA sequence was immediately recognized as a transformative technology in biology[7-10]. Within 10 years from the initial functional discovery, CRISPR-Cas had become a household name.

CRISPR-Cas Mechanisms and Biology

CRISPR-Cas utilizes a network of nucleic acid interacting enzymes to form an adaptive immunity system and is present in the majority of bacteria and archaea. Briefly, upon infection from a bacteriophage, the surveying Cas1 and Cas2 proteins identify the sequence and use their transposase and integrase functionalities to copy a section of the phage genome into the CRISPR array within the bacterial genome (Figure 1, All figures appear in Appendix)[11]. By integrating into the genome, the sequence is saved for future reference should the bacterium see the invading sequence again (the adaptive immune response). The CRISPR array is then transcribed into RNA and then processed by other accompanying Cas genes into mature crRNA sequences. These crRNAs then combine with the Cas nuclease complex, and the entire ribonucleoprotein (RNP) is directed to the foreign DNA (or RNA) sequence via sequence complementarity with the crRNA[12].

The diversity of CRISPR-Cas systems is classified into 2 classes, 6 types, and 24 subtypes (Figure 2)[1,12]. Class 1 systems are the multi-effector nucleases and are comprised of the Types I and III[13,14]. From phylogenetic analysis, Type IV is also classified into Class 1; however, the mechanism of this type remains cryptic[15-17]. Structurally, Type I and III are similar, as they both rely on a multi-effector complexes to induce cleavage but exhibit distinct functionalities[18]. Type I systems are defined by their multi-effector recognition complexes that bind the crRNA and DNA followed by recruitment of the Cas3 helicase/nuclease protein. In contrast, Type III systems rely on the Cas10 protein for transcribed RNA recognition and cleavage, and use recruitment of Csm/Cmr subunits to stabilize the structure in between recognition and cleavage[5,14,17,19].

Class 2 contains the single-effector nucleases of Type II, V, and VI[14,17]. These systems have been widely co-opted for heterologous expression due to their single-effector nature, which—along with the guide-RNA—requires just a single protein to be expressed for recognition and cleavage. This capability was quickly realized by multiple groups using the Cas9 (Type II) protein from *Streptococcus pyogenes* (spCas9, II-A)[*8,9,20*]. Coincidentally, this Cas9 proved to be the most robust across all multiple organisms and *in vitro* assays, making it the most well-studied system to-date[21]. Due to the large size of the spCas9 (160kDa) other smaller Cas9 orthologs have been sought to be optimized so they could be more readily packaged for lenti- or adenoviral delivery[22-24]. Common to all Cas9 is their use of trans-activating (tracrRNA) to complex with the crRNA and form a double-strand break (DSB) in the DNA substrate[8,24]. In contrast, Type V systems lack a tracrRNA, and create a staggered cut on the DNA. These proteins were discovered by the Zhang group at the Broad Institute[25] in the *Acidaminococcus* spp. and were given the name Cpf1 before being redesignated Cas12 to match the nomenclature of the expanding suite of CRISPR-Cas systems[17]. Type VI is the most recently discovered class and is quite unique from Type II (Cas9's) and Type V (Cas12's) as it cleaves single-stranded RNA (ssRNA), while also exhibiting *trans* nuclease activity on non-target ssRNA once activated via target binding[26,27].

Predicting CRISPR-Cas Behavior

Like its predecessors, the Zinc Finger Nucleases (ZFNs) and transcription activator like endonucleases (TALENs)[28], CRISPR-Cas directed targeting of genetic elements requires the design of a homing sequence. Prior to CRISPR-Cas, such tools took the shape of identifying the correct order of short DNA sequence recognition proteins and assembling them together to make more specific arrays. With CRISPR-Cas, the combinatorial space available (4^n, n = length of RNA guiding sequence) requires a different and more facile approach to rational design. In addition, the ease of producing a vast library of sequence variants enables the use of high-throughput screening methods to assay activity. Many such studies exist, using a variety of host organisms/cell lines, and more recently different Cas endonucleases[25,29-33]. While nuances exist in each study, the general experimental layout follows a similar procedure. First, a library of sequences is generated from oligo synthesis. This group of oligomers is then transformed into a cell line of interest, and their activity is assayed either through the existence of a certain phenotype (e.g. GFP fluorescence, cell morphology) or through a more quantitative sequencing of the target DNA or sgRNA distribution profile. Phenotypic assays are most commonly used to detect on-target activity, defined as the activity from a perfect sgRNA/DNA match[29,31]. On the other hand, off-target activity—activity at a non-perfect sgRNA/DNA site—is most commonly detected through DNA re-sequencing of closely matched target sites, or the more expensive but unbiased genome-wide mutation assay[34-36].

While the aforementioned studies were designed to identify design rules for making highly active guide-RNAs, their combined results seem to indicate that such a rational workflow for predicting what is referred to as on-target activity remains elusive. Elements of certain studies have identified the possible reasons for the difficulty in ascertaining general design rules[37]. In one study[38], the state of supercoiling of a plasmid was shown to drastically affect cleavage by a Cas9 ortholog, suggesting the state of DNA superstructure may play a role in governing activity. This is further supported by seminal FRET analyses by Doudna and colleagues that showed the R-loop formation—when the ribonucleoprotein complex binds to the double stranded DNA and unwinds it—as the energetically limiting factor to activity[39]. A DNA coiling that reduces this thermodynamic barrier would lead to more facile binding and therefore greater activity at such sites. Genomic DNA contains a great variation in superstructure, which is used in many organisms to regulate gene expression in concert with traditional protein-based transcriptional regulation[40]. Of note, nucleosomes (i.e., histones) are essential elements in packing genetic material, but have been shown to reduce DNA accessibility to CRISPR-Cas nucleases[41,42]. To further complicate the endeavor, global genomic structure is subject to the stages of cell growth, while local regions can be further modulated based on specific environmental conditions. Understanding the complex interplay of these factors in genomes of multiple organisms with multiple cell types is a near insurmountable task with current methodologies, so the focus on the relative activity of a particular guide-RNA

sequence is a necessity. Mendoza *et al.* (see Chapter 2) coined the terms relative and absolute activity, referring to the activity that can be assigned purely to the guide-RNA sequence holding all other factors constant, which is predictable, in contrast to the absolute activity, which is experimentally observable[43]. The ability to deconvolute the relative from the absolute activity is a subject of active research, with a variety of models using advanced machine learning techniques to more accurately assign relative activity scores based on canonical experimental datasets[44-47]. Ultimately, the ability to tile multiple guides across the desired locus target to determine the best performing sequence is a cost and time efficient experiment, which limits the utility of on-target prediction algorithms. However, the ability to identify potential off-target sites is of paramount importance, as assaying genome-wide off-target activity for a large number of guides can quickly become intractable experimentally[28].

Off-target activity, defined as the non-specific activity at sites that contain mismatches between the guide-RNA and targeted DNA, has been a key limitation in the deployment of CRISPR-Cas in many of its potential applications, particularly therapeutics[43,48]. Early studies began investigating the promiscuity of Cas enzymes, with the majority of the focus on spCas9[36]. Due to the variation in cell types and experimental conditions confounding absolute activity, assays from different groups led to different conclusions regarding Cas enzyme propensity for off-target activity[35,49,50].

The location of the potential off-target activity is relevant to its assessment as a potential risk. An identified off-target site that is located on a crucial gene that would result in silencing or harmful mutation, faces a much stricter cutoff than one that is located in an isolated topologically associated domain (TAD) on a non-coding sequence[28,51]. Streamlining the integration of this assessment in the well curated genomes (e.g., human, mouse) will be an important next step in off-target activity analysis by adding a risk variable to the prediction exercise.

Orthologs and applications of their functionalities

Despite the ubiquity of spCas9, other orthologs have been recently gaining attention for their unique chemistries that enable the development of novel applications (see Figure 2 for summary). The Type I and III systems show promise for harnessing endogenous systems. They also hold promise due to their decoupled nuclease activity from their recognition domain, allowing potential modulation via a recruitment effect[13,18]. Type V systems, of which Cas12 (formerly Cpf1) belongs, is quite similar to Cas9 in its single-effector structure and double-stranded DNA substrate[25]. Recently, the discovery of trans-nuclease activity in many Cas12 enzymes has suggested it holds promise in a variety of new applications, particularly that of antimicrobials and DNA identification and degradation[33]. Finally, the Type VI endonucleases (Cas13 family) targets a single-stranded RNA substrate and exhibits robust trans nuclease activity[26,52]. Unlike their Cas12 cousins, the trans nuclease domain is separate from the endonuclease domain, which is thought to contribute to its broader trans activity. Cas13s have

been deployed to great effect as sensitive nucleic acid biosensors, utilizing this trans activity to cleave synthetic indicator RNA analogs mixed into a sample[53,54]. This capability shows promise for a new generation of rapid biological detectors, with applications ranging from prompt pathogen identification to environmental profiling.

Engineering new types of Cas

CRISPR-Cas's "find" capability is arguably its most powerful. Despite ZFNs and TALENs also sharing in this ability, their functionality is limited due to their involved multimeric assembly. Single-effector CRISPR-Cas systems (Type II,V,VI), contain only 1-2 nuclease domains that can be easily inactivated with point mutations, rendering catalytically dead proteins that only bind the DNA[55]. Multi-effector systems (Type I,III) can also be rendered catalytically inactive through the knockout of the separated nuclease enzymes (e.g. Cas3)[13]. These dCas proteins are used to great effect in gene knockdown assays and can be combined with a variety of experimental conditions to identify relevant pathways for inducing certain phenotypes. So widespread are such assays that the procedure is commonly referred to as CRISPRi (CRISPR inactivation/inhibition)[56]. As expected, the placement of the target sequence relative to key elements in the operon (promoter, start codon, ribosome binding site, introns, etc.) is key to effective inhibition[57]. As such, a multitude of software tools have dedicated CRISPRi algorithms for designing appropriate targets.

Some studies have decided to take CRISPRi a step further, by fusing to the N-terminus a transcriptional repressor such as KRAB, enhancing the inhibitory effect. On the opposite end, activator proteins such as VP64 can be fused to dCas9, creating a gene-activator complex in a process known as CRISPRa (CRISPR activation)[58]. To add even further combinatorics, other Cas9/Cas12 orthologs have also been fused with repressors and activators and show varying degrees of activity relative to their spCas9 counterparts providing a near-analog toolkit for designing gene regulation elements with CRISPR-Cas[59].

Recently, the development of adenine and cytosine deaminase domains fused to a spCas9 nickase (one catalytic domain knocked out resulting in double-strand nicking) has produced a novel functionality of precisely directed point mutations. The unwinding of DNA by Cas9 and nickase activity allows the deaminases to target bases in the 12-16 nucleotide PAM-distal region of the DNA sequence[60]. The precise sequence requirements for base editing prompted further engineering to decrease PAM specificity to make available more sites in the genome able to be editing by a CRISPR-Cas base editor[61]. Due to the large number of human diseases associated with point mutations, this technology is a promising new direction for clinical deployment of CRISPR-Cas.

CRISPR-Cas as Antimicrobials

Targeting Bacteria with CRISPR-Cas

While CRISPR-Cas is native to prokaryotes, the immediate attention of its applicability was turned to the promise of heterologous deployment in relevant eukaryotes for facilitating gene editing based therapy. Despite promise in this area, turning the system against its bacterial origins is a significant opportunity for pathogen control and microbiome manipulation[62,63]. Because many bacteria already contain active CRISPR-Cas machinery, one can either harness this native system or express a heterologous system to enact the desired genome modifications. Each method comes with its own strengths and weaknesses, prompting a question of rational design depending on the desired application. This section will first focus on techniques for harnessing endogenous machinery before turning its attention to the heterologous approach.

The ability to harness native CRISPR-Cas systems for modification of the host's genome presents a different set of variables than heterologous eukaryotic manipulation. First, the expression of CRISPR-Cas systems is subject to environmental conditions and cellular stress responses, in particular the presence of foreign species and phages[64]. While this can be detrimental in some cases, this enacts a degree of control that may be desirable, particularly when dealing with modulating species within a population. In fact, the ability to control expression of certain genes with CRISPR-Cas mediated interference has been shown to be a natural phenomenon[62,65]. Multiple species' native CRISPR-Cas systems have been utilized for editing and gene knockdown via the introduction of self-targeting guides, notably *Pseudomonas aeruginosa* (Type I)[66], *Escherichia coli* (Type I)[65], *Clostridium pasteuranium* (Type I)[67], *Staphyloccous aureus* (Type II, Type III)[68]. In particular, due to the hyperactive nature of CRISPR-Cas in multidrug resistant *P. aeruginosa,* its system was turned against the *mexB* gene (efflux pump component) successfully re-sensitizing the strain to antibiotics[69]. Use of the native machinery in *C. pasteuranium* was also compared to the heterologous approach and shown to be significantly more efficient than use of a heterologous Cas9. While these reports highlight the potential of the endogenous approach, working in a new organism requires careful characterization of the endogenous machinery (e.g., PAM identification) to ensure proper targeting[70]. In contrast, the heterologous approach utilizes a well-characterized Cas system that has the potential to be rapidly deployed to a larger number of organisms, specifically eukaryotes as they contain no native CRISPR systems.

Heterologous expression and activity of CRISPR-Cas single-effector nucleases has shown to be surprisingly robust across a wide variety of hundreds of organisms[14,71]. Paradoxically, the domains exhibiting the greatest difficulty with heterologous Cas expression are those of bacteria and archaea. This is due to two factors: 1) the global cellular response to DNA DSBs and 2) the broad native defense systems (including native CRISPR-Cas) that inhibit foreign DNA uptake[72,73]. These barriers will be crucial to address when deploying CRISPR-Cas

in bacteria. Due to the DSB toxic response, many research endeavors into CRISPR-Cas in bacteria usually employs catalytically inactivated enzymes for gene knockdown[13]. Targeting genes with dCas9 has created a robust technique to screen genes for essentiality that are helpful in the identification of putative drug targets[74]. However, the use of CRISPR itself as an antimicrobial is promising, due to the tuning of the system to a unique sequence, and therefore unique sequence, limiting the undesirable targeting of commensurate species in a host or environmental microbiome[75]. However, many improvements must be made to CRISPR-Cas efficiency and delivery for such a system to be deployed in real-world applications[62,63,76].

Bacterial Responses to CRISPR-Cas

Evolution dictates the survival of a population over that of an individual cell, meaning that cellular adaptation and sacrifice is a means of preventing potentially harmful genetic elements (including self-targeting CRISPR-Cas machinery) from making their way through a population. The SOS response of bacterial is a highly regulated system that is induced by a variety of environmental factors. Briefly, LexA (repressor) and RecA (inducer) act as a control gate for the expression of downstream genes whose products are responsible for error-prone repair of a damaged DNA region[73]. In many bacteria, RecA governs a strong homologous recombination pathway that is highly favored due to the fact that rapidly dividing bacteria may harness up to 10 copies of its genome in a single cytoplasm. Further downstream, the error-prone polymerases (*polB, dinB, umuC, umuD*) are increasingly upregulated the greater the amount of DNA damage present. This genetic coordination can potentially have several advantages. First, it serves as an adaptive mechanism, allowing the induction of the SOS response to create mutations that could potentially serve as beneficial to the new environment[72]. Second, it also provides a means of mutating away from highly active CRISPR-Cas expression. Specifically, the greater number of breaks induced can increase the likelihood that an error-prone polymerase repairs the break by introducing mismatches, thereby rendering the expressed guide-RNA inactive. Therefore, managing the SOS response is key to achieving high levels of CRISPR-Cas activity. Inhibition of the activating RecA demonstrates this principle and is a key insight for designing CRISPR-Cas antimicrobials[73]. Future studies can be directed at such SOS response co-targeting to investigate the chokepoints that enable high CRISPR-Cas toxicity.

Not all bacteria exhibit an SOS response. In particular, the well-known pathogens *Campylobacter jejuni, Streptococcus pneumoniea, Streptococcus pyogenes, Legionella pneumophila, Helicobacter pylori, Neisseriae meningititis, and Neisseriae gonorherae* all lack an SOS response system[72]. These species are hypothesized to make up for this seeming lack of repair pathway by upregulating their competence mechanisms, taking in foreign DNA as a means of adaptation. Further, other pathways loosely regulated by SOS exist in many bacteria and archaea, namely primitive non-homologous end-joining (NHEJ) and

microhomology end-joining (EJ) proteins that can repair DNA double strand breaks in low quantities[28,77]. These two mechanisms can also serve to adapt away from self-targeting CRISPR-Cas[13]. Bacteria also co-opt valuable elements from their infecting phages. Relevant to the study of DNA repair is that of the highly active ligases that can efficiently repair DSBs[77]. These may induce mutations at the site of the break, preventing further targeting.

The diversity of DNA repair pathways throughout the bacteria and archaea kingdoms suggest that increasing CRISPR-Cas antibacterial potency can be achieved by weakening host defense systems but the choice is highly species- and environment-dependent. Even those species without an SOS response are sensitive to the stress associated with introducing exogenous CRISPR-Cas machinery, whether it be through plasmid delivery via transformation or conjugation, or through phage RNA/DNA delivery. This makes targeting the defensive machinery all the more crucial for designing efficient CRISPR-Cas antimicrobials.

CHAPTER TWO
IDENTIFICATION AND QUANTIFICATION OF CRISPR-CAS TARGETS

Summary: A version of this chapter was originally published by Brian Mendoza and Cong T. Trinh titled: “Enhanced guide-RNA design and targeting analysis for precise genome editing of single and consortia of industrially relevant and non-model organisms.”[43]

Introduction

Transforming biology into engineering practice has shaped the frontiers of synthetic biology and metabolic engineering [78,79]. This helps drive industrialization of biology with broad applications related to food, health, energy, and the environment from production of drugs to fight diseases to synthesis of renewable fuels to replace fossil fuels [80]. Genetic diversity of non-model organisms offers a repertoire of unique phenotypic features for exploration and cultivation for these applications [81,82]. To realize this enormous potential, it is critical to have an efficient genome editing tool for rapid strain engineering of these organisms to perform novel programmed functions. In the last several years, the CRISPR technology has emerged as a powerful genome editing tool for metabolic engineering and synthetic biology applications [9,13,48,83-91].

Using a CRISPR/Cas system for genome editing requires it to be both highly active and accurate. CRISPR tools that are used to design guide-RNAs (gRNAs) to form ribonucleoprotein (RNP) complexes for effective genome targeting focus on two types of evaluation. The first type is the prediction of “on-target” activity of a selected gRNA on a target DNA (or RNA) sequence on the genome. This on-target activity reflects the ability of such a gRNA to successfully find and bind the Cas endonuclease to the target sequence. The second type is the prediction of “off-target” activity of a selected gRNA, which pertains to the propensity of the RNP complex to interact with sequences on the genome similar to the target sequence. Choosing gRNA sequences with high activity is paramount to all genome editing endeavors for metabolic engineering and synthetic biology applications [57,92,93]. To predict which sequences lend themselves to efficient targeting, various algorithms have been developed based on high-throughput experimentation and validation [29,30,94]. By identifying prominent nucleotide features in highly active gRNAs and their target sequences, scoring tables were generated and a gRNA sequence may be cross-referenced to predict its on-target activity. Past experimental evidence has revealed that the CRISPR RNP complex can reach cleavage efficiencies of at least 80% for single site modification [83]. However, this efficiency might become problematic when applying a CRISPR/Cas system to modify multiple sites across the genome simultaneously using multiple gRNAs, known as multiplexing, for rapid strain engineering. For instance, the probability of achieving activity at 5 loci simultaneously drops to 32% (0.8^5). Therefore, developing new algorithms

capable of accurately designing gRNA sequences with high activity is important for multiplex CRISPR implementation.

In addition to the on-target analysis, it is very critical to make sure that a gRNA design has limited activity at non-targeted sites. The foundation for such off-target analysis for the spCas9 (Cas9 derived from *Streptococcus pyogenes*) system in eukaryotes was established by Hsu *et al.* [95] and later by Lin *et al.* [96]. Studies building on this foundation generally take two directions: one being the engineering or discovery of alternative Cas enzymes that exhibit lower off-target activity [97], and the other being algorithms that enhance prediction capabilities [96]. Recently developed software platforms have implemented various algorithms to determine off-target sites for a gRNA design, based on either biological principles or training a model to fit experimental data [30,34,98,99]. A collection of these programs is detailed at https://omictools.com/crispr-cas9-category. Unfortunately, some potential off-target sequences are unable to be reconciled by these algorithms that do not account for newly discovered interactions between the target DNA and RNP. For instance, a study by Malina *et al.* shows the presence of PAMs (Protospacer Adjacent Motifs) across the sequence to be inhibitory to Cas9 cleavage [100]. This underscores the need for additional experimental investigations into off-target effects and a robust algorithm to capture these effects for designing gRNAs, especially when applying CRISPR/Cas systems to non-model organisms.

Furthermore, the rapid discovery of novel, diverse Cas enzymes beyond spCas9 requires a flexible and robust algorithm that can accurately predict on- and off-target activities utilizing new CRISPR/Cas systems for genome editing. For instance, the discovery and use of another type of CRISPR class II endonuclease, *Acidaminococcus* sp. Cpf1 (asCpf1), has shown complementary activity to the Cas9 family with a T-rich PAM and “sticky-end” generation [25]. Based on the crystal structure of the asCpf1-RNA-DNA heterotriplex, it is possible to infer important nucleotides within the CRISPR-RNA (crRNA) for catalytic efficiency, and thus develop accurate targeting algorithms for genome editing [24,34].

While previous research has focused on identifying highly active and unique target sites using on- and off-target algorithms to assist in gRNA design, targeting non-unique (repeated) sites with CRISPR tools may lead to some interesting future directions for experimentation [101]. This strategy, called multitargeting, has powerful metabolic engineering and synthetic biology applications, but has not yet been fully explored. For instance, repeated sequences may serve as sites where a single gRNA could induce knockouts across multiple orthologs of a gene, reducing the amount of heterologous machinery required to achieve the same effect. Furthermore, engineering consortia of organisms has been a burgeoning field in the past decade [102]. CRISPR tools can potentially investigate and modify these consortia, which may not be achievable by conventional genetic manipulations. Software with the ability to implement these applications can pave the way for novel metabolic engineering and synthetic biology applications using CRISPR/Cas systems.

In this study, we developed a novel method, named CASPER (CRISPR associated software for pathway engineering and research) that implemented flexible algorithms to guide precise genome editing. Combining both experimental data and newly discovered biological principles, CASPER formulated an improved scoring method for enhanced prediction of on- and off-target activities. This prediction presents "relative" activities of a gRNA design that depend solely on target sequences, and hence are independent of experimental conditions (e.g., growth rate, Cas9 concentrations, and DNA supercoiling). These "relative" activities are different from "absolute" activities of a gRNA design that are condition-specific and must be determined experimentally [87,95]. Further, CASPER expanded novel applications of CRISPR/Cas systems in genome editing including multitargeting analysis (i.e. simultaneous multiple-site modification on a target genome with single gRNA requirement) and multipopulation analysis (i.e. gRNA design for genome editing of a consortium of organisms).

Methods

On-target activity formulation for CRISPR gRNA design

On-target activity is defined as the binding affinity and subsequent nuclease activity of a particular gRNA and endonuclease complex to its matching DNA (or RNA) target sequence. Such activity depends on many factors that have been identified such as the state of DNA supercoils, genome size, Cas endonuclease concentration, and growth rate [87]. These variables change based on the organism and the experimental conditions and thus make it impractical to integrate into an algorithm designed for assessing a large set of non-model organisms and alternative Cas endonucleases. Such factors have the same effect on every target within a genome and therefore have no impact on the difference in relative activity between targets found on the same genome.

To predict the on-target activity of the complex, our developed CASPER method calculates the score $S_{C,P}$ (Equation 2.1) that is defined as follows:

Equation 2.1

$$S_{C,P} = \frac{S_C}{P}$$

where S_C is the CRISPRscan [94] score and P is the penalty score. In Equation 2.1, $S_{C,P}$ has a value between 0-100 that is assigned to a gRNA seed sequence, with higher values indicating higher predicted activity. The CRISPRscan score S_C (Equation 2.2) is defined as follows:

Equation 2.2

$$S_{C,P} = \sum_{i=l-6}^{l+6} S_{C,i}$$

where l is the length of a seed sequence and PAM (see Figure 8 for indexing illustration), and $S_{C,I}$, is the score associated with the mono and dinucleotide features at position I (N_i or NN_i) of the seed sequence that has been experimentally determined to be relevant [94]. Each score, $S_{C,I}$, is the sum of the values of the features appearing at position I, and the sum of all positions' scores (S_C) is normalized with respect to the highest and lowest possible scores to obtain a normalized value from 0 to 100. The values of features at any given position can be seen in the CRISPRscan scoring table[31].

The penalty score, P, (Equation 2.3) is obtained by the combination of the PAM density score, s_{ij}, (Figure 9) and the score S_G (Equation 2.4):

Equation 2.3

$$P = \begin{cases} s_{ij} S_G & s_{ij} > 1 \\ 1 - S_G / 5 & s_{ij} = 1 \end{cases}$$

The score s_{ij} in Equation 2.3 is determined by the number of PAMs present in a given seed sequence, where i is the number of PAMs in the forward sequence and j is the number of PAMs in the reverse complement of the sequence. The score s_{ij} can be looked up in the i_{th} column of the j_{th} row in Figure 9, which is derived from experimental data [100]. Due to the length of the spCas9 PAM (3) and the length of a seed sequence (20), i + j < 8. In Equation 2.3, the score S_G, used to reinforce the importance of guanines and adenines to the stability/instability of the gRNA respectively, is formulated to account for the nucleotide composition of the seed sequence irrespective of position.

Equation 2.4

$$S_G = \sum_{i=1}^{l-PAM} n_i$$

$$n_i = \begin{cases} 1 & if \quad N_i = G \\ 0.5 & if \quad N_i = C \\ -0.1 & if \quad N_i = A \end{cases}$$

The values n_i were derived to maximize correlation between scores and on-target experimental data [94] further described in the Results. A visual summary of the process of obtaining an on-target score for a sequence is shown in Figure 8.

Off-target formulation for gRNA design

In contrast to the on-target activity assessment, the off-target activity is defined as the probability of a given gRNA sequence to interact with a non-matching sequence on the genome. Our developed CASPER method calculates the off-target score $S_{H,T,S}$ as follows:

Equation 2.5

$$S_{H,T,S} = \frac{\left(\sqrt{S_H} + S_T\right) R^2 S_S^6}{4}$$

where the off-target score $S_{H,T,S}$ lies between 0 and 1, with a higher value indicating higher probability of off-target activity, and is determined by combining four subscores S_H, S_T, S_S, and R discussed below. The appropriate arrangement of these scores in Equation 2.5 was determined by employing a genetic algorithm using the Pearson's coefficient between the output scores and experimental cleavage efficiency [95] as the fitness function. A more detailed description of the algorithm is presented in the Results. The subscore S_H accounts for the types of nucleotide mismatches and their location on the seed sequence, and is defined as follows:

Equation 2.6

$$S_H = \prod_{i=1}^{20} M_{ij}$$

where M_{ij} is the element of the Hsu matrix derived from experimental data gathered in Hsu *et al.* [95] represents the index of the mismatch, and j corresponds to the identity of the mismatch (e.g. C with A).

The subscore S_T is derived from the inverse relationship of the proximity of the mismatch to the PAM:

Equation 2.7

$$S_T = \frac{3.5477 - \sum_{i=1}^{20} \frac{1}{i}}{3.5477}$$

where S_T is valued from 0 to 1 with a higher score indicating a higher probability of off-target activity, and i is the index of the mismatch. This score was motivated by previous studies showing the farther a mismatch is from the PAM site, the less likely it is to interfere with activity [49]. As a result, sequences with PAM distal mismatches are more likely to be sites of off-target activity. Summing the values for a mismatch at every location across the seed sequence gives a value of 3.5477, thus this number is used to normalize S_T as formulated in Equation 2.7.

The subscore S_S also captures an inverse relationship of the proximity of the mismatch to the PAM but is formulated using a stepped scale:

Equation 2.8

$$S_S = 1 - \sum_{i=1}^{20} a_i \quad a_i = \begin{cases} 0.1 & i \le 6 \\ 0.05 & 6 < i \le 12 \\ 0.0125 & i > 12 \end{cases}$$

where i is the index of the mismatch, and a_i is defined by a step function. The step sizes were derived to agree with the previous experimental report where mismatches in regions closer to the PAM are more detrimental to activity [95].

Using the knowledge that some gRNAs are more stable/active than others, the on-target activity scores of the target sequence and the off-target sequence are assembled into a ratio, R:

Equation 2.9

$$R = \frac{S_{C,P}^{off}}{S_{C,P}^{on}}$$

where $S_{C,Poff}$ and $S_{C,Pon}$ in Equation 2.9 are the on-target scores for the target DNA sequences appearing at the undesirable and desirable sites of the genome, respectively. A ratio of R < 1 signifies that the sequence of interest (SOI) is more active compared to the potential off-target site, meaning less likelihood of the off-target site being hit compared to the SOI. A ratio R > 1 signifies the reverse, i.e. a highly active off-target site that has an increased likelihood of being hit compared to the SOI. The visual representation of gathering all these scores and combining them for a given two sequence comparison can be seen in Figure 8.

Multitargeting analysis

Multitargeting is the process of editing multiple sites simultaneously across the genome of an organism using a single gRNA. Instead of removing repeated gRNA target sequences, our developed CASPER method stores the data of repeated sequences for further analysis. To generate data of the non-unique seed sequences, CASPER references the set of sequences that appear more than once to a genome annotation file generated by either GenBank (http://www.ncbi.nlm.nih.gov/genbank/) or KEGG (genome.jp/kegg/). These sequences are then sorted based on the number of times they appear in an annotated region. Further analysis can be performed by investigating individual sequences in order to design gRNAs to target these repeated sequences.

Multipopulation analysis

To analyze multiple genomes simultaneously for use in the study and genetic modification of a consortium, CASPER collects the data by identifying targets

across a collection of genomes and performs on- and off-target activity analysis against the entire metagenome. This enables precise gRNA design for accurate genome editing within a consortium. To deal with polyploid organisms, the same algorithm can apply. Sequences shared across genomes can also be identified, thereby applying the concept of multitargeting to populations.

Results and Discussion

Development of CASPER algorithm for enhanced on-target activity prediction

In order to analyze sequences to be targeted by a gRNA/Cas complex, we developed an algorithm for CASPER to predict the gRNA design for the on-target activity (Equation 2.1). It was formulated to incorporate three guiding principles: i) the CRISPRscan features experimentally identified to be present in highly active guide sequences [94], ii) the density of the PAM in question across the guide sequence [100], and iii) the propensity for guanines (and to a lesser extent cytosines) over adenines in the gRNA sequence which has been shown to be a factor in gRNA stability [29,103]. To formulate a combination of these three principles, we combined the normalized CRISPRscan score, S_C (Equation 2.2) and divided it by a penalty score, P (Equation 2.3). The structure of P is determined by whether or not the number of PAMs in the sequence is inhibitory ($s_{ij} > 1$). If so, the value is multiplied by the score S_G; otherwise, it is disregarded, and P purely accounts for the nucleotide content of the sequence (the S_G score). While the values of s_{ij} are derived directly from experimental data [100], the score S_G needs to be constructed *de novo.*

Since previous studies showed guanines were favorable and adenines unfavorable to gRNA stability [29,103], we decided to value guanine most favorably (positive value) and adenine least favorably (negative value). The values assigned to guanine, adenine, and cytosine (Equation 2.4b) were obtained by varying their values between 0 to 1 (with a step of 0.1) and then evaluating which combination of values resulted in the optimal Pearson coefficient between experimental data from Moreno-Mateos *et al.* [94] and the score, $S_{C,P}$.

We applied CASPER's on-target activity algorithm to predict the on-target activities of gRNAs from the experimental study by Moreno- Mateos et al. [94]. This study investigated indel frequency from a pool of 1280 gRNAs in zebrafish single cell embryos at 9-hour post fertilization. The CASPER method provided a minor 2.4% improvement in the R_2 value as compared to the CRISPRscan model (Figure 3A) by taking into account PAM density and position-independent nucleotide content. Such a minor variation underscores the robustness of the CRISPRscan model. This improvement is statistically insignificant from the CRISPRscan model as determined by an F-test between the two data sets ($\alpha = 0.05$). However, CASPER successfully identified 3 out of the 4 largest CRISPRscan score outliers in an experimental dataset of 25 gRNAs (Figure 3A).

CASPER's on-target activity algorithm can be refined with subsequent data sets or further studies on spCas9's underlying mechanism of action. Additional on-target data sets generated by future studies will allow CASPER to develop parameters to assess gRNA activity in novel CRISPR/Cas systems (i.e. diverse organisms and Cas enzymes). For instance, CASPER's on-target activity algorithm can be applied directly to asCpf1 as soon as experimental data of on-target activity becomes available. In the meantime, to evaluate asCpf1 and other non-canonical endonucleases, CASPER focuses on the principles governing PAM density and guanine/adenine content to predict on-target activity.

Development of CASPER algorithm for enhanced off-target activity prediction

The objective of off-target analysis is to determine the sequences most likely to exhibit activity despite not being an identical match to the de-signed gRNA. Many factors contribute to off-target activity, including the number and location of mismatches, the number of times the PAM appears across the genome, and the concentration of the endonuclease. Only those variables relating to sequence identity are considered here, as other factors are considered to be either relatively constant (e.g. Cas9 concentration) or entirely unpredictable (e.g. state of DNA supercoils) when comparing one sequence relative to another. To per-form such an analysis, an algorithm was developed for CASPER to compare all potential target sites in the genome to the SOI. For each comparison, mismatches between the two sequences are identified. If the total number of mismatches across the two sequences exceeds 4 the pairing is given a score of 0, signifying there should be no appreciable activity at the off-target sequence in question. The mismatches are then scored according to Equations 2.6, 7, and 8a. In addition, the on-target scores for the two sequences are obtained and the ratio (Equation 2.9) is also incorporated into a final score, $S_{H,T,S}$ (Equation 2.5). This scoring method improved prediction capabilities by 30.2% in the R2 value (Figure 3B) over the canonical study[36].

The enhanced correlation to experimental values could be attributed to the combination of incorporating experimental data (subscore S_H)[36] and guiding principles derived from the CRISPR RNP complex's mechanisms of action (subscores S_T, S_S, R). A genetic algorithm was employed to determine the optimum arrangement of the subscores (S_H, S_T, S_S, and R) to create a single final score, $S_{H,T,S}$, and to determine if any subscore was redundant. The result from the combinations of these scores were compared to experimental values given in Hsu et al. and the resulting correlation (R_2) was used as the fitness parameter for the genetic algorithm. The algorithm was given free range over modifying the coefficients and exponents for each of the subscores, as well as the general form of the equation, i.e. whether the subscores were added/subtracted or multiplied/divided to each other. The algorithm was initialized with 100 possibilities and the fittest "parents" were chosen for crossover to create a new generation. Each generation spawned 100 children. By running the algorithm for 1,000

generations, the final format present in Equation 2.5 was achieved. To confirm the algorithm was not over trained on the experimental data set provided, a separate set of data[36] was used for determining the correlation of the algorithm's output with experimental off-target activity data (Figure 3B).

Off-target guiding design principles of asCpf1 are similar to those of Cas9 in that the PAM proximal nucleotides are important for binding and enzyme activity, particularly the first 8[25]. Due to the lack of off-target experimental data at the time of this study, asCpf1 off-target identification was simplified to the use of equations 7-9 as off-target data to generate a matrix for Equation 2.6. In addition, using the recently solved crystal structure of the asCpf1-RNA-DNA heterotriplex[24] enables us to exclude the 4 most PAM distal nucleotides of the 24-nucleotide gRNA sequence, as they are oriented away from pairing with the DNA when complexed with asCpf1. The parameters of Equation 2.8 were modified to give Equation 2.10:

Equation 2.10

$$S_S = 1 - \sum_{i=1}^{24} a_i$$

$$a_i = \begin{cases} 0.1 & i \le 8 \\ 0.0125 & 8 < i \le 20 \\ 0 & i > 20 \end{cases}$$

As in Equation 2.8b, the values for a_i in 2.10b are calculated such that a completely mismatched sequence is given a score (S_S) of 0. This equation accounts for experimental observations that the most significant nucleotides in the seed sequence for asCpf1 are the 8 nucleotides proximal to the PAM. In the absence of experimental data, to perform off-target activity analysis on non-canonical CRISPR endonucleases, such as asCpf1, CASPER subscores can be used independently to evaluate off-target activities. The 30.2% increase in R^2 value was evaluated for statistical significance by an F-test using the scores from the Hsu matrix. It was confirmed that the CASPER algorithm represents a statistically significant improvement ($p < 0.05$).

The newly developed CASPER off-target algorithm can be quickly adapted to other systems seamlessly by training the algorithm against a set of experimental data and accounting for observed features associated with the biophysical properties of the CRISPR RNP complex. CASPER therefore provides a foundation for establishing *in silico* off-target analysis for a variety of organism and endonuclease combinations, especially when dealing with non-model organisms.

Development multitargeting analysis

While sequences repeated in a genome have been traditionally discarded due to their inherent lack of specificity, we developed CASPER to exploit these

sequences as potential gRNAs that are capable of targeting multiple sites across a genome with useful applications in synthetic biology and metabolic engineering.

To demonstrate CASPER for multitargeting analysis, we compiled data for 34 genomes of model and non-model organisms with three different endonucleases (spCas9, asCpf1, and spCas9-VRER) to gain perspective on the number of target sites that appear across each genome (Figure 4). Genomes with a low GC content such as *Clostridium beijerinckii* (31.0%) and *C. saccharobutylicum* (29.4%) have a much greater number of target sites for asCpf1 (PAM: TTTN) than that of spCas9 (PAM: NGG). The spCas9-VRER variant (PAM: NGCG) in particular have only a couple thousand sequences appear for the mentioned organisms, a two order of magnitude difference compared to asCpf1. In addition, the greater PAM length of the VRER variant, which one would assume to appear less frequently than the canonical NGG, does indeed present less target sites across all the organisms investigated. These results show that some endonucleases are more useful than others depending on the genome and application. While this paper only presents a representative sample of non-model organisms, the algorithm is capable of analyzing any desired combination of organisms and endonucleases.

To understand where these repeated sequences appear, we mapped their locations on the annotated genomes. Figure 5 reveals that many of these sequences appeared in annotated regions, but a significant number also fell in unannotated ones, particularly sequences targeted by asCpf1 as shown by the significant number and height of black bars in Figure 5B, 5D. This discovery is valuable in that it helps reveal regions of the genome that may be related and subsequently probed through the precise CRISPR genome editing tool. Further analysis into non-unique targets in unannotated regions revealed a significant number of sequences in regions of completely unknown function. Sequences appearing in unannotated regions are particularly useful for two reasons. First, designing gRNAs to target sequences that appear in unannotated regions may reveal information about whether these regions are functional. Second, these sequences can be used for inserting multiple copies of genetic cargo with a reduced risk of unintentional disruption of cellular function. With multitargeting analysis, CASPER facilitates the investigation and manipulation of unannotated regions with the CRISPR/Cas system for rapid genome editing.

To discover cellular processes that lend themselves to multitargeting, we looked further into the identity of the annotated regions in which these sequences were appearing. A cursory examination of the repeated sequences across the genomes revealed that transposable elements appear frequently across all organisms and can provide a powerful platform for integration of multiple copies of genes or entire operons (Figure 6). Across the species investigated, between half and two-thirds of the non-unique sequences found on annotated regions were located on a transposon related feature. The promise of harnessing transposons for genome manipulation has been well documented by the utility of the Sleeping Beauty transposon system (e.g. SB10) [104,105]. In addition to transposable

elements, regions labeled as hypothetical proteins are also common sites to find repeated sequences (Figure 6). Targets within hypothetical protein regions may be useful in a similar manner to those appearing in unannotated regions because the function of these cryptic regions can be systematically investigated.

Of particular interest are also the repeated sequences that appear on gene loci. Figure 6 details the common motifs targeted by non-unique sequences across the genomes of *C. thermocellum, C. aceticum,* and *C. kluyveri.* Interestingly, elements such as the cellulosome anchoring protein of *C. thermocellum*, a defining feature of the organism, is targeted 197 times by repeated sequences, opening the possibility of interrogating this structure with a select number of gRNAs. In *S. cerevisiae,* hexose transporters (416 times) and heat-shock proteins (111 times) are cellular processes that also may be targeted simultaneously by a single gRNA and are of particular interest for metabolic engineers looking to probe the sugar metabolism and environmental sensitivity and adaptation in this organism. In general, CASPER is capable of performing multitargeting analysis in any organism and with any endonuclease desired.

Development multipopulation analysis

When editing a consortium of organisms with a CRISPR/Cas system, one must be mindful of the existence of similar or even shared sequences among genomes. Horizontal gene transfer of a plasmid containing the CRISPR/Cas system can exhibit activity in other species' genomes within a consortium. Additionally, direct transformation of the RNP into a consortium may result in unintended off-target activity in multiple species. It is therefore important to screen all genomes in a consortium for potential off-targets. We have developed the CASPER off-target algorithm to check for off-target sites across genomes in the consortium, thus minimizing unintended activity not just within the targeted organism but the consortium as a whole. In addition, CASPER can identify repeated sequences across the genomes that may lend themselves to multitargeting. This enables the identification of potential sites where multiple organisms in the consortium can be edited with the same gRNA.

To examine more closely the opportunities for multitargeting in a consortium, we ran CASPER against pairs of *Clostridial* species. Figure 7A-B details the number of sequences available for such targeting between the pairing of *C. thermocellum* and a selection of thermophilic species, as well as the pairing of *C. cellulolyticum* with other mesophiles. Additionally, CASPER is capable of performing an analysis on multiple organisms to identify repeated sequences shared across them. For instance, we applied CASPER to identify repeated sites across *C. kluyveri, C. cellulolyticum*, and *C. aceticum* (Figure 7C-D). This consortium was designed to utilize either biomass or a hydrogen/carbon dioxide feed for production of industrially relevant long-chain organic acids. This example demonstrates how CASPER can be used to identify off-targets and multitargeting analysis for genome editing of a consortium.

The number of sequences that may be used to target multiple organisms is quite small when compared to the size of the genomes. Thus, we postulate that the usefulness of multitargeting across organisms lies in niche insertions or deletions, such as those targeting a gene with high sequence similarity across the organisms in question. The opportunities for such manipulation will vary drastically depending on the identity of a consortium. Further, off-target analysis for a consortium will prove crucial in preventing unintended activity. Overall, CASPER's algorithms can facilitate the use of CRISPR tools to edit target genomes within a heterogeneous consortium.

The multipopulation analysis framework can also be utilized to analyze polyploid organisms. If individual allele sequences are known, they can be uploaded as separate 'genomes' and then compared via multitargeting on- and/or off-target analyses. This enables the selection of target sequences both unique and shared amongst alleles, facilitating partial or complete allelic editing with CRISPR/Cas.

Conclusion

The development of CASPER's flexible algorithms for analyzing on- and off-target activity in any organism with any Cas enzyme broadens the utility of CRISPR tools for genome editing of industrially relevant and non-model organisms. CASPER's multitargeting analysis facilitates simultaneous genetic manipulation of multiple loci using a single gRNA with novel potential applications including the investigation of large complex systems (e.g. the cellulosome of *C. thermocellum*) and unannotated genome regions. Further, CASPER's multipopulation analysis provides the ability to investigate microbial consortia and apply the CRISPR tools to perform genetic manipulations on single or multiple organisms within the consortia. We envision CASPER will assist the progress of CRISPR genome editing for metabolic engineering and synthetic biology applications.

CHAPTER THREE
A CRYSTAL STRUCTURE ENERGY CALCULATION ENSEMBLE FOR PREDICTING OFF-TARGET ACTIVITY OF SINGLE-EFFECTOR CAS NUCLEASES

Summary: Understanding the specificity of CRISPR-Cas systems is critical for advancing the technology as a safe genome editing technology. In order to reconcile wide-ranging results of experimental studies attempting to measure on- and off-target activity, this study employs a computational approach using the protein modeling software suite, Rosetta, to create a transition-state-like model that can approximate the thermodynamic and kinetic contributions (collectively coined relative activity) to the protein-RNA-DNA heterotriplex of multiple single-effector Cas nucleases. The resulting score parameter of this simulation out-competes the best performing convolutional neural networks with a Spearman correlation of 0.5 and a best-in-class AUC (area-under-curve) from a relativistic operating curve (ROC) of 0.860. This work represents the first full ternary (RNP, DNA) complex modeling for CRISPR-Cas activity, and the most comprehensive attempt to account for all sources of the relative thermodynamic activity. We hypothesize large deviations between computational results and experimental data are likely the result of experimental conditions and DNA target superstructure *in vivo* (coined absolute activity). To elucidate this absolute activity, we used experimental data using a Cas12 dataset[35] with degenerate guide sequences (multitargeting) to assist in labeling genomic sites as amenable or inhibitory to targeting.

Introduction

The power of CRISPR-Cas (Clustered Regularly Interspaced Palindromic Repeats, CRISPR Associated) as a genome editing platform is reflected in its ubiquity in biological research. Despite the promises (and fears) of a panacea of cures for cancer, super-crops, and embryonic manipulation, the biochemistry of the native CRISPR-Cas proteins exhibit certain degrees of non-specific activity, which is a major concern especially when considering engineering heritable modifications[1,36,48]. This “off-target” activity is a priority for those looking to commercialize CRISPR-Cas technologies for applications sensitive to non-specific changes in the genome. To better understand the underlying mechanisms of CRISPR-Cas enzymatic activity, a combination of single molecule mechanistic studies[39,106-108] and large library assays both *in vitro* and *in vivo* have been performed[36]. Single molecule crystallography and FRET analysis have illuminated kinetic factors important to Cas enzyme recognition, binding, and ultimately cleavage using spCas9[39,107]. These studies have identified the formation of the R-loop (DNA unwinding and RNA binding) as the thermodynamic barrier to Cas protein binding and cleavage. This principle has then been expanded upon in theoretical studies to identify correlation of a parametrized R-loop model to experimental cleavage assays investigating off-target activity[37,47]. Large library

assays have taken a broader view, looking at RNA/DNA sequences for insight into active guide-RNAs for future application[30,32]. These studies provide the foundation for a large number of predictive algorithms based on either first principles from mechanistic insight[87], machine learning from the generated datasets[29,31,36,109], or combinations of the two[43,49,50].

Recently, deep learning methods with neural networks have become popular in predicting on- and off-target activity in datasets[45,110]. While the ability for newer models to avoid false positives in a variety of datasets is promising, the unexplainable nature of these algorithms is currently a detriment to the advancement of better designed guides at the outset. Further, the need for comprehensive datasets with strong control for environmental, condition-specific variables can create variability in the predictive capability of deep-learning algorithms across datasets and Cas enzymes.

Macro trends can be observed in both experimental and computational studies, however, the field is riddled with conflicting information that has led to controversy. Particularly, reports have come out in close succession suggesting either a lack of or ubiquity of off-target activity, creating confusion as to the exact nature of the fidelity of Cas enzymes[34,35,55,97]. At the center of the apparent contradiction is the difficulty of separating *relative* from *absolute* activity. As previously defined[43], relative activity refers to the thermodynamics and kinetics associated with the Cas endonuclease, its guide-RNA and immediate DNA target. This information is known at the outset of experimental design and is the common form of input for generation of correlation functions, including machine learning exercises. The more complicated absolute activity refers to the entirety of factors influencing Cas activity including both controllable experimental factors and elements that are difficult to quantify, but not limited to: DNA supercoiling, DNA chromatin structure, transcription factor binding competition, polymerase binding, and DNA repair machinery protections. Recently, some of these factors (DNA supercoiling and chromatin structure) have been proven to play a key role in the kinetics of Cas enzymatic activity, while others have been known but their role relatively hard to quantify (DNA repair pathways)[37,42,111-113]. Due to the extreme variability in absolute activity, the model proposed in this work focuses on developing the most accurate assessment of relative activity that can be used irrespective of factors influencing absolute off-target activity. Often, the absolute activity factors can be parametrized during the design phase (e.g., titration of Cas9 concentration) that can maximize efficacy while minimizing off-target activity. The ability therefore to have an explainable model that accounts for all relative activity is incredibly helpful for informing the design of effective guides. In turn this information can be used to understand the state of activity across different genomic loci, as absolute activity is reintroduced into the equation, while relative activity is controlled for by the *in silico* model.

Use of computational tools to identify thermodynamic and kinetic parameters of protein interactions is widely used but suffers diminishing returns as the size of the protein and intensity of the simulation is taken into account. For

instance, molecular dynamics simulations (MD) are state-of-the-art for interrogating atomic-level interactions, but are so computationally intense they can only perform simulations up to the millisecond timescale on relatively small proteins (<30kDa) and peptides[114,115]. The Rosetta operating environment was developed in order to perform *ab initio* protein folding simulations to predict the most likely structure[116]. Briefly, it simplifies protein structures into degrees of freedom (e.g. rigid body rotation/translation) and employs physical scoring functions based on the positions of atoms in a Cartesian space. This framework is coupled with a Monte Carlo Minimization to sample conformational states with the lowest-energy state as an approximation for the native state of the structure. The substantial reduction in calculation time from MD simulations enables Rosetta to be performed on a greater number of larger structures. This study seeks to employ this modeling strategy for the investigation of a large number of DNA-RNA-Cas protein heterotriplex structures to identify the thermodynamic contributions of the entire complex and its correlation to on- and off-target activity.

Methods

Structure preparation

The starting Protein Data Bank (PDB) structures of Cas enzymes complexed with RNA and DNA molecules were first downloaded from RCSB. In the case of the Cas12 structures, before the 20-structure ensemble could be generated (Figure 10, the RNA and DNA molecules were extended with the DNA/RNA builder in MoE (Molecular Operating Environment, Chemical Computing Group) to the appropriate lengths used in the experimental studies (25 and 24 respectively). In order to create the mutant structures (Figure 10), the PDB structure was separated into its component chains. Nucleic acid chains were mutated with the *rna.thread.default* Rosetta algorithm developed by the Das Lab[117]. The PDB structures were then recombined with their respective chains and heteroatoms (e.g. Mg, Ca, H_2O). The recombined PDBs were in the case of the on-target structures relaxed then minimized, or just minimized in the case of off-target structures. Truncated structures were generated by removing residues from the minimized structures and then generating a new Pose score for the structure using the scoring algorithm. As Rosetta tags heteroatoms to the first residue by default, the flag “-in:auto_setup_metals” was used for structures with metal ions to account for ionic/covalent interactions between them and the protein residues. The validity of the default multipliers was inspected with a visual check of the ion position in PyMOL to ensure reasonable structure geometry was generated by the relax algorithm.

Rosetta algorithms

The flexible protein modeling software suite Rosetta (RosettaCommons.org) was used to analyze the crystal structures of the CRISPR-Cas systems in this study. To generate the relevant sequences from the experimental datasets, the

rna.default algorithm was used to mutate the DNA and RNA sequences of the starting Protein Data Bank (PDB) structures downloaded from RCSB (Table 1). Three main algorithms were used to compute energy profiles of the PDB structures. Briefly, the *Scoring* algorithm generates Rosetta Energy Units (REUs) for each residue based on the positions of the atoms and 19 sub-score algorithms assessing attractive, repulsive and hydrogen bond forces[116]. The per-residue energies are summed to generate the score of the entire structure. The *Minimization* algorithm chooses a descent vector, calling a variation of the scoring algorithms after each iteration to determine the next vector until it finds a local energy minimum. The algorithm ends when the fractional tolerance (deviation from true minimum value) is met, which was set to 0.0001 in this study. The chosen *Minimization* algorithm "lbfgs_armijo_monotone" uses inexact line search to allow for temporary increases in energy in order to avoid shallow local minima. The final algorithm, *Relax*, calls the *Minimization* algorithm and the Packer alternately. The Packer uses Monte Carlo sampling to optimize the positions of side chains. Because of its stochastic nature, the *Relax* algorithm is used in combination with ensembles of structures in our protocol to obtain a statistically relevant sampling. For the generation of the initial structure relaxation the *Relax* algorithm was repeated 10 times while for relaxation of on-target structures, the algorithm was called 5 times.

Ensemble generation

To generate an appropriate ensemble of structures in the Rosetta computational environment, the protocol of the Rosetta-Vienna ddG method was largely followed[118] and adapted to accommodate the larger structures and thus longer computational times. In summary, the initial PDB structure downloaded from RCSB (www.rcsb.org, Table 1) was relaxed to generate a 20-structure ensemble (Figure 10). The 5 structures with the lowest total energies were used to create a 5-structure ensemble. For each structure in the ensemble, a set of relaxed and minimized structures followed by successive truncation of the RNA molecule were generated (Figure 10).

In order to investigate the effect of larger ensemble size on the correlation to experimental values, a 20-structure ensemble for each on-target structure (Figure 10) was created for each of the 5 ensembles of the 5F9R structure, generating a total ensemble size of 100. The process of off-target structure minimization and subsequent truncation and scoring was followed for each ensemble.

Statistical analysis for correlation to experimental data

For spCas9, the MLE (maximum likelihood estimation) DNA cleavage values from Hsu *et al.* was used[36]. On-target investigation used the CRISPRscan sgRNA library dataset[94]. Cas12 (Cpf1) experimental values were obtained by normalizing read counts in Kleinstiver *et al.* 2016[35] to their on-target values.

Results and Discussion

On-target activity is not predicted by heterotriplex energetics

The first investigation into a protein modeling approach for investigating CRISPR-Cas single-effector activity was to determine potential correlation of the experimentally determined on-target activity with output from Rosetta's Pose calculations. The sgRNA sequences from the CRISPRscan dataset[94] were used as the basis for the training set. These structures were prepared for simulation using the standard *rna_mutate*, *minimiz*e, and *relax* Rosetta algorithms. Using the 19 subscores and total score (sum of 19 subscores) from each simulated structure as parameters, we performed a multivariate regression to identify potential correlation (Figure 11). No significant correlation between measured experimental on-target activity and the energetics of the structure could be detected. This result is in agreement with a variety of studies that point to the dominant role of "absolute" activity factors such as genomic superstructure, cell-state, and transcription control machinery that might either inhibit or enable the Cas protein from binding to the desired target site[38,111,112]. This lack of correlation was further anticipated by the nature of the Rosetta algorithm, because the simulation calculates conformation structure with REUs that may correlate but are not exact thermodynamic values. Finally, the value output from the Rosetta protocol is a final structure value and does not give insight into the transition state of R-loop formation that has been experimentally shown to be the governing factor for CRISPR-Cas activity.

Truncation scoring approximates R-loop formation energetics

To simulate the passage through R-loop formation, we estimated this dynamic by calculating truncated DNA structures in the static Rosetta environment. Each of these structures represents a snapshot of the intermediate binding states that the complex passes through as the R-loop is formed. The differences in the energy values between the matched (on-target) structure and the mismatched (off-target) structures is then calculated and used to approximate the ΔΔG at each stage of the R-loop transition phase. This procedure is used to account for the intrinsic energy change to the system in the Rosetta environment when more molecules are added (Figure 12). To identify only relevant forces and not random differences due to conformational state found with Monte Carlo minimization in distal protein residues, we used energy scores from only the RNA and DNA molecules.

Figure 12 demonstrates this procedure for both spCas9 and lbCas12 structures showing the resulting ΔΔG proxy values for a structure with a given mismatch profile. In many cases the proxy values align with exactly with the location of the mismatch, although destabilization of neighboring nucleotides can also be seen in the lbCas12 triple mismatch result. A distinct advantage of using this method is the inherently agnostic nature to the type of mismatches. In particular, previously developed algorithms[30,36,119,120] could not to systematically account for deletions or insertions, as they are difficult to account for in position matrix. The proxy scores resulting from the truncation exercise identify favorable

interactions, thereby probing movements of the entire DNA and RNA chains that are difficult to account for when relying on index specific mismatch matrices for scoring. Such generalizability is important for detecting the broadest range of off-target activity.

Correlation of ΔΔG proxy values to experimental off-target activity

We hypothesized that the ΔΔG proxy values could serve as an insight into the energy barriers faced by a particular structure as it passes through R-loop binding transition state. As signaled from other studies, the closer the barrier to the initial formation of the R-loop (PAM proximal barriers) the less likely binding and therefore cleavage would occur. This is due to the nature of the DNA unwinding by the Cas9 (and Cas12) protein complex, which starts its unwinding proximal to the PAM sequence. To quantify this effect using our Rosetta generated values, we defined energy barriers as ΔΔG proxy values that are deviated from a defined Z-score from the average across the entire structure (Z = 0.5 or 1.0). In addition, the cumulative effect of multiple barriers should serve as a larger detriment to binding. The larger the positive value of the ΔΔG proxy value provided by Rosetta is likely to indicate a taller thermodynamic wall. Finally, the more stable final configuration is, the greater the likelihood of structures makes it to the final conformation, as suggested by the principles of chemical equilibrium. Using our off target experimental dataset of 712 off-target structures[36], we then investigated various mathematical formulations of our hypothesized contributors to off-target activity.

The single largest contributor to correlation with experimental activity is the cumulative effect of ΔΔG barriers, defined as the sum of values with Z-scores above the threshold (Z>1.0). Further, dividing this by the location of the barrier relative to the PAM sequence weights barriers closer to the initiation of the structure because it is more inhibitory to Cas activity. Further addition of the total structures ΔG, calculated by summing all ΔΔG proxy values, achieves the maximum possible correlation (Spearman) and predictive performance. A natural logarithm on the sum of the ΔΔG proxy values was used to appropriately weigh the effect of the ΔΔG and ΔG scores into one equation due to the approximate exponential rise in relative ΔΔG proxy values as one moves farther from the PAM, i.e., the center of the protein structure opening up to Rosetta larger movements in residues). Predictive performance was calculated by using a cutoff of 0.01% MLE (maximum likelihood estimation) for cleavage, which is used to define the experimental values of the relativistic operating curve (ROC, Figure 13B). The area-under-curve calculation provides a singular metric by which to assess the true/false positive ratio of the Rosetta algorithm and derived equation on a spectrum of threshold values.

Equation 3.1

$$\mathrm{RosCas} = \sum_{i}^{n} \frac{ln\left(\Delta\Delta G_{i,Z>1}\right)}{L_i} + \Delta G$$

$$Z = \frac{\Delta\Delta G_i - \overline{\Delta\Delta G}}{\sigma_{\Delta\Delta G}}$$

$$\Delta G = \sum_{i}^{n} \Delta\Delta G_i \ , n = \ gRNA\ length$$

The final formulation of the equation used to correlate the Rosetta ΔΔG proxy value output, termed RosCas (Equation 3.1).

The results for Cas9 simulation outcompetes the best performing convolutional neural networks with a Spearman correlation of 0.5 and a best-in-class AUC (area-under-curve) from a relativistic operating curve of 0.860 (Figure 13). RosCas is unique because it does not rely on any sort of machine learning (i.e. training) but biophysical and chemical equilibrium first principles. This formulation can be applied for many different types of Cas systems, so long as a viable crystal structure has been resolved.

An important aspect of using Rosetta for assessing thermodynamic properties of structures—in particular snapshots of intermediate states—is the use of an ensemble average of structures to smooth the stochasticity of Monte Carlo algorithm results. To measure the impact of the size of the ensemble, we calculated the Spearman and Pearson correlation values of different ensemble configurations of the computed values to experimental data. Figure 14 shows that averaging the ensembles improves the correlation to experimental data. In particular, averaging the highest performing ensembles may offer slight improvements in Pearson correlation over a full ensemble average, suggesting that given the computational power, by running more ensembles and selecting the top performers before averaging may increase performance of the Rosetta-based prediction presented in this work.

In addition to its predictive power, the data generated through the simulations used by RosCas can also provide greater mechanistic insight into certain off-target pairings. Notably, the effect of a single mismatch is often felt by the proximal nucleotides, which can either result in a local stabilizing or destabilizing effect (Figure 12A). When a significant number of mismatches are located in close proximity, this local effect expands across all nucleotides in question, essentially cancelling out any stabilizing effect of the matched nucleotides in the region (Figure 12B). Further investigations into these truncation diagrams with different datasets and Cas nucleases will undoubtedly provide further insight into the complicated nature of non-specific CRISPR-Cas activity.

Cas12 dataset highlights genomic loci contribution to off-target activity

Due to the success of the modeling exercise with the spCas9 dataset, we then performed our RosCas protocol to two Cas12 (lbCas12, asCas12) structures with guide-RNA sequences from Kleinstiver et al. dataset[35]. This dataset was unique because many of the targeted sequences had degenerate sequences with multiple hits across the genome. It should be noted that only a single RosCas score is

given to the multiple off-target hits because the RosCas protocol only takes into account sequence identity.

The result of Cas12 simulation shows that truncation scoring can approximate R-loop formation energetics (Figure 15B). Destabilization of neighboring nucleotides can be significantly affected even by a single upstream mismatch that is close to PAM. We observed there is no correlation between RosCas scores and experimental values (Figure 15A). The main reason was likely due to the degeneracy of the Kleinstiver *et al.*'s dataset coupled with the large variability of experimental (GUIDEseq[25]) values coming from identical sequence mismatch profiles.

Due to the success of the spCas9 exercise, we sought to employ RosCas as a relative activity normalizing factor to probe the effects of genome location on the absolute activity of a given target. RosCas is not only generalizable to other Cas nuclease structures, but its lack of reliance on dataset-specific training allows it to be used as a relative activity guideline on any experiment spanning multiple experimental conditions and cell lines. To this end, the difference between the RosCas score and the normalized (via on-target) GUIDESeq value for each off-target hit was mapped to determine the targetability of each locus (Figure 15C). This exercise resulting in the successful mapping of the accessibility of genomic sites to Cas12 across the human HEK293T cell line. This novel capability of RosCas, enabled by its incorporation of all relative activity factors, shows the promise of using such assays to gain a greater understanding of genome accessibility. A couple obvious limitations exist that could be mitigated through experimental design. First, the lack of sgRNA tiling within a locus leaves the exact nature of the absolute activity a mystery. By investigating variation across kilobase-scale regions, the resolution of genomic accessibility could be further enhanced. Second, the use of larger variety of guide sequences could prevent confounding effects from on-target normalization. Ultimately, the use of RosCas for predicting relative activity with accompanied mechanistic insight is a valuable complement to experimental off-target activity screens for both validating off-target activity and further investigating CRISPR-Cas interactions with cell lines and genomes of interest.

Conclusions

This work has generated a computational model and subsequent workflow for evaluating off-target activity of any single-effector Cas enzyme. The use of truncated DNA structures to identify energy barriers during R-loop formation is the single greatest contributor to the fidelity of the model. Further, the strength of correlation allows the model presented to be used to interrogate the role of other factors influencing total (absolute) off-target activity by controlling for the relative activity in a precise and accurate manner. Showing this model is reliable with only a small ensemble increases the value of the approach as this reduces computational time, making it an attractive alternative to large-scale experimental screening approaches. This method can be coupled with sequence alignment

algorithms (e.g. BLAST, CASPER) to identify putative off-target sites, adding higher resolution to activity prediction. Finally, this method can act as a complement to unbiased genome-wide off target search to evaluate genomic sites that are more prone to CRISPR-Cas activity as shown in the Cas12 data. Taken collectively, this work results in the expansion of off-target analysis to a suite of unexplored Cas enzymes, both engineered and natural.

CHAPTER FOUR
IN SILICO PROCESSING OF THE COMPLETE CRISPR-CAS SPACER SPACE FOR IDENTIFICATION OF PAM SEQUENCES

Summary: A version of this chapter was originally published by Brian Mendoza and Cong T. Trinh: “In Silico Processing of the Complete CRISPR-Cas Spacer Space for Identification of PAM Sequences”[121]

Introduction

The prokaryotic adaptive immune systems, CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), are near-ubiquitous across the archaea and bacterial kingdoms and exhibit extensive diversity [17]. Despite this heterogeneity, there exists three fundamental biochemical behaviors shared among all CRISPR-Cas systems: acquisition, expression, and interference [18]. The first stage, acquisition, is relatively conserved across species and involves the process by which foreign DNA (or RNA) is acquired and inserted into the CRISPR array. The second step involves the expression of the acquired sequences of the array and processing them into short RNAs (crRNA). Third, the interference stage uses either a single (Class 2) or multi-unit (Class 1) effector complex to bind the processed crRNAs for recognition and cleavage of the foreign DNA/RNA. To avoid self-cleavage of the sequence within the CRISPR array, a short sequence known as a protospacer adjacent motif or PAM, appears just outside the recognition sequence. The PAM serves as a checkpoint to ensure the correct target has been found [3,70,122]. This step is critical to the interference step, as even a perfectly matched guide RNA sequence without the proper PAM sequence will fail to achieve any activity [3,36].

Significant research has been devoted to the classification of CRISPR-Cas systems to help define common functionalities [16,17]. The current system contains two main classes, distinguished by a single (Class 2) or multi (Class 1) protein effector complex and further classified into 6 types and 22 subtypes [17,123]. Experimental studies have shown however that PAM sequences are particularly diverse and may vary from species to species even within subtypes [124]. The knowledge of the PAM sequence is thus critical for harnessing CRISPR-Cas systems, and the current classification system offers incomplete insight into a species’ PAM sequence. As research increasingly branches out from the canonical Cas9 of *Streptococcus pyogenes* and its Type II relatives to discover novel functionalities in alternate CRISPR-Cas systems, the ability to study these systems will be limited to the speed in which PAM sequences can be determined [5,32].

Three common methods exist for determining the PAM sequence and have been used to varying degrees of efficacy: *in vivo* selection [19,21,25,125], *in vitro* assays [25,59,126], and *in silico* analysis [3,127,128]. The first two methods are experimentally driven and provide the most accurate determination of the PAM sequence. However, they require time-intensive experimental procedures, and rely on effective molecular biology tools suited to the organism in question to be previously

developed (*in vivo*), or the ability to clone out and express an active system heterologously (*in vitro*) [70]. In contrast, the *in silico* method uses sequence similarity of existing spacers in an organism's genome to that of invasive genetic elements to suggest a potential PAM sequence [127]. While the method is less accurate and complete than that of its experimental counterparts, it requires simply a query of CRISPR array(s) against potential target genomes, which are readily available in ever-expanding sequencing databases. Although this method has been performed for a number of organisms to elucidate potential PAM sequences, a comprehensive PAM analysis of all available CRISPR-carrying organisms has not previously been performed.

We have recently presented the CRISPR Associated Software for Pathway Engineering and Research (CASPER) for enhanced on- and off-target prediction of gRNA designs and multi-targeting analysis for single and consortium of organisms using various Cas enzymes [43]. In this study, we have developed the CASPERpam algorithm to systematically and comprehensively perform a complete PAM analysis for all prokaryotes in the NCBI database for which sufficient data exists to investigate potential protospacer matches. From the generated PAM sequence database, we were able to infer certain design principles that are relevant for understanding PAM diversity and provide a baseline for further experimental studies including PAM assays. The PAM sequence database and categories have been integrated into the CASPER platform as a new feature. We anticipate the CASPERpam tool will enable further research and engineering of prokaryotes using their native CRISPR machinery and aid in the development of novel CRISPR-Cas systems for heterologous expression in eukaryotes.

Methods

This section presents the CASPERpam algorithm for processing the complete CRISPR-Cas spacer space to identify PAM sequences to complement the resources of CASPER [43] for research in novel CRISPR-Cas systems. The CASPERpam algorithm contains 5 steps for spacer collection, genome collection, and PAM detection and analysis (Figure 16).

Spacer collection for PAM analysis

Based on the 720,391 spacers detected across 42,352 CRISPR arrays from 30,389 assemblies collected and analyzed by Shmakov *et al.* [129], a total of 26,364 hits were detected by a BLASTN search across the viral (13,885 assemblies), plasmid (11,218 assemblies), and bacterial/prokaryotic (129,209 assemblies) genome databases of NCBI, using the 95% sequence identity and 95% coverage constraints to gather only hits with sufficient similarity (Step 1, Figure 17). Self-hits were removed by discarding protospacers with identical assembly and location coordinates as the spacer. There were no spacers that generated multiple hits. To identify a PAM sequence, we collected the protospacer sequence and included an additional 10 nucleotides on either side (the "collected" sequence). Because the protospacer sequence itself may only be a partial match to the spacer, we

extrapolated endpoints of the spacer within the collected sequence. Specifically, the spacer and collected sequences were aligned by finding the largest perfectly matched sequence shared between the two sequences. Then the indices of the spacer sequence's start and end within the collected sequence were used to define the locations of the flanking sequences.

Genome collection for PAM analysis

From the spacer collection suitable for PAM analysis, we could narrow the number of assemblies from 30,389 to 6,955 (Step 2, Figure 17). Since multiple assemblies can belong to the same species, we clustered them at the species level to create a total of 1,493 species (Step 3, Figure 17). By choosing species with at least 2 spacer hits, we further reduced the number of species from 1,493 to 1,049 for which we could search for PAMs (Step 4, Figure 17). In our analysis, strains and substrains were grouped together under a common species. While in some rare cases it is possible that PAM sequences might differ between strains and substrains due to random mutation or perhaps horizontal gene transfer of CRISPR-Cas machinery among strains [16], we would not have been able to perform further analysis in a vast majority of species due to the limited number of protospacers belonging to each substrain.

PAM detection and analysis

The Cas type was used to determine whether a PAM was located 5' (upstream) or 3' (downstream) of the protospacer (Step 5, Figure 17). For type I, III (rPAM), IV, and V systems, the PAM has been either shown or inferred to occur 5' on the same strand of the protospacer sequence [12,130,131]. Type II systems, however, have PAMs 3' along the same strand of the protospacer sequence. Each species was assigned a CRISPR-Cas system based on the data provided by Makarova *et al.* [16,17]. For those species that did not appear in the prior classification studies, the genus was used to match the species to a relative and its Cas type was inferred from the type of its relative.

To obtain the PAM for each of the 1,049 species, bitwise scoring [132] was used to determine the important positions (Step 5, Figure 17). Positions with an R-score higher than a half a standard deviation above the mean were considered significant. As an additional constraint, the R-scores for positions farther than 5 nucleotides away from the spacer sequence were subject to the more stringent requirement of full standard deviation above the mean to further filter out noise from positions unlikely to contribute to the PAM sequence. For Type II-C systems where nucleotide positions 5-8 often contain the PAM sequence, the R-scores were high enough such that the full standard deviation requirement did not alter the predicted PAM sequence. Within each significant position, if a nucleotide had a relative frequency higher than 0.5, it was considered to be the consensus nucleotide. If no nucleotide was higher than 0.5, the nucleotide(s) that had relative frequency greater than 0.25 were considered to be part of the PAM.

The R tool ggseqlogo was used to create a bitwise graph for each species to determine the PAM from the appropriate flanking sequence (5' or 3') [132]. In many cases, sequences obtained were shorter than 10 nucleotides due to the spacer index starting in the flanking sequence and/or appearing at the end of an assembly or near an ambiguous sequence (set of 'N's). To satisfy the symmetrical sequence length requirement of ggseqlogo, flanking sequences smaller than 6 were discarded and the remaining were truncated to the length of the smallest sequence.

Computation

NCBI servers were used solely for query purposes (https://www.ncbi.nlm.nih.gov). All genomes needed to identify flanking sequences of the protospacers were downloaded from NCBI (RefSeq/GenBank) onto an external hard drive (as of 11/24/2017). Python (ver. 3.6) was used to process the data and sequences. R (ver. 3.4.3) statistical programming language was used to run ggseqlogo and generate the sequence bitwise plots [132]. All computational work was performed on a Macintosh MacBook Pro computer (Apple Inc., 2013). Code is available at https://github.com/TrinhLab/CASPERpam.

Results and Discussion

CASPERpam is a valuable tool when attempting to harness the power of CRISPR-Cas in a poorly studied organism. It can be used as a starting point for determining the PAM sequence necessary for interference with the native Cas machinery, which may be more efficient than heterologously expressed Cas proteins and arrays [128]. It also provides another level of classification that can be used as a benchmark for comparing and understanding functionality with other systems with similarly predicted or experimental PAMs. Our results highlight the algorithms utility in these areas and offers unique insights as a result of its large-scale analysis.

Predictive PAM sequences match experimental evidence

To validate the CASPERpam algorithm, we compared the predicted and experimentally derived PAM sequences for a set of 12 species spanning both CRISPR classes and multiple types/subtypes (Figure 18). Our result shows that CASPERpam could generate the exact match (green boxes) for the experimentally derived PAM sequences in 3 of the 12 species including *P. aeruginosa* (Figure 2A), *Acinetobacter baumannii* (Figure 18B), and *E. coli* (Figure 18C). In addition, we found strong correlation (blue boxes) between the experimental and predicted PAM sequences of 4 species: *S. pyogenes* (Figure 18D), *C. difficile* (Figure 18E), *L. bacterium* (Figure 18F), and *B. halodurans* (Figure 18G). An additional two species had weak correlation (yellow boxes): *S. thermophilus* (Figure 18H) and *T. denticola* (Figure 18I). Weak correlation species are defined here as partial matches where half or less of the predicted PAM matched with the experimentally derived sequence. Finally, we did not find a match (red boxes) for the 3 remaining

species: *N. meningitidis* (Figure 18J), *C. jejuni* (Figure 18K), and *S. aureus* (Figure 18L).

In our analysis, we observed that the number of complete and strong partial matches appeared in species that have relatively high protospacers (hits) and consensus Cas-types (Table 2). In contrast, the non-matches correlated to a smaller number of protospacers. A notable exception is *Bacillus halodurans*, which despite a relatively low number of protospacers returned a strong match to its experimental PAM. Interestingly, type II species seem particularly prone to inaccurate predictions of the PAM sequence, despite large amounts of protospacers and a consensus Cas-type in certain scenarios. Because sequencing results belong to existing phages and other mobile genetic elements, i.e. those that have survived and escaped CRISPR-Cas interference, protospacer-carrying phages may have acquired favorable mutations in the PAM identification region, contributing to an incorrect PAM sequence being derived by CASPERpam. This scenario may be exaggerated in with type II systems due to the significant differences in the mechanisms of both acquisition and interference between type II (Class 2) and other Class I systems [6, 12-14][122,133,134]. Further studies on common means of escape (e.g. anti-CRISPRs) could further elucidate this phenomenon [135,136].

Overall, the CASPERpam algorithm provides putative PAM sequences with confidence of prediction (PamCoS) scores for 1,049 species, revealing consensus motifs particularly among Class 1 systems. The result also highlights the importance of a more quantitative metric to determine whether the information available for a given species is sufficient to predict its putative PAMs, as presented in the next section.

Quantification of the quality of in silico PAM prediction

To quantify the accuracy of PAM prediction for a given species, we developed a metric, the PAM confidence score (PamCoS), to determine whether the information obtained from the database (NCBI RefSeq) is of sufficient quantity and quality. PamCoS is calculated as follows:

Equation 4.1

$$PamCoS = h \times \left(\frac{C_m}{C_T} \times \frac{1}{N_S} - H_{avgdev}\right)$$

In Equation 4.1, the value *h* represents the number of protospacers (hits) from all the spacers. C_m and C_T refer to the number of detected Cas arrays that match the consensus type for a designated species and the total number of arrays, respectively[17]. The consensus type is defined as the Cas subtype that appears the most. For example, an organism with three arrays, two of which are of Type I-B and one Type II-A would have a consensus type of Type I-B with $C_m/C_T = 0.67$. $N_s = 1$ for a species with a well-defined consensus Cas type. However, if a species is not found in the PAMCoS classification database[121], the value of N_S is increased by 1 for every level up in the phylogenetic tree that the species must be traced until

matching a relative. The value H_{avgdev} is the average deviation of the Shannon entropy term for each of the positions deemed as significant contributors to the PAM sequence (denoted by the subscript "sig"). H_{avgdev} is calculated as follows:

Equation 4.2

$$H_{avgdev} = \frac{\sum H_{i,sig} - H_{avg}}{n_{sig}}$$

Equation 4.3

$$H_i = -\sum f_{a,i} \times log_2 f_{a,i}$$

In Equation 4.2, $H_{i,sig}$ is the Shannon entropy of a significant position; H_{avg} represents the average Shannon entropy across all positions in the flanking sequence; and n_{sig} is the number of significant positions. The $f_{a,i}$ term in Equation 4.3 is the relative frequency of a given nucleotide "a" at position i.

We used PamCoS to score all 1049 species in the CASPERpam produced database (Figure 17 and Table 2). The species with the top 40 PamCoS scores are presented in Table 2, along with a select number of additional species for which an experimentally derived PAM sequence is available. The relative PamCoS score is shown in this table, where the scores have been normalized (0 to 100) based on the highest PamCoS score of *Pseudomonas aeruginosa* (2851.61) in the database. High PamCoS scores correlate well with accurate predictions by CASPERpam of the species with experimentally derived PAMs, with the exception of type II species. It does seem however that after a certain PamCoS threshold, the prediction for type II species becomes effective. For instance, *S. agalactiae* and *S. suis*, two species with type II-A CRISPR-Cas systems, have PamCoS scores over 400 (Table 2) and exhibit the NGG motif as their predicted PAM, the same as their experimentally proven close relative *S. pyogenes*. This suggests that a PamCoS score of 400 can be used as a minimum threshold to guarantee accurate PAM prediction for all Cas types. While only 8 species in the NCBI RefSeq database currently satisfy this criteria, further sequencing efforts of microbial genomes will aid in providing the necessary information for accurate *in silico* PAM prediction.

To assign a more flexible cutoff for the currently available data and using the PamCoS scores of species with known PAMs, a score above 10 (0.35 relative PamCoS to *P. aeruginosa*) seemed a reasonable threshold for PAM prediction. This however is not a strict cutoff as the prediction of the PAM for *B. halodurans* proves lower PamCoS scoring species may still lead to reasonably accurate PAMs. Lower PamCoS rated species must be viewed with some skepticism, however. Due to the smaller number of protospacers, even when accompanied by relatively low Shannon entropy values, these predicted sequences are more sensitive to false assignment of significant positions and nucleotides. Further, false sequence noise may be compounded by potential Cas type ambiguity, as a

non-consensus Cas-type hit may be incorporated into the PAM sequence prediction.

Overall, use of the PamCoS score enhances *in silico* PAM analysis. Notably, PamCoS allows the PAM sequences produced by CASPERpam to be assessed for their relative accuracy and aid in narrowing the search space of sequences for experimental validation. For instance, species with low PamCoS scores should be validated with a comprehensive *in vitro* or *in vivo* methodology, using the results of CASPERpam as supplemental validation of experimental discovery. On the other hand, species with relatively high PamCoS scores may allow the CASPERpam prediction to be supplemented with only minimal experimentation to provide confirmation of the predicted sequence.

Based on CASPERpam analysis, the number of species with PamCoS scores over the threshold value of 10 (221) represents a very small portion of the tens of thousands of bacterial and archaeal species that have been sequenced. It does however represent an order of magnitude increase in the number of CRISPR-Cas systems that have had a PAM sequence predicted or validated. The result provides a starting point to harness the extensive diversity present across the CRISPR-Cas space.

Of the 221 species above the threshold value, 78% had a Class 1 system, an understudied relative to their more popular single effector Class 2 cousins. Even though the Type I system appears to have a greater variety in their PAM sequences [137], only a select number of nucleotide patterns of PAMs are commonly used. These select PAM motifs are not confined to a single Cas type, rather they are spread across systems (Figure 19A). One potential explanation is that the PAM motif of Class 1 multi-effector systems is likely subject to some degree of horizontal gene transfer. Particularly, the Cas8/10 and Cas1/2 proteins, likely responsible for PAM recognition during the interference and acquisition steps respectively, may be traded within microbial communities, creating within a species hybrid Class 1 systems with a clearly assigned Cas-type using a PAM largely associated with another [15,17,138].

PAM prediction highlights diversity and patterns within and between Cas types

To analyze the PAM diversity across different Cas-types, we gathered all PAM motifs of the 221 species above the threshold value and grouped them into families of closely associated motifs. We identified common patterns and assigned 5' and 3' PAM sequence categories and/or subcategories to these sets for 218 species (Figures 19B and 19C). Each category/subcategory is defined by a consensus motif that captures the general pattern. Three species from the original set of 221 species, including *Proteus mirabilis, Bifidobacterium longum,* and the *Eubacterium* sp., were removed from the analysis because CASPERpam could not assign either a consensus Cas-type or the PAM motif was clearly the result of multiple Cas-types.

The 5' PAM sequences contain 8 categories (J, K, L, M, N, O, P, and Q) (Figure 19A, 19B). The K and Q categories represent 61% of PAMs located on the 5' end. The 'CC' PAM of the Q category and 'TTC' PAM of K are heavily represented (Figure 19A, 19B). These groups are disproportionately associated with types I-F and I-B respectively; however, they are not exclusive and also appear in species with other Cas-types. The PAM sequence space of the 170 species with reliable 5' PAMs (as defined by the confidence score) can be seen mapped together in Figure 19A.

It is striking to observe that none of the Cas types nor subtypes map to a unique PAM category, with the exception of a few types with less than 5 species represented in the sample (e.g. Cas-III-B and Cas-I-A). This result supports the view that the current CRISPR-Cas classification method is imperfect likely due to the CRISPR-Cas systems' similarity and modularity prone to horizontal gene transfer [17]. It also prevents a simple classification based on sequence homology or phylogenetics from perfectly partitioning the heterogeneity of these systems.

The 3' PAM sequences consists of 3 categories (X, Y, and Z) with the X category broken into a further 4 subcategories (X1, X2, X3, X4) (Figure 19C). Among the 3' PAM sequences, the category Y, with the NGG consensus motif, is quite common, representing 25% of identified PAMs (Figures 19C, 19D). The category X, although representing two-thirds of identified PAMs, has a much more broadly defined consensus motif $(N)_nW$ and lent itself to subcategorization of which none reached a representation comparable to the Y category (Figures 19C, 19D). The 'NGG' PAM found in the Y category is the most represented PAM motif, but CASPERpam only identified two such species, with an exact 'NGG' PAM. Other PAMs in the category were defined more broadly, with adenines and sometimes thymines also highly represented at the 2 and 3 positions creating an NRR, and NKK PAMs. This result is consistent with experimental evidence that type II-A systems such as the canonical *S. pyogenes* exhibited some affinity for a 'NAG' PAM. Even though this detected promiscuity may be overstated by the *in silico* analysis, it highlights the importance of future experimental studies on species with a PAM falling in the Y category to focus on the relative affinity for guanine over adenine or thymine.

The PAM categorization exercise of the highest PamCoS scoring organisms is useful in highlighting both heterogeneity of PAMs as well as some clear categorical delineations, particularly for Class 1 systems (namely K, L, M and Q). These categories can be used to reinforce or place further doubt upon the CASPERpam predicted PAM motif of organisms with substantially lower PamCoS scores. For example, despite a low relative PamCoS score of 0.105, the putative PAM for *A. suryakundensis* is TTN, which agrees well with other Cas I-B systems as a group K motif, lending further confidence to the prediction.

Based on the 5' PAM space and the X and Z 3' PAM categories, we clearly observed that there exists a high level of non-specificity in many PAM interacting domains. It cannot be discounted that this diversity may be underrepresented by the *in silico* technique, as only those with particularly effective targeting will appear

across the targeted genomes [70]. On the other hand, PAM sequences are likely to be mutated for phages to survive, and thus sequencing assemblies may not represent the true form of the PAM sequence, but rather an overrepresentation of a random set of mutated sequences despite a close spacer match. Such a hypothesis would result in the *in silico* process revealing greater non-specificity than would be shown experimentally. Likely, these two effects are both present, and result in some overly specific as well as some overly promiscuous sequence predictions from the *in silico* method.

Conclusion

To harness a large, diverse class of alternative CRISPR-Cas systems outside of the group of well studied Cas9 and Cas12 variants, it is critical to identify PAM sequences for characterizing these novel CRISPR-Cas systems. Development of CASPERpam provides a powerful tool to predict PAMs for a given species and allows subsequent experiments to change from a determination to a validation study. While some previous studies have performed such an *in silico* PAM analysis for a select number of species to guide experimental studies of native CRISPR-Cas systems, this study has performed an exhaustive PAM analysis for every CRISPR-Cas carrying bacterial and archaeal genome in the NCBI database. As research moves into investigating a greater number of CRISPR-Cas systems with novel functionalities, CASPERpam will be a useful bioinformatic tool for those seeking to understand and harness the diversity of CRISPR-Cas systems.

CHAPTER FIVE
DEVELOPMENT OF CRISPR-CAS ANTIFUNGALS

Summary: A version of this chapter is to be submitted to PNAS under the title: “Potency of CRISPR-Cas antifungals is enhanced by co-targeting DNA repair and growth regulatory machinery at the genetic level” by Brian Mendoza, Xianliang Zheng, Jared Clements, and Cong T. Trinh.

Introduction

Fungi represent a particularly resistant class of pathogens that are the source of dysbiosis in a variety of societally relevant hosts and environments[139,140]. Damage to agricultural and sensitive ecosystems due to climate change have exacerbated the propensity and subsequent damage of outbreaks of *Magnaporthe oryzae* (rice and wheat pathogen) and *Ascophaera* and *Aspergillus* genus of bee broods[141]. Recent examples of fungal outbreaks including the *Candida auris* “superbug” scare in 2019 and has prompted increasing research into these relatively rare, but deadly pathogens[142]. Even common fungal pathogens such as *Candida albicans* can cause significant damage to the host and make them susceptible to dangerous secondary bacterial infections (e.g. methicillin resistant *S. aureus,* MRSA)[143]. These co-infection scenarios are especially problematic as the fungi provide a safe haven for the bacteria to propagate, protecting them from treatment with traditional antibiotics.

At the center of fungal persistence is the robust and complex regulators of growth that allow them to adapt to a multitude of environments. In particular, hyphal growth stages, where multinucleated cells elongate and form thick extracellular matrices, reduce the efficiency of chemical antifungal treatments[143-145]. Current research has identified many key genes that regulate such growth with the hopes of employing novel treatments to re-sensitize fungi to traditional antifungals[146]. While this approach can identify relevant targets, the ability of deploying such systems *in vivo* and with successive chemical antifungal treatment is complicated, not least due to the fact that many of the current antifungal therapies have moderate levels of toxicity to their animal/plant hosts. Furthermore, fungal infections often occur in tandem with a bacterial outgrowth, resulting in a multi-kingdom pathogenesis. The fungal infections provide a favorable environment in which the bacteria can grow, protecting them from standard antibiotics. This symbiotic resistance mechanism is particularly recalcitrant, requiring treatment of the fungal infection before the bacterial co-infection can even be addressed.

CRISPR-Cas, the method usually employed for knocking out key genes in the growth regulation cycle, can also be employed as an antifungal in and of itself, and has shown such promise in bacterial systems[62,63,76,139]. Unfortunately, these studies usually rely on the high lethality of DNA double-strand-breaks (DSBs) induced by the Cas enzyme. Such a method is ineffective in fungi, as the robust eukaryotic DNA repair mechanisms can easily counteract DSB activity rendering

Cas enzymes only mildly (if at all) toxic[48]. Therefore, targeting essential genes with such machinery and ensuring their complete or near-complete knockout is vital to the efficacy of a CRISPR-Cas based antifungal. The cataloged diversity of preferred repair mechanisms and their dependence on growth stage in the various fungal pathogens makes understanding such interactions when faced with CRISPR-Cas antifungals a large combinatorial problem, prompting the use of a model organism to first understand the significance of this relationship.

The yeast *Saccharomyces cerevisiae* is a widely studied model organism for understanding yeast genetics and biology. Due to its highly curated genome, the essentiality of its genes is well documented, providing a baseline for understanding the effectiveness of a transiently expressed CRISPR-Cas system when targeting such loci. In addition, much is known of its DNA repair pathways and growth cycles, which exhibit many similarities to the clinically relevant pathogenic yeasts of the *Candida* genus[147,148]. Through screening multiple target sites across essential genes, this study investigates promising CRISPR-Cas sequences for use as antifungals. In addition, multiplexing with DNA repair (i.e. "defensive") genes, hopes to improve upon the overall lethality of CRISPR-Cas by removing the ability of the organism to effectively repair the site of the DSB, resulting in increased lethality of essential gene targeting.

Materials and Methods

Strains and Plasmids

The vectors used were adapted from DiCarlo *et al.*'s study[10]. In all studies (excluding the library verification) a two plasmid system was used, which comprise of 1) a uracil selectable marker on the plasmid carrying the sgRNA under the SNR52 promoter on a p426 backbone and 2) Cas9 under the galactose inducible promoter, GAL1-10 with a leucine selectable marker on a p415 backbone. The derived vectors containing the sgRNA sequences were built via Phusion PCR in a 40 µL reaction using the p426 plasmid from DiCarlo *et al.* as a template, sgRNA Gibson primers with homology for the template and the insert, and a generic primer with homology for the plasmid. Inserts were constructed using the same template and complementary primers to the backbone (i.e., one unique to an sgRNA and one generic). Linker sequences were used to construct an insert with one sgRNA sequence followed by the SUP4t terminator, a short spacing sequence, and then another SNR52 promoter with the second sgRNA sequence on the end. These inserts were inserted into the p426 backbone using the same Gibson assembly protocol described. Transformations were performed with Top10 *E. coli* using 5 µL of reacted Gibson Assembly. Upon addition, cells were left on ice for 3 hours followed by a 45-seconds heat shock at 42˚C. After heat shock, cells were inoculated in 1 mL of LB and shaken for an hour at 37$_0$C. Cultures were spun down and concentrated for plating on 100 mg/L Amp plates, where they grew overnight at 37˚C. Plasmids were run on a gel electrophoresis and confirmed by Sanger sequencing. About 1 µg of each confirmed plasmid with the Cas9 carrying plasmid

were transformed into *S. cerevisiae* BY4741 using electroporation (2.1kV, 200 ohm, 25 uF).

Construction of the tetracycline inducible single plasmid utilized the p426 backbone with the SNR52 promoter on the sgRNA operon, but the GAL1-10 promoter controlling Cas9 expression was swapped for the stronger TEF promoter with a TetO sequence just upstream of the start codon. The sgRNA cloning procedure was also simplified by removing the operon and adding *KpnI* and *Avr*I sites in its place. Operons expressing sgRNAs were then constructed by overlap extension PCR using the primers containing the sgRNA sequences followed by digestion with *KpnI* and *XbaI* and then ligation into the *KpnI/AvrII* cloning sites. This cloning strategy could be reconstituted by use of an upstream overlap extension primer that recreated the site on the inserted operon, and hence allowed for subsequent addition of sgRNA operons via simple digestion and re-ligation of the plasmid libraries.

Media and culturing conditions

The parent strain used in this study was S288C BY4741. Strains carrying the Cas9 and sgRNA plasmids were grown up in selective media with the appropriate double auxotrophy (SC-Leu$^-$Ura$^-$) and the desired carbon source (20 g/L glucose for control, 20 g/L galactose for induction of Cas9). For agar plate serial dilution assay, strains were grown up in SC-Leu$^-$Ura$^+$ glucose media overnight until exponential phase was reached. They were then washed 1x with PBS and diluted 10x, 100x, and 500x in PBS before spotting 1 μL of sample on the agar plate. For growth assays in broth, culture prep followed the same incubation followed by PBS wash. Cultures were then diluted with culture media (SC-Ura$^-$Leu$^+$ carbon source) to the appropriate starting OD (0.05 or 0.2). Strains were cultured in a Duetz plate with a working volume of 500 μL. OD measurements were taken with a plate reader (Bio-Tek).

Flow cytometry

To perform viability assessment with flow cytometry, samples were taken and diluted to within a range of OD 0.1-1.0 to ensure accurate counting by the flow cytometer. Samples were then further diluted 20x to a total volume of 200 μL by the Guava ViaCount staining solution (EMD Millipore #4000-0041), mixed by pipetting, and allowed to sit in the dark for 5 minutes. Samples were taken in a 96-well round bottom plate and placed in the Guava EasyCyte 6HT flow cytometer. Gating voltages were calibrated on a 1:1 mixture of exponential phase S288C strain and a 20-second microwaved sample of the same strain.

Library generation and sequencing

A library of 184 targets was generated using the CASPER (Casper2.org) library generation feature. Oligos were ordered and annealed with their reverse complement, then ligated into pBM1000 via golden gate assembly with BsmBI. The ligation mixture was transformed using NEB DH10B High Efficiency *E. coli*

(C3019I) using the heat-shock protocol described by NEB with the following modifications: after recovery cells were added directly to selective outgrowth media (LB + 100 μg/mL ampicillin). 5 mL cells were cultured for 18 hours followed by plasmid extraction with Omega Biotek E.Z.N.A plasmid mini kit (SKU:D6942-01). Transformation into *S. cerevisiae* was performed by electroporation described above with the following modification. After the 1-hour recovery in YPD, cells were washed twice with PBS and added to a 250 ml baffled Erlenmeyer flask with 25 mL selective media (SC-Uracil) with and without 5 μg/mL tetracycline in triplicate. Samples were taken at 24, 48, and 72 hours post inoculation. Colony PCRs were performed on the extracted samples, gel confirmed, cleaned and prepped for run on an Oxford Nanopore MinION R9.4.1 flow cell using the LSK109 ligation sequencing kit with the native barcoding kit (EXP-NBD104). Results were analyzed with a custom python script available on GitHub (github.com/TrinhLab/OxfordAlgorithms/) to obtain library sequence distribution.

Results and Discussion

Screening target combinations in S. cerevisiae

The robust DNA repair mechanisms of eukaryotes present a significant hindrance to the development of CRISPR-Cas based antimicrobials. To investigate the capability of multiplexing sgRNAs targeting both essential and defensive machinery, we selected a set of three DNA repair genes (RAD51, RAD52, LIF1) that are key elements to either homologous recombination (HR) or non-homologous end-joining (NHEJ) pathways. We then combined these genes with a select subset of essential genes (n=25) that span multiple functional domains including central metabolism, replication, and transcription regulation. A serial dilution assay of strains carrying sgRNAs targeting either essential, defensive, 2x essential, 2x defensive, or an essential and defensive sgRNA revealed that double targeting with an essential and defensive gene proved most effective at inhibiting growth. Figure 20 details the rationale for co-targeting essential and defensive machinery for CRISPR-Cas antifungal design. In the case of sgRNAs targeting a single essential locus, the robust DNA repair machinery of eukaryotes can repair any breaks in at the essential locus, resulting in only a mild toxicity. Targeting a single defensive gene can result in a knockout but one that still results in a viable phenotype. Dual or multiplex targeting essential genes can further increase toxicity but fall short of the efficacy found by targeting essential and defensive machinery simultaneously, which leaves the cell unable to reliably repair a break at the essential gene site.

Characterization of high-performing CRISPR-Cas targets

Despite positive results with spotting, further characterization of 18 of the most promising strains was performed in liquid culture to identify if the lack of growth on agar could be recreated in the more forgiving liquid monoculture conditions. The added dimension of growth allowed escaped cells to propagate through the entire

culture, exemplified by the OD (600 nm) growth curves in Figure 21B. Some strains however exhibited flat growth curves, suggesting greater efficacy of CRISPR-Cas growth inhibition. These findings led to the investigation of the state of viability of these cultures.

To probe the viability of the cultures, a reseeding of the high-performing Vip60 (RPS21, RAD51) culture was performed to see if the viability could be recovered. Indeed, after 48 hours, placement of the induced strain in non-inducing media (glucose as carbon source), was able to recover growth, suggesting that the CRISPR-Cas is rendering cells dormant. This inference is further supported by flow cytometry viability assays. Figure 21C highlights the viability results from 6 different strains under 4 different conditions. Strains in the control environmental scenario (glucose) grow quickly in culture, and by the 48-hour mark when the sample was taken have exited exponential phase, resulting in a high count of viable cells but low percentage viability (~25%). Most strains in the galactose induction media exhibit a profile similar to that of the no-carbon growth condition. In this case, viability of the cells remains high, but growth is arrested, leaving the culture in a dormant state comprised of persistent cells. In fact, the galactose induction of Cas9 does seem to slightly impact the overall viability, particularly in Vip60 and Vip57. This implies that the CRISPR-Cas antifungal design broadly creates a dormant phenotype but select combinations of targets can act fungicidal. Overall, the ability to induce cell death is less pronounced by CRISPR-Cas as it is using a traditional geneticin (G418) antibiotic treatment, although some combinations get close on either percent viability (Vip52, Vip76) or viable cell count (Vip60) metrics. Strains that perform less well exhibit viability profiles similar to that of growth of that of Vip46, which targets 2 defensive genes (RAD51/RAD51) in galactose media. In Vip46 there is a slight reduction in both cell viability and total cell count even when compared with the no carbon scenario, suggesting some toxicity, although the degree to which this is the case is much less than 5 the co-targeting essential/defensive strains, particularly on the total viable cell count metric. Viability assay via flow cytometry indicates that identifying target combinations that cause apoptosis rather than predominantly cellular dormancy will be critical in future developments of CRISPR-Cas antifungals.

High cell-loading results in more pronounced community escape

In order to understand the robustness of growth suppression in the best performing strains, we differentially seeded cultures with either an OD of 0.05 (~$5.5x10^5$ cells/ml) as in previous experiments, or 0.2 ($2.5x10^6$ cells/ml). The increase in cells resulted in a pronounced increase in growth rate in all media types, including growth in the antibiotic (G418, Geneticin) and the glucose control medium (Figure 22). The resistance to G418 is likely due to the greater number of cells producing a greater effective concentration of aminoglycoside 3'-phosphotransferases, which provide the innate resistance via degradation of the antibiotic[139].This broad escape phenomenon induced by higher initial cell concentration—and therefore higher loading of enzymes responsible for counteracting antimicrobial activity—suggests

that ongoing kinetics of the culture is also a major driving factor of the escape to CRISPR-Cas antimicrobials, in contrast to an initial inhibition/inefficiency of the machinery that eventually allows for escaped cells to proliferate. Further, the more pronounced rate of growth in the CRISPR-Cas inducing media compared to G418 antibiotic further demonstrates the reduced potency of the CRISPR-Cas system compared to traditional antibiotics. However, the differential inoculation results underscore the similarity in escape profiles between traditional small-molecule antibiotics and CRISPR-Cas, requiring deeper investigation into the roles of cellular repair machinery and stress response mechanisms to this phenotype.

Targeted amplicon sequencing reveals escape genotypes

As part of our in-depth characterization, we have performed both amplicon and plasmid re-sequencing using Oxford Nanopore sequencing with the MinION device. Our results showed that escape populations were comprised of cells that have somehow avoided mutation to the targeted essential genes as well as any mutations on the plasmid sequence. A deep sequencing (average sequencing depth of 20,000x) of 10 essential genes revealed mutation frequencies of less than 5% for essential genes and 8%, 6%, and 7% for RAD51, RAD52, and LIF1, respectively (Figure 23). These mutations were dominated by large deletions (>20bp) resulting in the destruction of the target site but also resulting in large deletions to the coding sequence of the target gene, implying gene knockout. However, as these mutated sequences represent a small minority of the total population, it suggests that the homologous recombination (HR) dominance over non-homologous end-joining (NHEJ) prevents significant edits even at nonessential gene loci. The rapid division of cells provides a significant amount of template for HR machinery, which could correspond with the higher inoculated cultures escaping more quickly: a shortened lag phase enabled high rates of division to overcome the CRISPR-Cas rate of double strand breaks.

In the co-targeted strains, targeted sequencing revealed that the defensive gene loci (LIF1/RAD51/RAD52) still experienced a small amount of mutation while mutations in essential genes were still undetectable, with the exception of a small deletion (2bp) in the ACC2 gene in the Vip48 strain. The results in the co-targeted strains are consistent with the single-targeted strains and implies that the surviving populations, while having an active CRISPR-Cas system, are those that are able to repair their breaks in essential genes with the accurate HR method. Additionally, because the defensive genes do not experience a full knockout, a lowered but sufficient amount of repair proteins are expressed for DNA damage repair of essential loci. This creates a more toxic effect, but only in select cases does this eliminate the eventual growth of the population. Taken all together, the competition among kinetics of growth, DNA repair, and CRISPR-Cas activity are likely the key factors to control antifungal effectiveness. Increasing the efficiency of knockout through protein/sgRNA selection in conjunction with more effective ways to reduce DNA repair pathway expression will be important steps

moving forward should CRISPR-Cas be deployed effectively as an antifungal therapy.

Library depletion assay for selecting CRISPR-Cas essential gene targets

On-target algorithms are notoriously inaccurate in predicting cleavage efficiency and only marginally better at identifying binding affinity with dCas9 assays[55]. While DNAse sensitivity assays can clue into certain regions that are more amenable to editing with decent probability[36], granularity within a gene or region of interest is still unattainable. Therefore, the tiling of the genes of interest is important to select highly active sgRNA sequences when identifying essential genes for use as CRISPR-Cas antifungal targets. To determine which sgRNAs targeting essential genes are the most effective inhibitors of growth, we performed a library screen with 185 sgRNAs targeting 37 essential genes (Figure 24) using our constructed single plasmid tetracycline inducible Cas9 system (see Methods). The construction and use of a single plasmid with a tetracycline induced Cas9 expression can deconvolute the growth lag from galactose metabolism from the CRISP-Cas toxicity. The use of a stronger TEF promoter controlled by the Tet operator (TetO) also increases Cas9 expression while offering tight control via the introduction of tetracycline. The newly designed system also provides a more streamlined cloning method, which was necessary for constructing a larger sgRNA library.

Briefly, log2 fold-changes were generated by comparing cultures with Cas9 induced by 5 μg/mL tetracycline against control samples (no tetracycline) at the same timepoint. The volcano plot (Figure 24A) highlights the 19 sgRNAs with a high confidence of depletion ($p < 0.05$). The diversity of gene loci in the subset of highly depleted guides underscores the importance of sgRNA target selection necessarily over the selection of essential locus. The result is consistent with previous studies that suggest the efficiency of the sgRNA is a stronger contributing factor to antimicrobial efficiency than the target gene itself[149]. One gene RPL37A (60S ribosomal subunit protein), however, had 3/5 guide sequences appear in the highly depleted cutoff region. This could be due either to the sensitivity of *S. cerevisiae* to perturbations to this locus or be the result of high accessibility of Cas9 complexes to that locus. Understanding such underlying mechanisms of efficiency will be important in future studies of promising CRISPR-Cas antifungal targets.

Characterization of promising targets with tetracycline inducible construct

In order to assess both the efficacy of the newly designed system as well as the efficiency of our library-identified targets, we transformed these constructs individually into our BY4741 parent strain for further investigation. Initially, we characterized our single-targeting strains in varying concentrations of tetracycline while also varying the temperature to identify the optimal conditions for induction and CRISPR-Cas induced growth suppression. We identified that consistent with previous studies, tetracycline concentrations above 3 μg/mL are sufficient for strong induction. In addition, growth at 37^{o}C (as opposed to 30^{o}C) only makes a

marginal increase in the effectiveness of the system, suggesting that the improvement in Cas9 efficiency at its ideal operation temperature does not result in significant change to growth profile. Additionally, the results of the library assay were broadly consistent with the individual strain characterization. For example, ORC2-3 exhibited the highest log2 fold depletion in the assay and consistently showed the highest level of growth suppression across all conditions. Three of the most promising targets: ORC2-3 (origin recognition complex subunit 2), FUN12-3 (translation initiation factor eIF5B), and RPL37A-4 (ribosomal 60S subunit protein), had co-targeting constructs with RAD51 also build and assessed with high-resolution liquid culture measurements using 5 μg/ml tetracycline concentration at 30ºC (Figure 25C). Indeed, the co-targeting strains exhibited a substantial decrease in growth compared to the single-target essential strains over the course of 60 hours. This result emphasizes the strength of the library and co-targeting approach for designing the most effective CRISPR-Cas antifungals.

Conclusions

The development of novel antifungal therapies is an area of concern for public health. This study has highlighted significant design principles and identified important factors and areas of improvement for harnessing CRISPR-Cas as a potent and effective antifungal therapy. Importantly, the robust DNA repair mechanisms encountered in eukaryotic pathogens must be addressed by means of direct targeting or utilization as a “Trojan Horse” to counteract their role in directing cells to a persistor phenotype. This study underscores the importance of such a co-targeting approach to increase CRISPR-Cas antifungal activity. In addition, informed selection of these defensive genes is needed based on the pathogen in question, as certain pathways are favored in different species. Finally, selection of a highly active essential gene-targeting sgRNA plays a significant role in efficacy, as the kinetic interplay between DNA cleavage and repair governs the fate of the population. Future work should focus on evaluating the temporal dimension to determine means of preventing escaped persistor cells from re-colonizing a host.

CHAPTER SIX
IDENTIFICATION OF TARGET AND CAS NUCLEASE COMBINATIONS FOR ANTIMICROBIALS

Summary: Diversity in prokaryotic DNA repair pathways and the rapid rate of division generates new problems and opportunities for the design of CRISPR-Cas antimicrobials. This chapter presents research into combinations of different techniques within different organisms (*Bacillus subtilis, Escherichia coli, Staphylococcus aureus)* and develops certain design rules and delivery strategies for employing CRISPR-Cas as an antimicrobial. Two key insights are highlighted: first, the SOS response is a significant barrier to CRISPR-Cas antibacterial efficacy, but co-targeting relevant regulators is an effective strategy for increasing potency. Second, the kinetics of rapid cellular division, DNA replication and DNA repair rates in exponentially dividing bacterial populations outcompete CRISPR-Cas cleavage kinetics. Using multiple turnover enzymes, preferably with trans nuclease activity, can help to counter this kinetic imbalance in favor of growth suppression and cell death.

Introduction

The rise of antimicrobial resistance coupled with the slowing rate of antibiotic development is a severe concern for public health[62,75,76]. This potential crisis has been partially brought about by the overuse of currently available antibiotics in both human healthcare and agriculture/animal husbandry. With diminishing financial incentives to probe antibiotic design space, the problem becomes even worse. The prokaryotic adaptive immune CRISPR-Cas systems are near-ubiquitous across the archaea and bacterial kingdoms[17]. Their inherent selectivity and adaptability based on the designed guide-RNA sequence can be harnessed for a new generation of highly selective antimicrobials. The differentiation of CRISPR-based antimicrobials from the traditional chemicals opens up a new design space while complementing traditional treatments for effective cocktail therapies.

Previous work into the capabilities of the CRISPR-based antimicrobial system range across a variety of organisms including clinically relevant pathogens, suggesting this approach is broadly applicable. In general, there exist two strategies, including CRISPR-Cas induced cell death and the re-sensitization of the targeted microbe to antibiotics[62,63]. The first approach (and the focus of this chapter) requires the CRISPR-Cas machinery to be delivered efficiently and highly active to prevent the proliferation of untargeted cells. Early studies focused on the transformation efficiency of CRISPR-Cas systems designed with guide-RNAs targeting essential elements of the genome or antibiotic resistance markers. While these studies established the proof-of-concept, the ability to understand escape mechanisms was limited due to the high level of selective pressure in the experimental design[13]. Further studies explored more free systems, and found that the SOS response in *E. coli* inhibits the lethality of CRISPR-Cas systems, which can subsequently be reactivated by using recA knockout mutants[73]. Finally,

the delivery of these systems in a clinical setting needed to be explored, with both conjugative and phage delivery techniques showing promise. In particular, Marraffini and colleagues demonstrated the ability of a phage-adapted *S. aureus* strain to be killed by introduction of a self-targeting CRISPR-Cas machinery via phagemid[75,150].

The specificity of CRISPR-Cas antimicrobials enables the control of a species within a microbiome. Tailoring CRISPR-Cas delivery systems also imparts an additional level of control (e.g., using a species-specific phage). As microbiomes gain more prominence as crucial ecological cornerstones, the ability to modulate and selectively remove species will be an important technique that is impossible with the current suite of antibiotics[63]. A classic example is the treatment of *Clostridium difficile* infections in the human gastrointestinal system[151]. Such infections require a high dose of a cocktail of antibiotics that leave the host microbiome weakened. *C. difficile* persistor cells often remain, leaving the host microbiome susceptible to the slightest perturbations that will then prompt reinfection, resulting in *C. difficile* often being classified as a chronic infection. Treatment that can selectively target *C. difficile* while leaving the host microbiome undisturbed and therefore able to contribute to suppressing *C. difficile* growth. This effect of chronic infection resulting from a weakened microbiome is also found in other well-known pathogens including certain strains of *E. coli* and *S. aureus*[13,67,75,143,151,152]. In order to address the key challenges towards deployment of CRISPR-Cas antimicrobials, this chapter goes through a series of experiments to investigate promising design strategies for maximizing CRISPR-Cas antimicrobial potential.

Materials and Methods

Strains and Constructs

BPK2101 from the Joung's Lab at Harvard University was used as the base vector for all *E. coli* experiments[97]. To create the vectors with other Cas proteins, the vector was amplified with generic primers. The swap primers for each protein was used to amplify the sequence found in their respective vectors. The backbone fragment was then combined with each of the Cas protein PCR fragments via Gibson Assembly. Each of these products were then amplified with BM_198.f/r and the respective RNA scaffold fragments were combined with the appropriate PCR product to give the final base plasmids. In the case of the native ecoCASCADE, the Cas3 sequence replaced the saCas9 sequence to increase expression of the exonuclease. Cloning procedure for inserting the guide sequences followed the protocol described in Joung *et al*.[97] To introduce a second guide sequence, either a second operon was cloned in using Gibson assembly in the case of those using sgRNA, or the respective guide combination was ordered as a spacer-repeat-spacer oligo with the appropriate BsaI overhangs and cloned in via ligation in the same manner as the single guides. *E. coli* K-12 (CGSC# 4401)

was used for all characterization, owing to its F' genotype for use in M13 phagemid work.

All work for *B. subtilis* was performed in the strain NRRL-765. The dual plasmid system from the Gross Lab was used to integrate CRISPR-Cas machinery at the *lacA* (pJMP1, dspCas9 containing) and *amyE* (pJMP2, sgRNA containing) loci, respectively[74]. The dspCas9 sequence in pJMP1 plasmid was replaced with a catalytically active spCas9 sequence. Natural competence transformation was used to deliver the vectors in a serialized method, first creating the BMbs001 strain with the dCas9 operon genome integrated followed by creation of the remaining strains with natural competence transformation of pJMP2-based plasmids.

To build a library of sgRNAs targeting essential genes for *S. aureus*, the backbone of the pCasSA plasmid developed by the Ji Lab at ShanghaiTech[153] was used. This plasmid has the benefit of replicating and expressing in *S. aureus* strains episomally. It was then modified by inserting a 2.3 kb section of the phiNM1 genome PCR'd from the Newman strain obtained from ATCC (#25904). We inserted this section onto pCasSA in between the CAT (chloramphenicol resistance) and KanR sequences, resulting in the pCasSAvip plasmid. All cloning procedures for initial construction of the pCasSAvip plasmid have taken place in *E. coli* 10-beta (NEB) cells to allow for the successful molecular cloning of the larger constructs. Transformation of single gRNAs into pCasSA followed the protocol described in Chen *et al.*[153]

Media and culture conditions

All experiments in *B. subtilis* were performed with Luria Bertani (LB) media. 1% xylose was used for the induction of the Cas9 protein from its site of genome integration. Initial screen of *E. coli* native sgRNAs was performed in LB broth supplemented with 0.2% arabinose for induction of the crRNA array. All growth of *S. aureus* Sa113 was done in tryptic soy broth (TSB) with 5 μg/mL chloramphenicol for selection after transformation of the pCasSA plasmid (and derivatives).

Agar plate screening assay with M13 phage

To obtain CFUs from treatment with Cas-carrying M13 phages, *E. coli* K-12 strain was mixed with a MOI (multiplicity of infection) of 1000 M13 phages, translating to an effective TU (transforming units) ratio of approximately 10:1. Mixtures were inoculated in media with and without antibiotic (100 μg/mL ampicillin or 30 μg/ml chloramphenicol) and with and without 1 μg/mL IPTG (for inducing native ecoCASCADE crRNA array or Cas9) and kept at 37ºC with shaking at 250rpm. Cultures were then diluted 10x and 50 μL was plated onto an agar plate. Colony counts were taken 16 hours post-spreading.

Flow cytometry measurements

Samples of *E. coli* were taken immediately following the 12-hour outgrowth in liquid media (LB +/- 100 μg/mL ampicillin, +/- 1 μg/ml IPTG). Flow cytometry was performed on a Guava EasyCyte HT-6 using the InCyte software for analysis.

Invitrogen LIVE/DEAD BacLight staining was used according to manufacturer's protocol. Cutoffs were calibrated based on a sample of exponentially grown (OD ~ 0.5) *E. coli* K-12 and a sample placed in 200 μg/ml ampicillin for 3 hours.

Library screens in S. aureus

To perform the library screen, the initial plasmid library was sequenced by singly-digesting the plasmid to linearize, and then using the ligation sequencing kit (LSK109) from Oxford Nanopore Technologies and a MinION flow cell to perform sequencing. Downstream analysis used guppy for base-calling and a custom alignment search algorithm to identify sequence hits. The method for the library depletion study is as follows: the plasmid library was transformed via electroporation into Sa113 cells. Following electroporation, the cells were added to SOB recovery media for 1.5 hours and then placed in 25 ml TSB media with 5 ng/ml chloramphenicol in triplicate as recommended by Chen *et al.*[153] Cells were grown at 37$_0$C and 5 ml samples were extracted at 24 hours for plasmid purification. Isolated and purified plasmids were then taken through the same sequencing preparation procedure as the initial library.

Results and Discussion

Bacillus subtilis is prone to escape when targeted with spCas9

As *B. subtilis* does not carry a native CRISPR system, it serves as an ideal platform to investigate in isolation the toxic effect of CRISPR-Cas induced double strand breaks. *B. subtilis* also has a well-studied set of essential genes, which provide a list of potential target sites. Previous studies have investigated the effect of targeting essential genes with catalytically inactive (dCas9)[89]. Such targeting resulted in an increase in the lag phase to various degrees depending on the target. To determine whether introduction of a catalytically active Cas9 changed the growth response, a similar growth study was performed here with a subset of 31 essential gene targets. The result showed that culture growth behaved similarly to that of dCas9, exhibiting increases in lag time in strains with sgRNAs targeting essential genes (Figure 26A). This behavior was also observed to a much lesser degree on agar plates (Figure 26B). By the 36-hour mark, there was a small but noticeable change in the colony fullness on 1% xylose (induced) plates, suggesting a small change due to Cas9 expression. However, this is likely due to the overall expression of Cas9, which even in the absence of a targeting sgRNA, can still complex with the empty scaffold RNA and bind randomly to PAM sites throughout the genome, inducing mild toxicity. This hypothesis is corroborated by the control colony, which exhibits similar colony size suppression on the 1% xylose plate.

Use of spCas9 in *B. subtilis* is a helpful exercise to identify areas of improvements for effective design of CRISPR-Cas antimicrobials. First, the use of a genome integrated Cas9 locus is likely to express a lesser amount of protein than from a high-copy plasmid, reducing the overall Cas9 concentration, thereby reducing activity. Second, this exercise targeted essential genes without

addressing the SOS response or other DNA repair machinery. Finally, previous studies and investigation from this work in using CRISPR-Cas as an antifungal highlight the importance of balancing the kinetic factors in favor of CRISPR-Cas. Use of the single-turnover spCas9 enzyme is — from a kinetic perspective — the poorest choice of enzyme for antimicrobial design. Investigation of these three areas for potential improvement are the subjects of the following sections.

Screening E. coli targets with endogenous CRISPR

The limitation of spCas9 as a single turnover enzyme makes it a poor choice for CRISPR-Cas antimicrobials. Once the DNA target is cleaved, spCas9 remains in its cleaved conformation, bound to the remnants of the DNA substrate and unable to bind to and cleave further targets. In contrast, other Cas types (and indeed other Cas9 orthologs) will bind and unbind to targets acting as multiple turnover enzymes. In addition, the Cas3 endonucleases of Type I systems also exhibit exonuclease activity, first cutting the DNA and then further degrading the DNA strand at the target site. The principal limiting factor to using Type I (and Type III) systems is their multi-effector nature that makes them difficult to express heterologously.

E. coli (K-12 strain) is a suitable organism to investigate Type I systems as antimicrobials, as it possesses a characterized native Type I CRISPR-Cas system, termed ecoCASCADE. By expressing a heterologous synthetic CRISPR array with crRNA sequences targeting GFP gene sequence in a Cas3 knockout strain, Luo *et al.* showed substantial suppression of GFP fluorescence[13]. To study the effect of essential gene targeting crRNAs, we co-opted the expression system on the developed pcrRNA.ind vector (Figure 27A), where the heterologous array was under the inducible arabinose (AraBAD) promoter. To compare the results of the native system with spCas9, the plasmid backbone of BPK2101 was used to express spCas9 under a T7 promoter controlled by the *lac* operon, with sgRNAs expressed under another T7 promoter (Figure 27A). Growth in selective LB broth (+100 μg/mL ampicillin, +30 μg/mL chloramphenicol, respectively) with induction molecules (0.2% arabinose, 0.1mM IPTG) showed the ecoCASCADE increasing growth suppression relative to the Cas9 control subset (Figure 27B). Interestingly, ecoCASCADE not only increased lag time of the strains but also reduced the growth rate of the strains in many cases. Because the *lac* operon is notoriously leaky, the Cas9 expression may have caused bias due to the cloned strains only containing mutants that were primed for survival in the presence of CRISPR-Cas. To reduce this bias and to provide a more clinically relevant mechanism of delivery, we investigated the ability of both systems to be packaged in an M13 lysogenic phage system.

Phagemid delivery and co-targeting SOS response with CRISPR-Cas

In complex environments such as that of the site of infection in a host or broader ecosystem, delivery of CRISPR-Cas antimicrobials cannot be achieved through basic laboratory transformation techniques. The use of the M13 phage for DNA

delivery into *E. coli* with an F genotype is a common laboratory technique that can also be harnessed for packaging and delivering CRISPR-Cas machinery. An added benefit of using the M13 is that it is lysogenic in standard laboratory conditions. Results from the assays of phage induction show the ability of co-targeting to increase antibacterial activity but underscore the substantial variability between target sequences (Figure 27C). Further, keeping selection pressure for the plasmid containing the CRISPR-Cas machinery leads to a lower colony count. This result is expected because the longer the CRISPR-Cas machinery persists, the more likely the SOS system becomes overwhelmed with DNA damage and the cell resorts to either dormancy or apoptosis.

The M13 phage delivery method was also used when comparing a Cas9 ortholog — saCas9 (from *S. aureus)* — to the native ecoCASCADE. In this case, the BPK2101 backbone was used and therefore the selective antibiotic was switched to chloramphenicol (Figure 27D). Broadly, saCas9 showed a similar pattern of depletion to the ecoCASCADE, with similar levels of reduction in colony count. The principal difference between saCas9 and spCas9 is the known multiple turnover nature of saCas9. Following the kinetic theory of CRISPR-Cas antimicrobial efficacy, the saCas9 performs as expected, exhibiting similar levels of targeting to ecoCASCADE, which in the previous growth assay was shown to outperform spCas9 as an antimicrobial. These results highlight the importance of using a multiple turnover Cas system for increasing the antimicrobial capability of CRISPR-Cas.

To identify both loci that are amenable to targeting and the most active target sequence for sgRNA design requires investigating a substantially larger combinatorial space than can be reasonably performed with the techniques used in the previous growth assays. However, selection based on growth (or lack thereof) presents the opportunity for using a library selection method to identify promising sgRNA targets.

Library screen to identify common targets and promising alternative endonucleases in the pathogen Staphylococcus aureus

To investigate a library of 125 guides in the pathogen *S. aureus,* we followed the workflow described in Figure 28A. Briefly, oligos containing the desired target sequence with golden gate overlaps were pooled and cloned into the pCasSA plasmid. This plasmid library was then transformed into *S. aureus* Sa113 strain. Log2 fold changes were calculated by taking the average read count of the three samples for each sequence and comparing them to the initial pooled sample. The library screen identified key sequences with significant levels of depletion (Figure 28B). The distribution of sequences with log2 fold depletions > 2 shows that some genomic loci are more prone to targeting than others (Table 3). For example, dnaJ (3x), teichB (2x), and gyrB (2x) have multiple sgRNAs that are significantly depleted in the top 13 sequences shown in Table 3. The remaining 6 sgRNA sequences in Table 3 appear on separate loci. So while the repeatedly targeted loci may be more amenable to targeting, the activity of a given sgRNA sequence

,irrespective of targeted gene, is an equally important metric. Such a diverse distribution of sgRNAs validates the use of library depletion assays to efficiently tile sequences across multiple genomic sites.

Characterization of both multiple turnover kinetics and trans nuclease activity in the Type V family of Cas enzymes is promising for their deployment of antimicrobials. In particular the asCas12 enzyme has shown a significant amount of trans activity *in vitro*[53]. To investigate its performance as a potential antimicrobial relative to spCas9, we built a library of 68 asCas12 crRNAs and compared it to a high performing subset of spCas9 sequences (log2 fold depletion > 0.5). As highlighted by Figure 28C, the asCas12 library (median sgRNA rank = 53) outperformed the high performing spCas9 library (median sgRNA rank = 70) with 12 of the 15 most significantly depleted sequences coming from the asCas12 library. This result supports the use of Cas12 family of nucleases for antimicrobial design. Their multiple turnover nature coupled with trans ssDNA nuclease boost the rate of DNA damage, thereby rebalancing the kinetic equilibrium in the favor of the CRISPR-Cas system and cell death.

Conclusions

The design of CRISPR-Cas for antimicrobial applications requires appropriate selection of a highly active target sequence on an available genome sequence with a multiple turnover enzyme preferably with trans nuclease activity. In addition, the co-targeting of defense systems, namely the SOS response, is important for increasing potency to relevant levels. Finally, the delivery via phage represents a unique ability to multiplex the lytic properties of some phages with the growth arrest and/or apoptosis caused by self-targeting CRISPR-Cas. This study lays out a systematic workflow for screening and characterizing antimicrobial targets. Using a library screen of tiled essential genes can narrow down the selection of guide-RNA sequences to a select number that can then be characterized with combinations of pre-determined highly active defensive (SOS) targeting guides. These combinations can be then packaged into a phage and assayed for their killing efficiency in various environments, particularly non-selective conditions that are more relevant to clinical application.

Ultimately, the ability of bacteria to inevitably evolve away via mutations to target sites will require multiple guide-RNA designs to be developed as a sort of cocktail therapy to prevent the formation of persistent mutant genotypes. This is an area of particular strength for CRISPR-Cas antimicrobial design and implementation, as the synthesis of a variety of guides is both rapid and cheap, making them ideal for combating time-sensitive pathogen outbreaks.

CHAPTER SEVEN CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

In this dissertation a pipeline was built for rapid identification of putative target sites for manipulation by CRISPR-Cas enzymes in varied types of genomic data, from population-level metagenomes to well-annotated model organisms. Further, the capability of protein modeling to enhance off-target activity prediction was investigated. Mining of the entirety of the available prokaryotic genomic data in NCBI's database to identify features of novel CRISPR-Cas systems in unstudied organisms completes the computational portion of the work. These exercises laid the groundwork to apply CRISPR-Cas as an antimicrobial. When designed properly, with careful attention to the target organism, guide-RNA(s) selection, Cas protein, and delivery mechanism, CRISPR-Cas is a highly capable system with inherent specificity. The importance of these features was further underscored in various experimental conditions and assay types that identified key mechanisms of resistance and persistence.

This work opens the combinatorial space of CRISPR-Cas enzyme and target design to increase antimicrobial potency of these systems, particularly as it applies to modulation of diverse microbiomes. In order to enhance the toolkit, key improvements at each step of the design process are suggested in the sections that follow; these will build upon a strong foundation of work set about by this dissertation.

Expanding population engineering computational workflow

The ability to design CRISPR-Cas tools for engineering microbiomes and other mixed populations is a powerful feature of the CASPER platform developed in this work. As understanding and capabilities to sequence and annotate metagenomes more precisely become widespread, the need for a more robust and specialized set of algorithms for dealing with these datasets will be necessary. The integration of CASPER into a user-friendly software platform for a broad audience is a critical development. The wide-ranging applications of CRISPR-Cas also demand consistent feedback from the research community to keep the algorithms relevant to experimentalists. Of note, the use of base-editors for modifying SNPs and the use of Cas12/13 nucleases for sequence detection in unknown samples are two applications uniquely suited to the bioinformatic capabilities of CASPER[60,61]. In addition, the foundations of CASPER's target identification and on/off-target activity algorithms can be combined with techniques in phylogenetic analysis to more precisely predict potential off-target activity within a microbiome. This will facilitate the design process for microbiome modulation, a promising new area of research that is still in its infancy.

Increased modeling accuracy with Rosetta parameterization

Using a Rosetta framework to predict R-loop energetics without the need for computationally costly molecular dynamics (MD) simulations allows such ternary

complex simulations to be used for approximating off-target activity[46]. The work presented in this dissertation highlights the capabilities of such a method, demonstrating the ability to rely on biophysical principles rather than opaque machine learning rules to evaluate activities of guide-RNAs. This enables the methodology to be expanded to other single-effector Cas proteins that may require re-training in the machine learning paradigm. However, the work here can be greatly expanded upon in three key directions: 1) increasing the number of Cas nucleases and accompanying datasets, 2) re-parameterization of Rosetta algorithms to streamline computation and more accurately assess R-loop energetics, and 3) establish an accessible pipeline for accessing Rosetta-enabled predictions.

First, the current study relied on two seminal datasets using three Cas nucleases. There currently exists a substantial number of off-target assays that can be examined using this framework but requires significantly more computational time to perform such analysis. By investigating other experiments, further insight can be obtained on both the accuracy of the model and the role of certain factors in contributing to absolute off-target activity. Future experimental assays can be designed with these parameters in mind to design more precise and accurate off-target assays that can be utilized for clinical applications.

Second, the use of the default settings in Rosetta were chosen to limit the vast combinatorial space of parameter decisions. Recently, updates to the Rosetta suite—including an upgraded RNA structure mutation algorithm—open the possibility for this space to be systematically explored[154]. Optimistically, these improved algorithms may further reduce the need for structure averages to be used to minimize the degree of local energy well fluctuations.

Finally, this prediction methodology is most useful when combined with an easily accessible portal to investigate potential off-targets of a guide-RNA sequence. The CASPER software developed as part of this dissertation provides an excellent platform to integrate such a feature, but due to the limited number of sequences currently analyzed by Rosetta, is not useful for most applications. The combination of increasing the number of guides analyzed with future access to a customized Rosetta/CRISPR simulation server will ultimately provide a valuable research to the wider research community.

Expanding principles to pathogenic fungi and validating mechanisms of delivery

The model yeast *Saccharomyces cerevisiae* used in this work is the ideal organism for investigating principles of design by providing a well-curated genome for investigation. This facilitated the identification of key features of CRISPR-Cas antifungal design, principally the importance of kinetic factors when such systems are inducible or transiently expressed, as would be relevant for deployment in clinical applications. Counteracting the DNA repair machinery is of importance to increase antifungal potency, particularly with homologous recombination machinery in *S. cerevisiae.* Other pathogenic fungi present their own unique

challenges with respect to their DNA repair mechanisms. Large ploidy numbers and multi-nucleated cells are of concern, presenting a greater challenge for CRISPR-Cas induced gene disruption[147,155,156]. Therefore, in these new pathogens, targeting replication/reproduction/and growth phenotype machinery may be equally important towards ensuring the maximum antifungal potential. Library screens such as the ones utilized in this work will play a critical role in these investigations, as they are a rapid and cheap way to reduce the design space for potential target combinations.

Controlling the microbiome through selective species

The design rules developed in this work for increasing the efficacy of CRISPR-Cas antimicrobials demonstrates key areas for systematic improvement, focusing on further tilting the kinetic equilibrium to favor the induction of cell death. A key variable that must be explored is the interplay of such perturbation within the scope of the microbial community. In particular, the highly selective removal of a certain species has been confined to synthetically reconstructed communities but is now possible in wild communities[157]. While pathogen/invasive species removal is a clear need, the more nuanced modulation of relative species abundance is a new research direction enabled by the CRISPR-Cas technology. Developing a greater understanding of the full spectrum of CRISPR-Cas mediated growth phenotypes, particularly in the context of a microbiome will need to be further explored to actualize microbial community control. Ultimately, the combination of the versatile CRISPR-Cas system with the increased understanding of the ability of microbial communities to transform ecosystems has the potential to be a transformative pairing.

REFERENCES

1 Jackson, S. A. *et al.* CRISPR-Cas: Adapting to change. *Science* **356**, eaal5056, doi:10.1126/science.aal5056 (2017).
2 Mojica, F. J. M., Díez-Villaseñor, C. s., García-Martínez, J. & Soria, E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution* **60**, 174-182, doi:10.1007/s00239-004-0046-3 (2005).
3 Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**, 733-740, doi:doi:10.1099/mic.0.023960-0 (2009).
4 Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709, doi:10.1126/science.1138140 (2007).
5 Horvath, P. *et al.* Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J Bacteriol* **190**, 1401-1412, doi:10.1128/JB.01415-07 (2008).
6 Garneau, J. E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71, doi:10.1038/nature09523 (2010).
7 Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences* **109**, E2579, doi:10.1073/pnas.1208507109 (2012).
8 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).
9 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823 (2013).
10 DiCarlo, J. E. *et al.* Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic acids research* **41**, 4336-4343 (2013).
11 Semenova, E. & Severinov, K. 51-62 (2017).
12 van der Oost, J., Westra, E. R., Jackson, R. N. & Wiedenheft, B. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. *Nature Reviews Microbiology* **12**, 479, doi:10.1038/nrmicro3279 https://www.nature.com/articles/nrmicro3279#supplementary-information (2014).
13 Gomaa, A. A. *et al.* Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *MBio* **5**, e00928-00913 (2014).
14 Chen, J. S. & Doudna, J. A. The chemistry of Cas9 and its CRISPR colleagues. *Nature Reviews Chemistry* **1**, doi:10.1038/s41570-017-0078 (2017).
15 Makarova, Kira S., Wolf, Yuri I. & Koonin, Eugene V. The basic building blocks and evolution of CRISPR–Cas systems. *Biochemical Society Transactions* **41**, 1392 (2013).

16 Makarova, K. S. & Koonin, E. V. in *CRISPR: Methods and Protocols* (eds Magnus Lundgren, Emmanuelle Charpentier, & Peter C. Fineran) 47-75 (Springer New York, 2015).
17 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Micro* **13**, 722-736, doi:10.1038/nrmicro3569 http://www.nature.com/nrmicro/journal/v13/n11/abs/nrmicro3569.html#supplementary-information (2015).
18 Barrangou, R. Diversity of CRISPR-Cas immune systems and molecular machines. *Genome Biol* **16**, 247, doi:10.1186/s13059-015-0816-9 (2015).
19 Leenay, Ryan T. *et al.* Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell* **62**, 137-147, doi:10.1016/j.molcel.2016.02.031 (2016).
20 Sapranauskas, R. *et al.* The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res* **39**, 9275-9282, doi:10.1093/nar/gkr606 (2011).
21 Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature methods* **10**, 1116-1121, doi:10.1038/nmeth.2681 (2013).
22 Wright, A. V. *et al.* Structures of the CRISPR genome integration complex. *Science* **357**, 1113-1118, doi:10.1126/science.aao0679 (2017).
23 Nishimasu, H. *et al.* Crystal Structure of Staphylococcus aureus Cas9. **162**, 1113-1126, doi:10.1016/j.cell.2015.08.007 (2015).
24 Yamano, T. *et al.* Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **165**, 949-962, doi:http://dx.doi.org/10.1016/j.cell.2016.04.003 (2016).
25 Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-771, doi:10.1016/j.cell.2015.09.038 (2015).
26 East-Seletsky, A., O'Connell, M. R., Burstein, D., Knott, G. J. & Doudna, J. A. RNA Targeting by Functionally Orthogonal Type VI-A CRISPR-Cas Enzymes. *Mol Cell* **66**, 373-383 e373, doi:10.1016/j.molcel.2017.04.008 (2017).
27 Liu, L. *et al.* The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. *Cell* **170**, 714-726 e710, doi:10.1016/j.cell.2017.06.050 (2017).
28 Carroll, D. Genome Engineering with Targetable Nucleases. *Annual Review of Biochemistry* **83**, 409-439, doi:10.1146/annurev-biochem-060713-035418 (2014).
29 Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* **32**, 1-13, doi:10.1038/nbt.3026 (2014).
30 Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotech* **34**, 184-191, doi:10.1038/nbt.3437

http://www.nature.com/nbt/journal/v34/n2/abs/nbt.3437.html#supplementary-information (2016).
31 Moreno-Mateos, M. A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods* **12**, 982-988, doi:10.1038/nmeth.3543 (2015).
32 Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186-191, doi:10.1038/nature14299 (2015).
33 Jeon, Y. *et al.* Direct observation of DNA target searching and cleavage by CRISPR-Cas12a. *Nat Commun* **9**, 2777, doi:10.1038/s41467-018-05245-x (2018).
34 Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490-495, doi:10.1038/nature16526
http://www.nature.com/nature/journal/v529/n7587/abs/nature16526.html#supplementary-information (2016).
35 Kleinstiver, B. P. *et al.* Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat Biotechnol* **34**, 869-874, doi:10.1038/nbt.3620 (2016).
36 Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827-832, doi:10.1038/nbt.2647 (2013).
37 Raper, A. T., Stephenson, A. A. & Suo, Z. Functional Insights Revealed by the Kinetic Mechanism of CRISPR/Cas9. *J Am Chem Soc* **140**, 2971-2984, doi:10.1021/jacs.7b13047 (2018).
38 Schmidt, S. T., Yu, F. B., Blainey, P. C., May, A. P. & Quake, S. R. Nucleic acid cleavage with a hyperthermophilic Cas9 from an uncultured Ignavibacterium. *Proceedings of the National Academy of Sciences* **116**, 23100, doi:10.1073/pnas.1904273116 (2019).
39 Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat Commun* **7**, 12778, doi:10.1038/ncomms12778 (2016).
40 Tsui, T. K. M., Hand, T. H., Duboy, E. C. & Li, H. The Impact of DNA Topology and Guide Length on Target Selection by a Cytosine-Specific Cas9. *ACS Synth Biol* **6**, 1103-1113, doi:10.1021/acssynbio.7b00050 (2017).
41 Yarrington, R. M., Verma, S., Schwartz, S., Trautman, J. K. & Carroll, D. Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proceedings of the National Academy of Sciences* **115**, 9351, doi:10.1073/pnas.1810062115 (2018).
42 Kallimasioti-Pazi, E. M. *et al.* Heterochromatin delays CRISPR-Cas9 mutagenesis but does not influence the outcome of mutagenic DNA repair. *PLOS Biology* **16**, e2005595, doi:10.1371/journal.pbio.2005595 (2018).

43 Mendoza, B. J. & Trinh, C. T. Enhanced guide-RNA design and targeting analysis for precise CRISPR genome editing of single and consortia of industrially relevant and non-model organisms. *Bioinformatics* **34**, 16-23, doi:10.1093/bioinformatics/btx564 (2018).
44 Chuai, G. *et al.* DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol* **19**, 80, doi:10.1186/s13059-018-1459-4 (2018).
45 Lin, J. & Wong, K. C. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics* **34**, i656-i663, doi:10.1093/bioinformatics/bty554 (2018).
46 Ribeiro, J. M. L., Tsai, S. T., Pramanik, D., Wang, Y. & Tiwary, P. Kinetics of Ligand-Protein Dissociation from All-Atom Simulations: Are We There Yet? *Biochemistry* **58**, 156-165, doi:10.1021/acs.biochem.8b00977 (2019).
47 Klein, M., Eslami-Mossallam, B., Arroyo, D. G. & Depken, M. Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Rep* **22**, 1413-1423, doi:10.1016/j.celrep.2018.01.045 (2018).
48 Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262-1278, doi:10.1016/j.cell.2014.05.010 (2014).
49 Anderson, E. M. *et al.* Systematic analysis of CRISPR–Cas9 mismatch tolerance reveals low levels of of off-target activity. *Journal of Biotechnology* **211**, 56-65, doi:10.1016/j.jbiotec.2015.06.427 (2015).
50 Anderson, K. R. *et al.* CRISPR off-target analysis in genetically engineered rats and mice. *Nature Methods* **15**, 512-514, doi:10.1038/s41592-018-0011-5 (2018).
51 Cheng, Y. & Tsai, S. Q. Illuminating the genome-wide activity of genome editors for safe and effective therapeutics. *Genome Biol* **19**, 226, doi:10.1186/s13059-018-1610-2 (2018).
52 Slaymaker, I. M. *et al.* High-Resolution Structure of Cas13b and Biochemical Characterization of RNA Targeting and Cleavage. *Cell Rep* **26**, 3741-3751.e3745, doi:10.1016/j.celrep.2019.02.094 (2019).
53 Kellner, M. J., Koob, J. G., Gootenberg, J. S., Abudayyeh, O. O. & Zhang, F. SHERLOCK: nucleic acid detection with CRISPR nucleases. *Nat Protoc* **14**, 2986-3012, doi:10.1038/s41596-019-0210-2 (2019).
54 Gootenberg, J. S. *et al.* Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* **360**, 439, doi:10.1126/science.aaq0179 (2018).
55 Boyle, E. A. *et al.* High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc Natl Acad Sci U S A* **114**, 5461-5466, doi:10.1073/pnas.1700557114 (2017).
56 Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc* **8**, 2180-2196, doi:10.1038/nprot.2013.132 (2013).

57 Ryan, O. W. *et al.* Selection of chromosomal DNA libraries using a multiplex CRISPR system. *Elife* **3**, e03703 (2014).
58 Zalatan, J. G. *et al.* Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell* **160**, 339-350, doi:10.1016/j.cell.2014.11.052 (2015).
59 Marshall, R. *et al.* Rapid and Scalable Characterization of CRISPR Technologies Using an E. coli Cell-Free Transcription-Translation System. *Molecular Cell* **69**, 146-157.e143, doi:https://doi.org/10.1016/j.molcel.2017.12.007 (2018).
60 Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420-424, doi:10.1038/nature17946 (2016).
61 Eid, A., Alshareef, S. & Mahfouz, M. M. CRISPR base editors: genome editing without double-stranded breaks. *Biochem J* **475**, 1955-1964, doi:10.1042/BCJ20170793 (2018).
62 Fagen, J. R., Collias, D., Singh, A. K. & Beisel, C. L. Advancing the design and delivery of CRISPR antimicrobials. *Current Opinion in Biomedical Engineering* **4**, 57-64, doi:https://doi.org/10.1016/j.cobme.2017.10.001 (2017).
63 Pursey, E., Sünderhauf, D., Gaze, W. H., Westra, E. R. & van Houte, S. CRISPR-Cas antimicrobials: Challenges and future prospects. *PLoS Pathog* **14**, e1006990, doi:10.1371/journal.ppat.1006990 (2018).
64 Høyland-Kroghsbo, N. M. *et al.* Quorum sensing controls the Pseudomonas aeruginosa CRISPR-Cas adaptive immune system. *Proceedings of the National Academy of Sciences* **114**, 131, doi:10.1073/pnas.1617415113 (2017).
65 Luo, M. L., Mullis, A. S., Leenay, R. T. & Beisel, C. L. Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression. *Nucleic Acids Res* **43**, 674-681, doi:10.1093/nar/gku971 (2015).
66 Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O'Toole, G. A. The CRISPR/Cas adaptive immune system of Pseudomonas aeruginosa mediates resistance to naturally occurring and engineered phages. *J Bacteriol* **194**, 5728-5738, doi:10.1128/JB.01184-12 (2012).
67 Pyne, M. E., Bruder, M. R., Moo-Young, M., Chung, D. A. & Chou, C. P. Harnessing heterologous and endogenous CRISPR-Cas machineries for efficient markerless genome editing in Clostridium. *Sci Rep* **6**, 25666, doi:10.1038/srep25666 (2016).
68 Guan, J., Wang, W. & Sun, B. Chromosomal Targeting by the Type III-A CRISPR-Cas System Can Reshape Genomes in Staphylococcus aureus. *mSphere* **2**, e00403-00417, doi:10.1128/mSphere.00403-17 (2017).

69 Xu, Z. *et al.* Native CRISPR-Cas-Mediated Genome Editing Enables Dissecting and Sensitizing Clinical Multidrug-Resistant P. aeruginosa. *Cell Rep* **29**, 1707-1717.e1703, doi:10.1016/j.celrep.2019.10.006 (2019).
70 Leenay, R. T. & Beisel, C. L. Deciphering, Communicating, and Engineering the CRISPR PAM. *Journal of Molecular Biology* **429**, 177-191, doi:https://doi.org/10.1016/j.jmb.2016.11.024 (2017).
71 Knight, S. C. *et al.* Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350**, 823-826, doi:10.1126/science.aac6572 (2015).
72 Žgur-Bertok, D. DNA damage repair and bacterial pathogens. *PLoS Pathog* **9**, e1003711, doi:10.1371/journal.ppat.1003711 (2013).
73 Moreb, E. A. *et al.* Managing the SOS Response for Enhanced CRISPR-Cas-Based Recombineering in E. coli through Transient Inhibition of Host RecA Activity. *ACS Synth Biol* **6**, 2209-2218, doi:10.1021/acssynbio.7b00174 (2017).
74 Peters, J. M. *et al.* A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. *Cell* **165**, 1493-1506, doi:https://doi.org/10.1016/j.cell.2016.05.003 (2016).
75 Bikard, D. *et al.* Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. *Nat Biotechnol* **32**, 1146-1150, doi:10.1038/nbt.3043 (2014).
76 Citorik, R. J., Mimee, M. & Lu, T. K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nat Biotechnol* **32**, 1141-1145, doi:10.1038/nbt.3011 (2014).
77 Ciccia, A. & Elledge, S. J. The DNA damage response: making it safe to play with knives. *Mol Cell* **40**, 179-204, doi:10.1016/j.molcel.2010.09.019 (2010).
78 Connelly, T. *et al.* Industrialization of Biology: A Roadmap to Accelerate the Advanced Manufacturing of Chemicals. (2015).
79 Trinh, C. T. & Mendoza, B. Modular cell design for rapid, efficient strain engineering toward industrialization of biology. *Current Opinion in Chemical Engineering* **14**, 18-25, doi:http://dx.doi.org/10.1016/j.coche.2016.07.005 (2016).
80 Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164**, 1185-1197 (2016).
81 Jullesson, D., David, F., Pfleger, B. & Nielsen, J. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnology Advances* **33**, 1395-1402 (2015).
82 Lee, J. W. *et al.* Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nature chemical biology* **8**, 536-546 (2012).
83 Jakočiūnas, T. *et al.* Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. *Metabolic Engineering*, 1-10, doi:10.1016/j.ymben.2015.01.008 (2015).

84 Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* **32**, 347-355 (2014).
85 Li, J.-F. *et al.* Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nature Biotechnology* **31**, 688-691 (2013).
86 Platt, R. J. *et al.* CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* **159**, 440-455 (2014).
87 Farasat, I. & Salis, H. M. A Biophysical Model of CRISPR/Cas9 Activity for Rational Design of Genome Editing and Gene Regulation. *PLoS Comput Biol* **12**, e1004724, doi:10.1371/journal.pcbi.1004724 (2016).
88 Qi, L., Haurwitz, R. E., Shao, W., Doudna, J. A. & Arkin, A. P. RNA processing enables predictable programming of gene expression. *Nature Biotechnology* **30**, 1002-1006 (2012).
89 Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173-1183 (2013).
90 Zeitoun, R. I. *et al.* Multiplexed tracking of combinatorial genomic mutations in engineered cell populations. *Nat Biotech* **33**, 631-637, doi:10.1038/nbt.3177
http://www.nature.com/nbt/journal/v33/n6/abs/nbt.3177.html#supplementary-information (2015).
91 Garst, A. D. *et al.* Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat Biotech* **advance online publication**, doi:10.1038/nbt.3718
http://www.nature.com/nbt/journal/vaop/ncurrent/abs/nbt.3718.html#supplementary-information (2016).
92 Stovicek, V., Borodina, I. & Forster, J. CRISPR–Cas system enables fast and simple genome editing of industrial Saccharomyces cerevisiae strains. *Metabolic Engineering Communications* **2**, 13-22, doi:http://dx.doi.org/10.1016/j.meteno.2015.03.001 (2015).
93 Feng, X., Lian, J. & Zhao, H. Metabolic engineering of Saccharomyces cerevisiae to improve 1-hexadecanol production. *Metabolic Engineering* **27**, 10-19, doi:10.1016/j.ymben.2014.10.001 (2015).
94 Moreno-Mateos, M. A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Meth* **12**, 982-988, doi:10.1038/nmeth.3543
http://www.nature.com/nmeth/journal/v12/n10/abs/nmeth.3543.html#supplementary-information (2015).
95 Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature biotechnology* **31**, 827-832, doi:10.1038/nbt.2647 (2013).
96 Lin, Y. *et al.* CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Research*, doi:10.1093/nar/gku402 (2014).

97 Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481-485, doi:10.1038/nature14592 (2015).
98 Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Research* (2016).
99 Hendel, A. *et al.* Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nature Biotechnology* **33**, doi:10.1038/nbt.3290 (2015).
100 Malina, A. *et al.* PAM multiplicity marks genomic target sites as inhibitory to CRISPR-Cas9 editing. *Nature Communications* **6**, 10124, doi:10.1038/ncomms10124 (2015).
101 Prykhozhij, S. V., Rajan, V., Gaston, D. & Berman, J. N. CRISPR MultiTargeter: A Web Tool to Find Common and Unique CRISPR Single Guide RNA Targets in a Set of Similar Sequences. *PLoS ONE* **10**, e0119372, doi:10.1371/journal.pone.0119372 (2015).
102 Andrianantoandro, E., Basu, S., Karig, D. K. & Weiss, R. Synthetic biology: new engineering rules for an emerging discipline. *Molecular Systems Biology* **2** (2006).
103 Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* **343**, 80 (2014).
104 Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvák, Z. Molecular Reconstruction of Sleeping Beauty, a Tc1-like Transposon from Fish, and Its Transposition in Human Cells. *Cell* **91**, 501-510, doi:10.1016/S0092-8674(00)80436-5 (1997).
105 Geurts, A. M. *et al.* Gene Transfer into Genomes of Human Cells by the Sleeping Beauty Transposon System. *Mol Ther* **8**, 108-117 (2003).
106 Shibata, M. *et al.* Real-space and real-time dynamics of CRISPR-Cas9 visualized by high-speed atomic force microscopy. *Nat Commun* **8**, 1430, doi:10.1038/s41467-017-01466-8 (2017).
107 Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Sci Adv* **3**, eaao0027, doi:10.1126/sciadv.aao0027 (2017).
108 Jiang, F. & Doudna, J. A. CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys* **46**, 505-529, doi:10.1146/annurev-biophys-062215-010822 (2017).
109 Tycko, J. *et al.* Pairwise library screen systematically interrogates Staphylococcus aureus Cas9 specificity in human cells. *Nat Commun* **9**, 2962, doi:10.1038/s41467-018-05391-2 (2018).
110 Amee, S. i. Z., Mubarak, A. S., Süleyman, A. & Mehmet, O. in *10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions - ICSCCW-2019.* (eds Rafik A. Aliev *et al.*) 697-703 (Springer International Publishing).
111 Tsui, T. K. M., Hand, T. H., Duboy, E. C. & Li, H. The Impact of DNA Topology and Guide Length on Target Selection by a Cytosine-Specific

Cas9. *ACS synthetic biology* **6**, 1103-1113, doi:10.1021/acssynbio.7b00050 (2017).
112 van Aelst, K., Martínez-Santiago, C. J., Cross, S. J. & Szczelkun, M. D. The Effect of DNA Topology on Observed Rates of R-Loop Formation and DNA Strand Cleavage by CRISPR Cas12a. *Genes (Basel)* **10**, 169, doi:10.3390/genes10020169 (2019).
113 Brinkman, E. K. *et al.* Kinetics and Fidelity of the Repair of Cas9-Induced Double-Strand DNA Breaks. *Molecular cell* **70**, 801-813.e806, doi:10.1016/j.molcel.2018.04.016 (2018).
114 Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S. & Shaw, D. E. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J Am Chem Soc* **134**, 3787-3791, doi:10.1021/ja209931w (2012).
115 Robustelli, P., Trbovic, N., Friesner, R. A. & Palmer, A. G. Conformational Dynamics of the Partially Disordered Yeast Transcription Factor GCN4. *J Chem Theory Comput* **9**, doi:10.1021/ct400654r (2013).
116 Leaver-Fay, A. *et al.* ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* **487**, 545-574, doi:10.1016/B978-0-12-381270-4.00019-6 (2011).
117 Cheng, C. Y., Chou, F. C. & Das, R. Modeling complex RNA tertiary folds with Rosetta. *Methods Enzymol* **553**, 35-64, doi:10.1016/bs.mie.2014.10.051 (2015).
118 Barlow, K. A. *et al.* Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein–Protein Binding Affinity upon Mutation. *The Journal of Physical Chemistry B* **122**, 5389-5399, doi:10.1021/acs.jpcb.7b11367 (2018).
119 Bae, S., Park, J. & Kim, J. S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473-1475, doi:10.1093/bioinformatics/btu048 (2014).
120 Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLoS One* **10**, e0124633, doi:10.1371/journal.pone.0124633 (2015).
121 Mendoza, B. J. & Trinh, C. T. In Silico Processing of the Complete CRISPR-Cas Spacer Space for Identification of PAM Sequences. *Biotechnol J* **13**, e1700595, doi:10.1002/biot.201700595 (2018).
122 Shah, S. A., Erdmann, S., Mojica, F. J. M. & Garrett, R. A. Protospacer recognition motifs: Mixed identities and functional diversity. *RNA Biology* **10**, 891-899, doi:10.4161/rna.23764 (2013).
123 Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Current Opinion in Microbiology* **37**, 67-78 (2017).

124 Cebrian-Serrano, A. & Davies, B. CRISPR-Cas orthologues and variants: optimizing the repertoire, specificity and delivery of genome engineering tools. *Mamm Genome* **28**, 247-261, doi:10.1007/s00335-017-9697-4 (2017).
125 Elmore, J., Deighan, T., Westpheling, J., Terns, R. M. & Terns, M. P. DNA targeting by the type I-G and type I-A CRISPR–Cas systems of Pyrococcus furiosus. *Nucleic Acids Research* **43**, 10353-10363, doi:10.1093/nar/gkv1140 (2015).
126 Karvelis, T. *et al.* Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. *Genome Biology* **16**, 253, doi:10.1186/s13059-015-0818-7 (2015).
127 Biswas, A., Gagnon, J. N., Brouns, S. J. J., Fineran, P. C. & Brown, C. M. CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. *RNA Biology* **10**, 817-827, doi:10.4161/rna.24046 (2013).
128 Pyne, M. E., Bruder, M. R., Moo-Young, M., Chung, D. A. & Chou, C. P. Harnessing heterologous and endogenous CRISPR-Cas machineries for efficient markerless genome editing in Clostridium. **6**, 25666, doi:10.1038/srep25666
https://www.nature.com/articles/srep25666#supplementary-information (2016).
129 Shmakov, S. A. *et al.* The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio* **8**, e01397-01317, doi:10.1128/mBio.01397-17 (2017).
130 Xue, C. *et al.* CRISPR interference and priming varies with individual spacer sequences. *Nucleic Acids Research* **43**, 10831-10847, doi:10.1093/nar/gkv1259 (2015).
131 Westra, Edze R. *et al.* CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Molecular Cell* **46**, 595-605, doi:10.1016/j.molcel.2012.03.018.
132 Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645-3647, doi:10.1093/bioinformatics/btx469 (2017).
133 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature* **519**, 199 (2015).
134 Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. *Nature* **550**, 137 (2017).
135 Bondy-Denomy, J. *et al.* Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature* **526**, 136 (2015).
136 Rauch, B. J. *et al.* Inhibition of CRISPR-Cas9 with bacteriophage proteins. *Cell* **168**, 150-158. e110 (2017).
137 Fu, B. X., Wainberg, M., Kundaje, A. & Fire, A. Z. High-Throughput Characterization of Cascade type I-E CRISPR Guide Efficacy Reveals Unexpected PAM Diversity and Target Sequence Preferences. *Genetics* **206**, 1727-1738, doi:10.1534/genetics.117.202580 (2017).

138 Mohanraju, P. *et al.* Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* **353** (2016).
139 Scorzoni, L. *et al.* Antifungal Therapy: New Advances in the Understanding and Treatment of Mycosis. *Front Microbiol* **8**, 36, doi:10.3389/fmicb.2017.00036 (2017).
140 Vyas, V. K. *et al.* New CRISPR Mutagenesis Strategies Reveal Variation in Repair Mechanisms among Fungi. *mSphere* **3**, doi:10.1128/mSphere.00154-18 (2018).
141 Almeida, F., Rodrigues, M. L. & Coelho, C. The Still Underestimated Problem of Fungal Diseases Worldwide. *Frontiers in Microbiology* **10**, 214 (2019).
142 Taori, S. K. *et al.* Candida auris outbreak: Mortality, interventions and cost of sustaining control. *J Infect* **79**, 601-611, doi:10.1016/j.jinf.2019.09.007 (2019).
143 Schlecht, L. M. *et al.* Systemic Staphylococcus aureus infection mediated by Candida albicans hyphal invasion of mucosal tissue. *Microbiology* **161**, 168-181, doi:10.1099/mic.0.083485-0 (2015).
144 Lu, Y., Su, C. & Liu, H. Candida albicans hyphal initiation and elongation. *Trends Microbiol* **22**, 707-714, doi:10.1016/j.tim.2014.09.001 (2014).
145 Kornitzer, D. Regulation of. *J Fungi (Basel)* **5**, doi:10.3390/jof5010021 (2019).
146 Bijlani, S., Thevandavakkam, M. A., Tsai, H. J. & Berman, J. Autonomously Replicating Linear Plasmids That Facilitate the Analysis of Replication Origin Function in Candida albicans. *mSphere* **4**, doi:10.1128/mSphere.00103-19 (2019).
147 Motaung, T. E., Ells, R., Pohl, C. H., Albertyn, J. & Tsilo, T. J. Genome-wide functional analysis in Candida albicans. *Virulence* **8**, 1563-1579, doi:10.1080/21505594.2017.1292198 (2017).
148 Roemer, T. *et al.* Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. *Mol Microbiol* **50**, 167-181, doi:10.1046/j.1365-2958.2003.03697.x (2003).
149 Hamilton, T. A. *et al.* Efficient inter-species conjugative transfer of a CRISPR nuclease for targeted bacterial killing. *Nature Communications* **10**, 4544, doi:10.1038/s41467-019-12448-3 (2019).
150 Heler, R. *et al.* Cas9 specifies functional viral targets during CRISPR–Cas adaptation. *Nature* **519**, 199, doi:10.1038/nature14245 (2015).
151 Spigaglia, P., Mastrantonio, P. & Barbanti, F. Antibiotic Resistances of Clostridium difficile. *Adv Exp Med Biol* **1050**, 137-159, doi:10.1007/978-3-319-72799-8_9 (2018).
152 Isidro, J. *et al.* Genomic Study of a. *Front Microbiol* **9**, 2994, doi:10.3389/fmicb.2018.02994 (2018).
153 Chen, W., Zhang, Y., Yeo, W.-S., Bae, T. & Ji, Q. Rapid and efficient genome editing in Staphylococcus aureus by using an engineered

CRISPR/Cas9 system. *Journal of the American Chemical Society* **139**, 3790-3795 (2017).
154 Kappel, K. *et al.* Blind tests of RNA-protein binding affinity prediction. *Proc Natl Acad Sci U S A* **116**, 8336-8341, doi:10.1073/pnas.1819047116 (2019).
155 Min, K., Ichikawa, Y., Woolford, C. A. & Mitchell, A. P. Candida albicans Gene Deletion with a Transient CRISPR-Cas9 System. *mSphere* **1**, doi:10.1128/mSphere.00130-16 (2016).
156 Ng, H. D., N. Dramatic Improvement of CRISPR/Cas9 Editing in *Candida albicans* by increased Single Guide RNA Expression. *mSphere*, doi:10.1128/mSphere.00385-16 (2017).
157 Li, X. *et al.* Legacy of land use history determines reprogramming of plant physiology by soil microbiome. *The ISME Journal* **13**, 738-751, doi:10.1038/s41396-018-0300-0 (2019).

APPENDIX

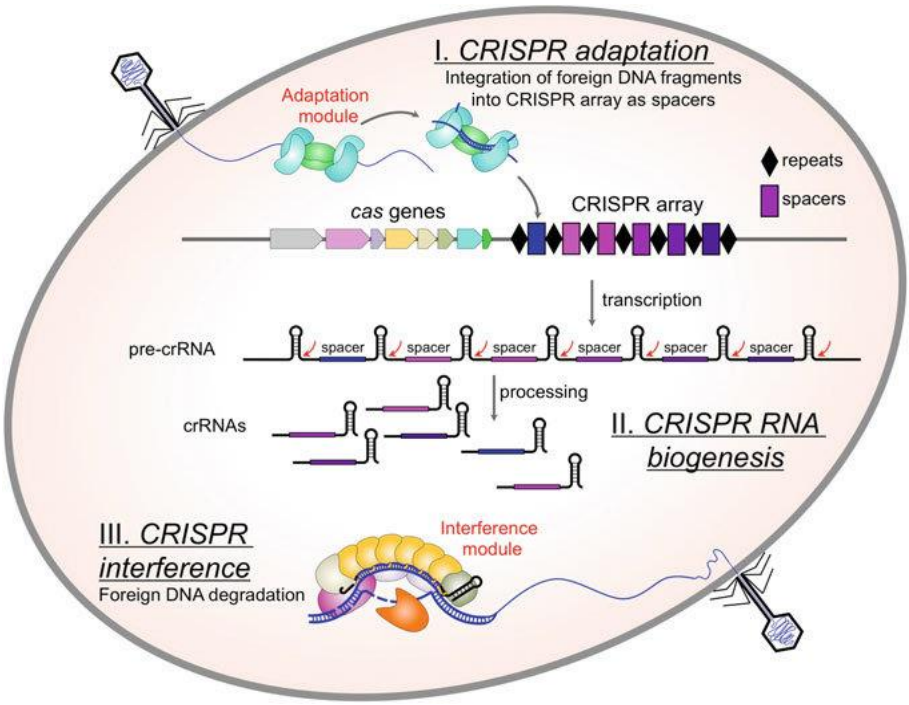


Figure 1 CRISPR-Cas adaptive immunity.

Figure adapted from Semenova and Severinov in "Evolutionary Biology: Self/Nonself Evolution, Species, and Complex Traits Evolution, Methods and Concepts"

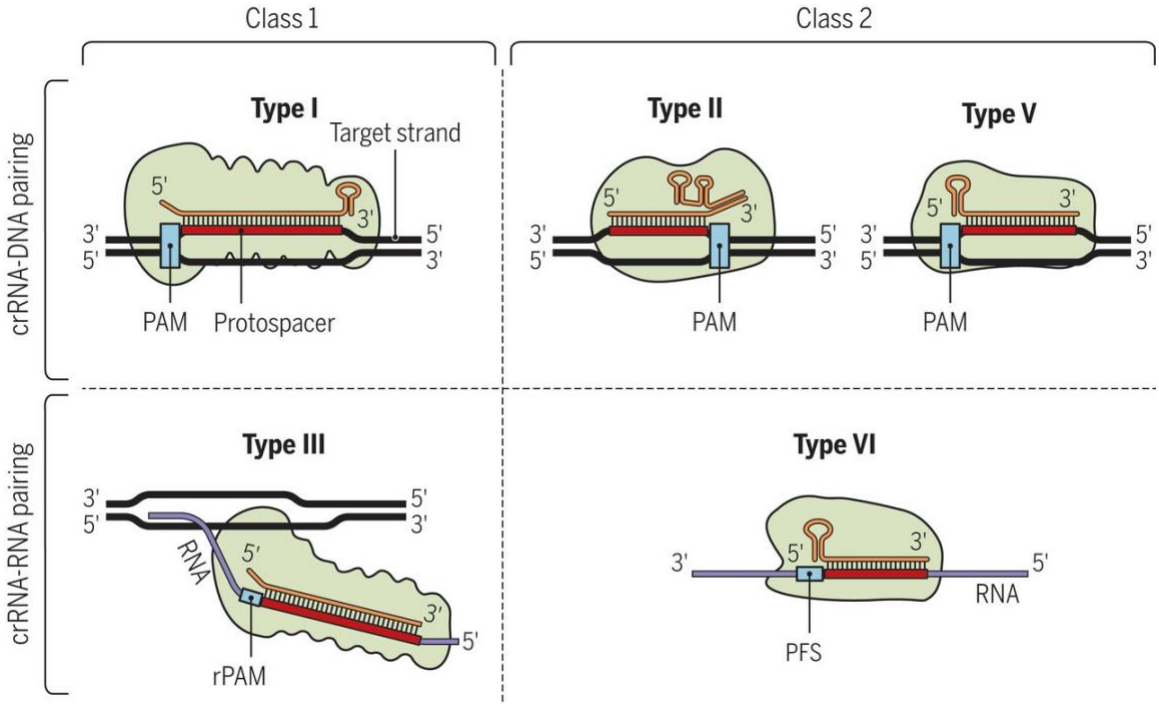


Figure 2 CRISPR-Cas types.

Types of CRISPR-Cas systems and their mechanisms of interaction with target DNA. Adapted from Jackson et. al. (2017) “CRISPR-Cas: Adapting to change.”

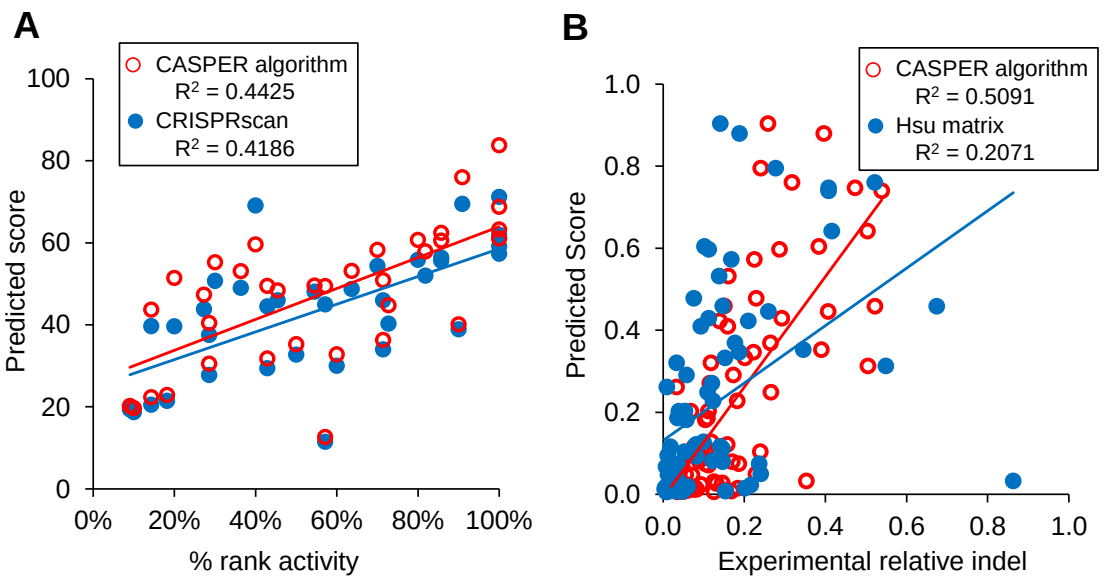


Figure 3 Correlation of CASPER algorithms.

(A) Comparison of CASPER on-target algorithm with CRISPRscan against experimental data obtained from Moreno-Mateos *et al.* [94]. The % rank activity (on-target) is obtained by ranking the activity of a sequence with respect to the rest of other sequences in the sample set. **(B)** Comparison of CASPER off-target algorithm with Hsu matrix method against experimental data obtained from Hsu *et al.* (off-target) [95]. The experimental relative indel (off-target) is determined by dividing the absolute indel % of the off-target site by the indel % of the targeted site. For example, a gRNA that has a 25.0% indel at the targeted site and 1.0% at the off-target site gives a 4% (or 0.04) relative indel.

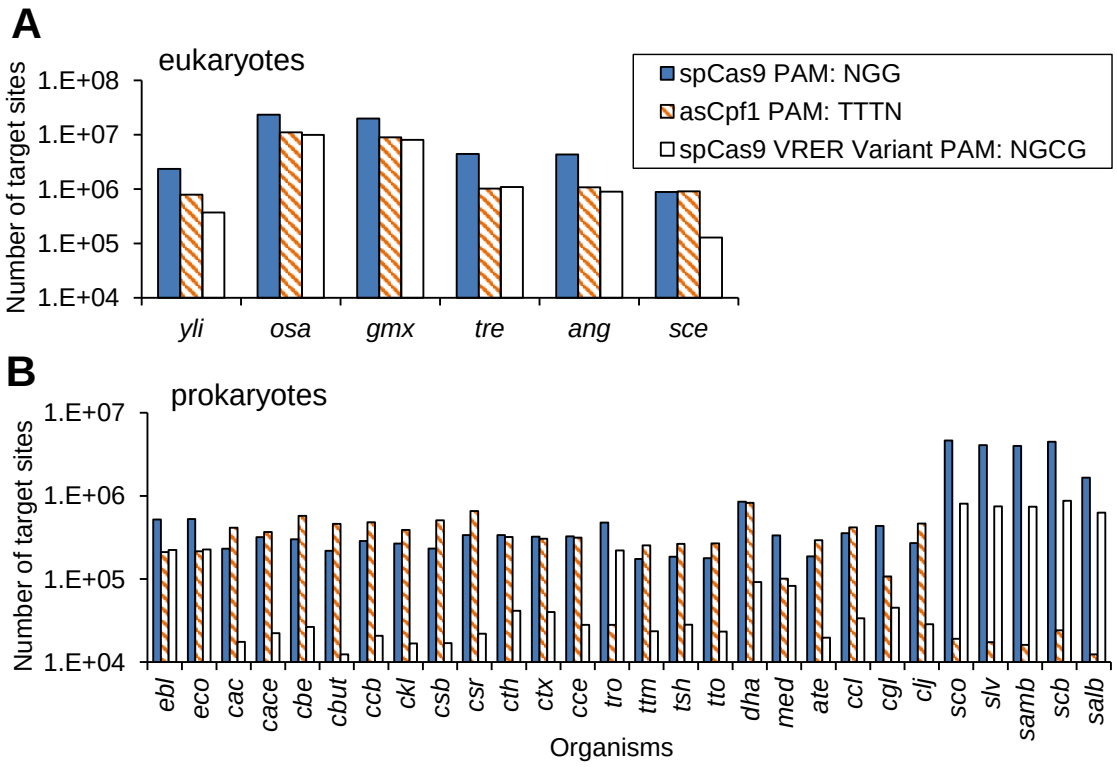


Figure 4 PAM motif appearances.

Appearances of PAM motifs for three common Cas endonucleases (spCas9, asCpf1, and spCas9-VRER variant) across a select number of non-model organisms including **(A)** eukaryotes and **(B)** prokaryotes. Organism abbreviations: *yli: Yarrowia lipolytica, ang: Aspergillus niger, sce: Saccharomyces cerevisiae gmx: Glycine max (soybean) osa: Oryza sativa Japonica (rice) tre: Trichoderma reesei, cac: Clostridium acetobutylicum, cace: Clostridium aceticum, cbe: Clostridium beijerinckii, cbut: Clostridium butylicum, ccb: Clostridium cellulovorans, ckl: Clostridium kluyveri, csb: Clostridium saccharobutylicum, csr: Clostridium saccharoperbuytlacetonium, cth: Clostridium thermocellum (ATCC 27405), ctx: Clostridium thermocellum (DSM 1313), cce: Clostridium cellulolyticum, tro: Thermomicrobium roseum, ttm: Thermoanaerobacterium thermosaccharolyticum (DSM 571), tsh: Thermoanerobacterium saccharolyticum, tto: Thermoanaerobacterium thermosaccharolyticum (M0795), dha: Debaryomyces hanseii, med: Megasphaera elsdenii, ate: Caldicellulosiruptor bescii, ccl: Clostridium clariflavum, cgl: Cornyebacterium glutamicum, clj: Clostridium ljungdahlii, sco: Streptomyces coelicolor, slv: Streptomyces lividans, samb: Streptomyces ambofaciens, scb: Streptomyces scabiei, salb: Streptomyces albus.*

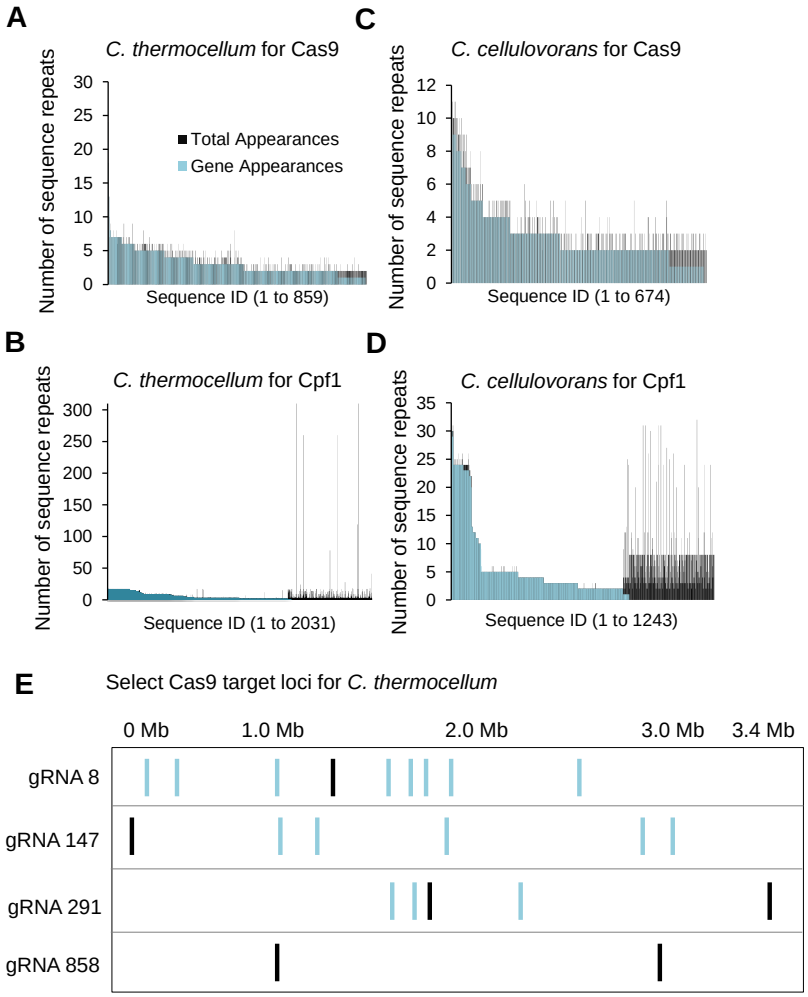


Figure 5 Multitargeting analysis.

Multitargeting analysis for each repeated sequence plotted by the number of times it appears within a gene (blue-green) and throughout the genome (black) for **(A)** *C. thermocellum* using spCas9, **(B)** *C. thermocellum* using asCpf1, **(C)** *C. cellulovorans* using spCas9, and **(D)** *C. cellulovorans* using asCpf1. Parentheses indicate the number of repeated sequences. **(E)** The distribution across the genome of *C. thermocellum* using a set of 4 representative non-unique sequences, including gRNA8: ACGTCAAACCTTTGGGCATT, gRNA147: GCAGAGCTGGATGAACATCT, gRNA291: TCCGTCGGTATCGGCTCCTC, and gRNA858: GACGGAGTCGGTTCATCAGA. Blue-green lines indicate appearance within a gene sequence while black lines represent appearances in un-annotated regions.

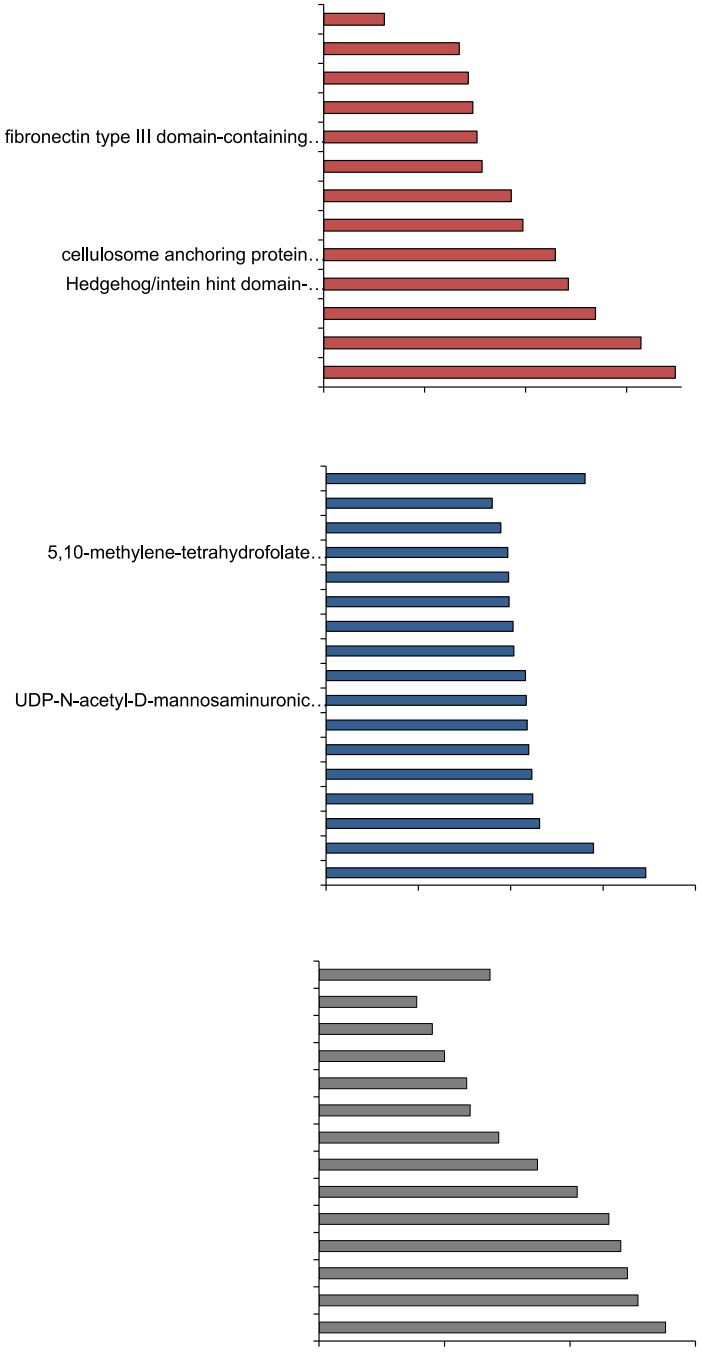


Figure 6 Multitargeting cellular processes.

Annotated cellular processes targeted by repeated sequences for **(A)** *C. thermocellum,* **(B)** *C. acetobutylicum,* and **(C)** *C. kluyveri*.

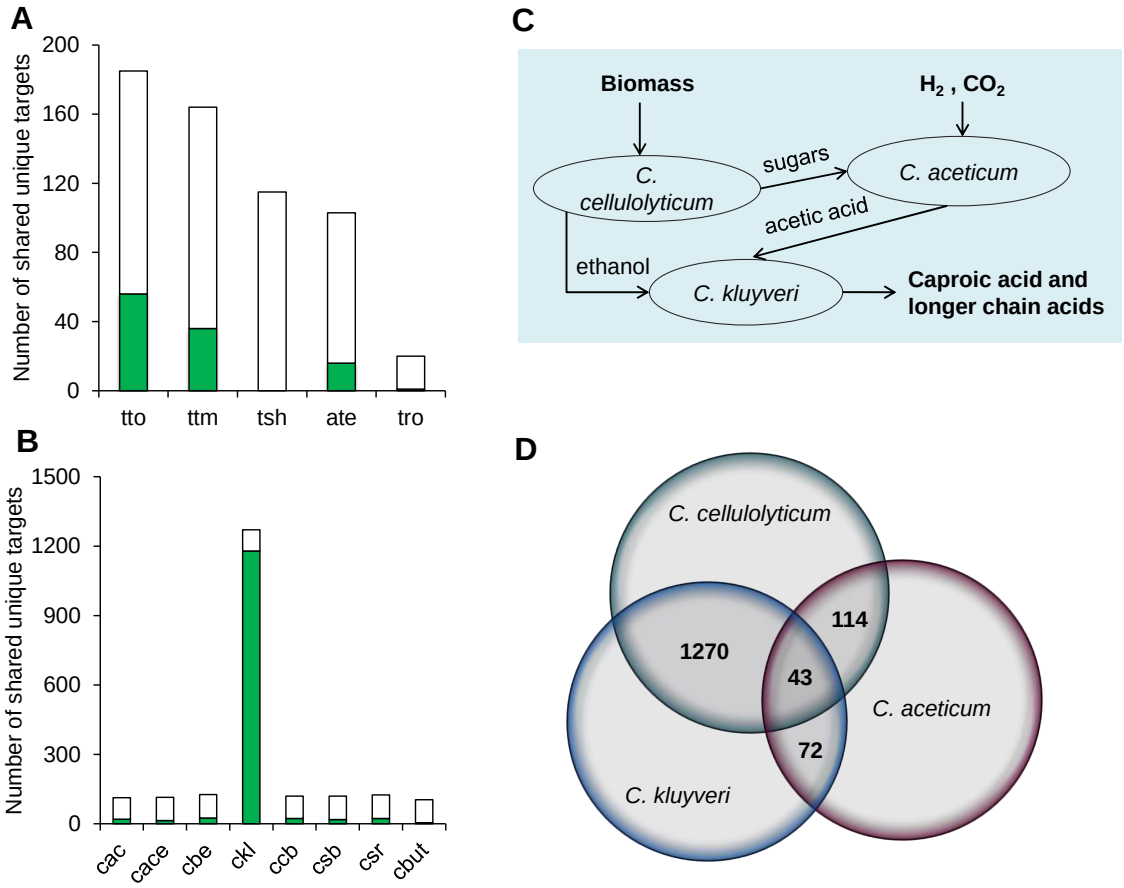


Figure 7 Population targeting.

Comparison of the number of shared targets between **(A)** *C. thermocellum* and other thermophiles and **(B)** *C. cellulolyticum* and other mesophiles. The green portion of the bar represents the number of sequences that appear in both genomes but are unique within each genome. White portion of the bar represents the number of sequences that appear in both genomes and is repeated within one or both genomes being compared. **(C)** A hypothetical synthetic consortium of three different *Clostridial* species with synergistic metabolisms. **(D)** The Venn diagram represents a three-way comparison of these organisms with the numbers of sequences shared among the genomes shown at the intersections.

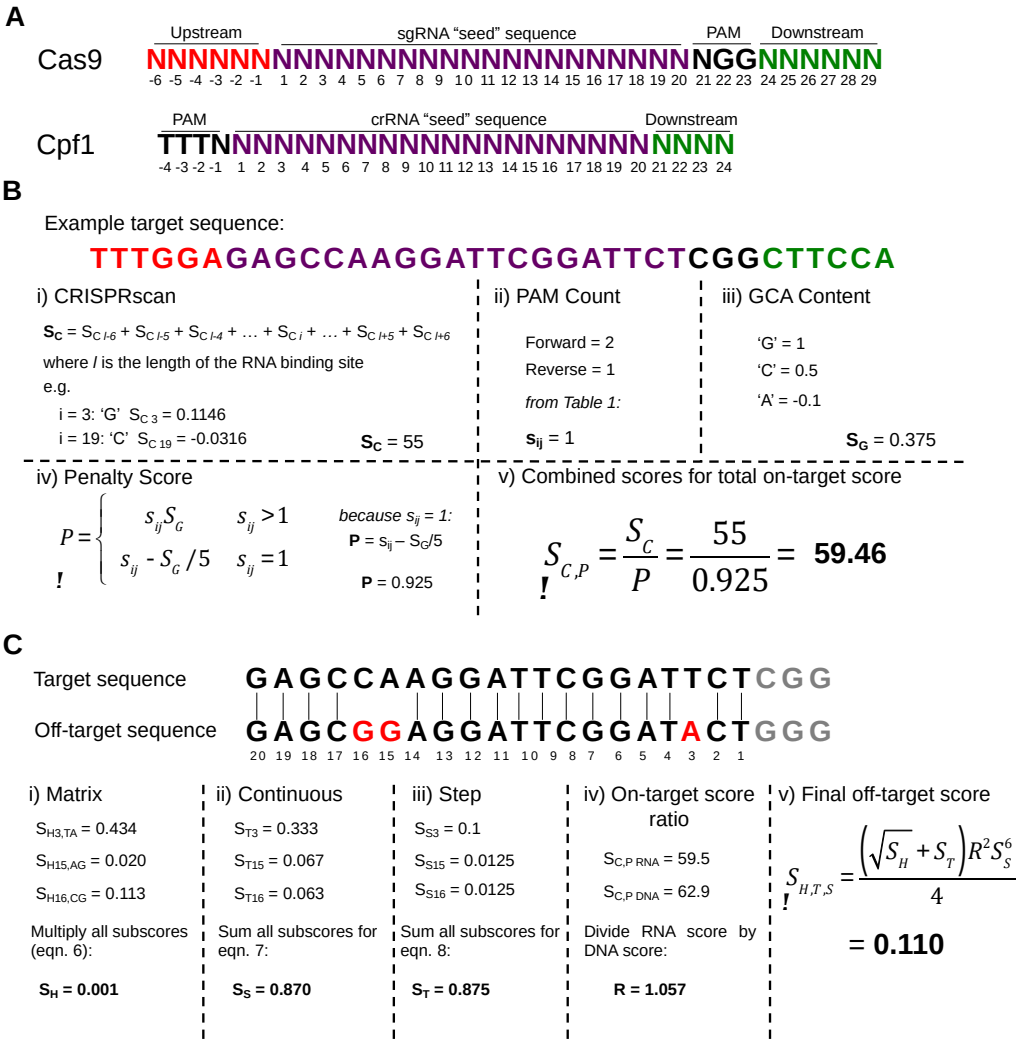


Figure 8 Algorithm schematic.

(A) Nucleotide indexing of the target sequences for Cas9 and Cpf1 used in the development of the CASPER on- and off-target algorithms. 'N' stands for any nucleotide. **(B)** Demonstration of the on-target algorithm on a sample target sequence. The on-target score has a range of 0-100, with 100 being most active. **(C)** Demonstration of the off-target algorithm on a sample off-target sequence. The off-target has a range of 0-1, with 1 indicating a high probability of off-target activity.

PAM density in the reverse 5'→ 3' sequence	PAM density in the forward 5'→ 3' sequence						
	0	1	2	3	4	5	6
0	1	1	1	2	2	8	10
1	1	1	1	2	2	8	10
2	1	1	1	2.5	2.5	10	N/A
3	1	1	1.66	2.5	2.5	N/A	N/A
4	1.66	1.66	3.33	3.33	N/A	N/A	N/A
5	2.5	2.5	3.33	N/A	N/A	N/A	N/A
6	5	5	N/A	N/A	N/A	N/A	N/A

Figure 9 PAM density scoring table.

Scores are the derived from the fold-change in cleavage. Efficiency of the sequence from the 0 PAM control[100].

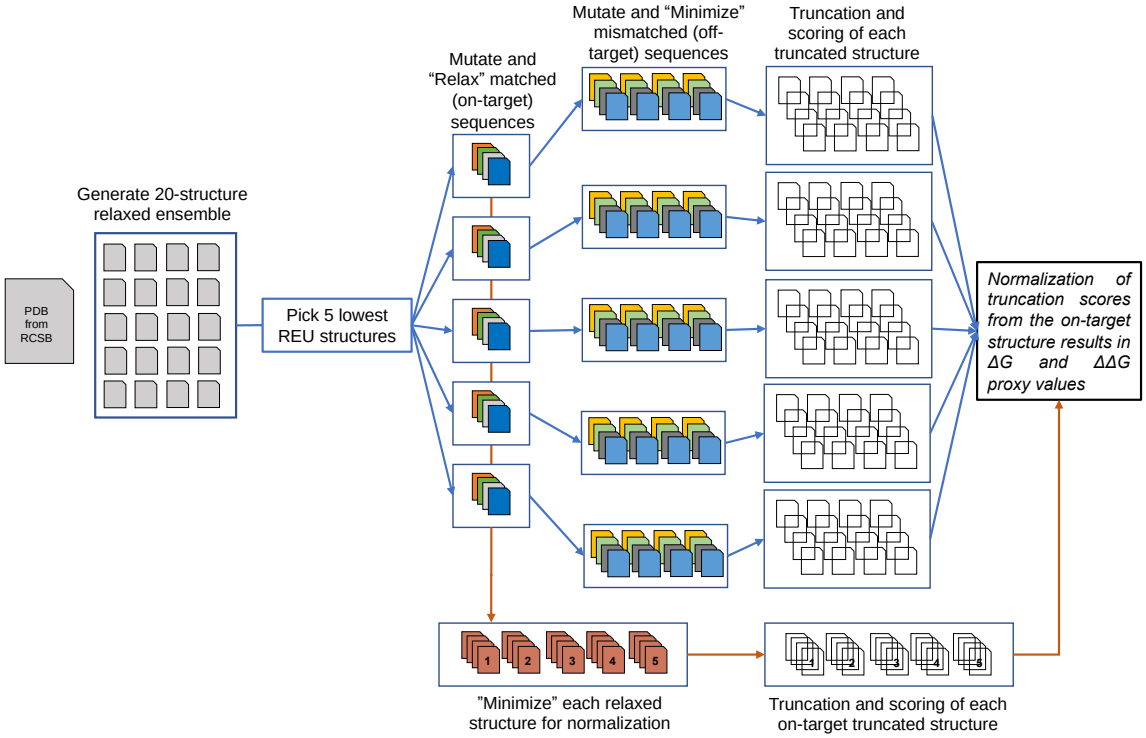


Figure 10 Rosetta computational workflow.

A 20-structure was first generated with the Rosetta 'relax' algorithm. The structures with the lowest Rosetta Energy Units (REU) for the total structure were then used as base structures for 5 ensembles. The base structure of each ensemble was then mutated with the 'rna_mutate' algorithm to match the on-target structures. These structures were then passed through both 'relax' and 'minimize' algorithms. The relaxed structure was passed through another round of mutation to generate the off-target structures, which were further minimized. The DNA molecules of the minimized structures were then successively truncated and scored to generate the proxy for the thermodynamic path of R-loop formation.

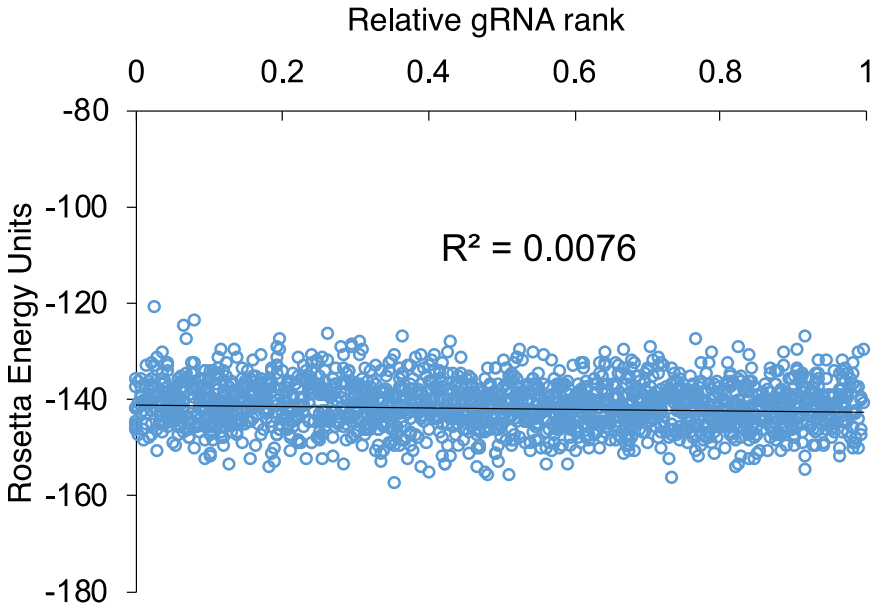


Figure 11 On target activity lack of correlation.

Using the dataset from Moreno-Mateos et al.[31] neither RNA, DNA, or total structure scores were able to recreate correlation of the experimental results from this study, suggesting the protein simulation is a poor approximation of absolute activity and should be used only to determine relative off-target activity where an R-loop simulation proxy can be utilized. Shown here is the cumulative nucleic acid Rosetta Energy Units and the relative sgRNA activity rank.

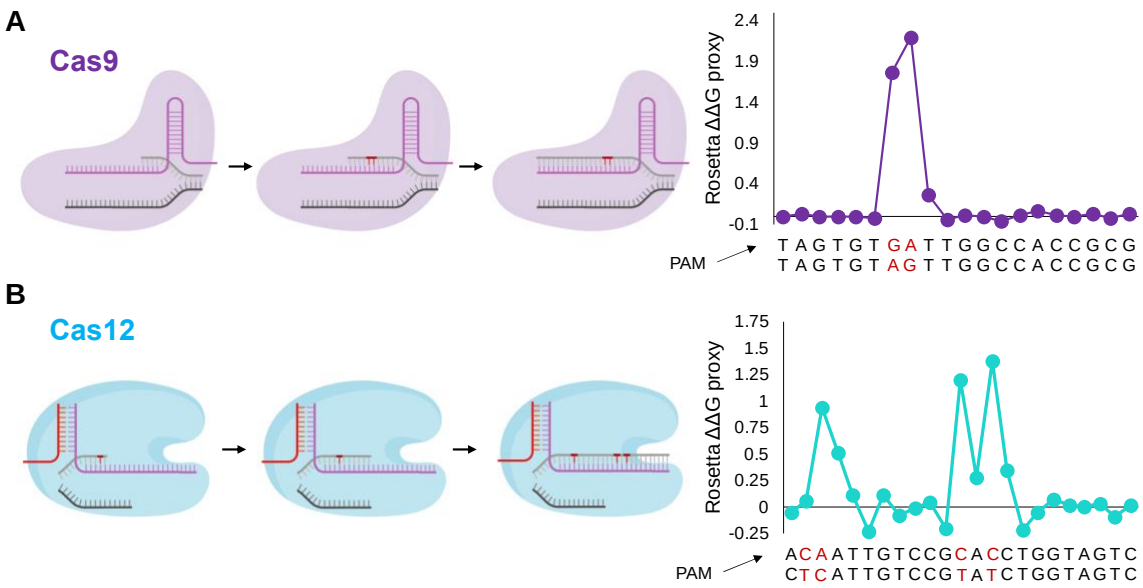
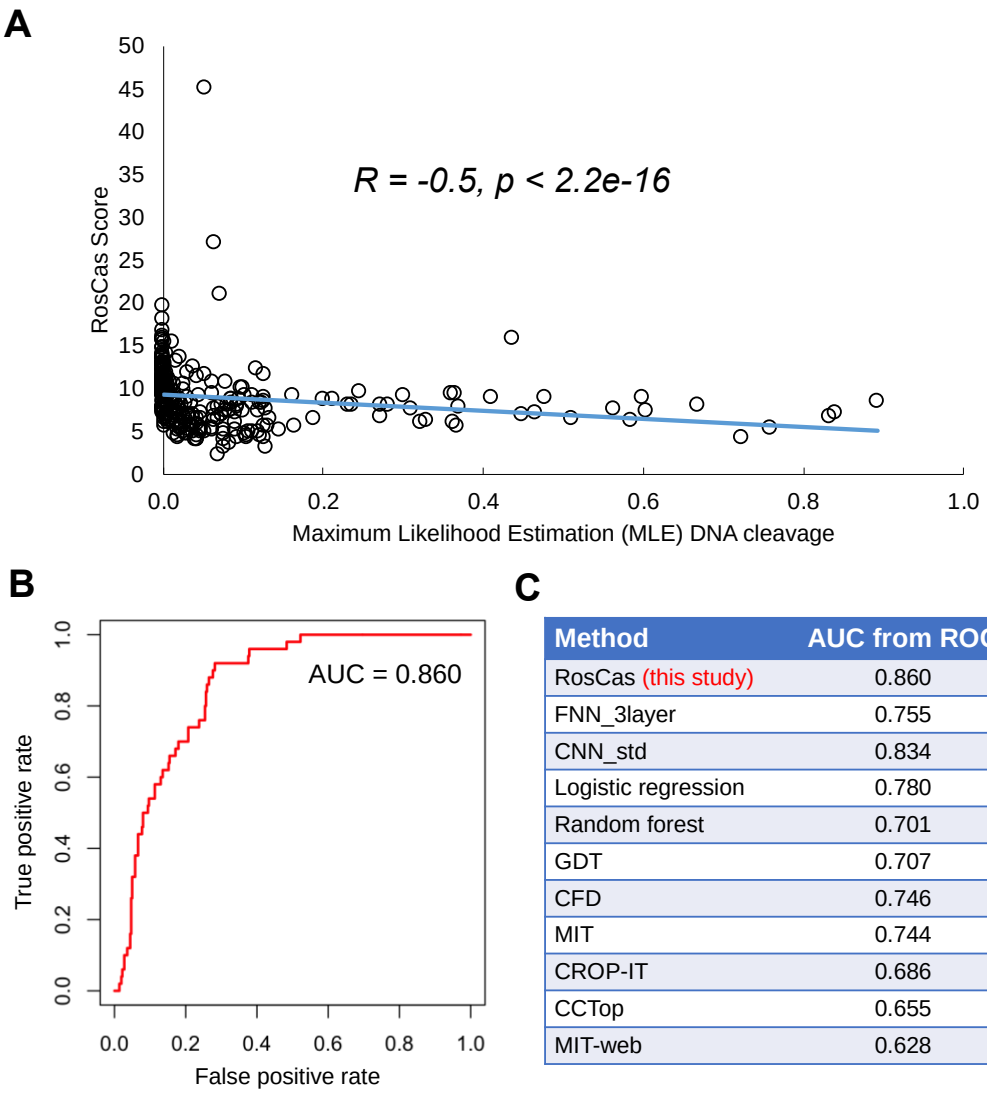


Figure 12 Truncation schematic.

RNA truncation method to simulate movement across the R-loop transition for **(A)** Cas9 binding and **(B)** Cas12 binding with sample truncation structures. Truncating the RNA-binding DNA can be used as an approximation of R-loop formation. The difference between the on-target structure at the corresponding truncation is used to subtract the contributions to the REU value from the addition of additional residues to the structure. The ΔΔG and ΔG equivalents used in the truncation plots are generated from the difference between the change of the two structures and summation of the changes respectively.

Table 1: PDBs used in Rosetta study.

RCSB	Cas Enzyme	PAM	RNA length	Targeted DNA length	Complementary DNA length
5F9R	spCas9	TGG	116	30	30
5XUS	lbCas12	TTTA	39	29	9
5XUT	lbCas12	TCTA	39	29	9
5XH6	asCas12	TATA	42	29	10
5XH7	asCas12	TCCA	42	29	10



Method	AUC from ROC
RosCas (this study)	0.860
FNN_3layer	0.755
CNN_std	0.834
Logistic regression	0.780
Random forest	0.701
GDT	0.707
CFD	0.746
MIT	0.744
CROP-IT	0.686
CCTop	0.655
MIT-web	0.628

Figure 13 RosCas correlation statistics.

Comparison between RosCas prediction and experimental data for off-target analysis using Cas9 structures. Using the Rosetta parameters described in Equation 3.1 a relatively low spearman correlation (R=0.5) but a high area under curve (AUC) from the relativistic operating curve (ROC) was obtained that outperforms previous algorithms, including state-of-the-art convolutional neural networks. **(A)** Correlation between RosCas and MLE. **(B)** Relative operating characteristic curve. **(C)** Comparison between RosCas algorithm with other state-of-the-art techniques using the same Hsu *et al.* dataset[30,36,44,45,120].

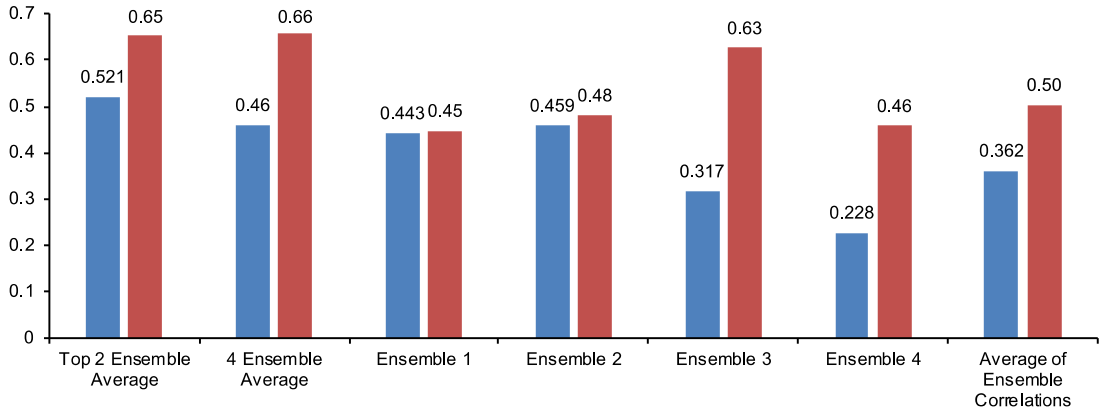


Figure 14 Ensemble sensitivity.

Pearson (Blue) and Spearman (Red) Ensemble Correlations of spCas9 Single Mismatch Dataset (Hsu *et al*, 2013)[36].

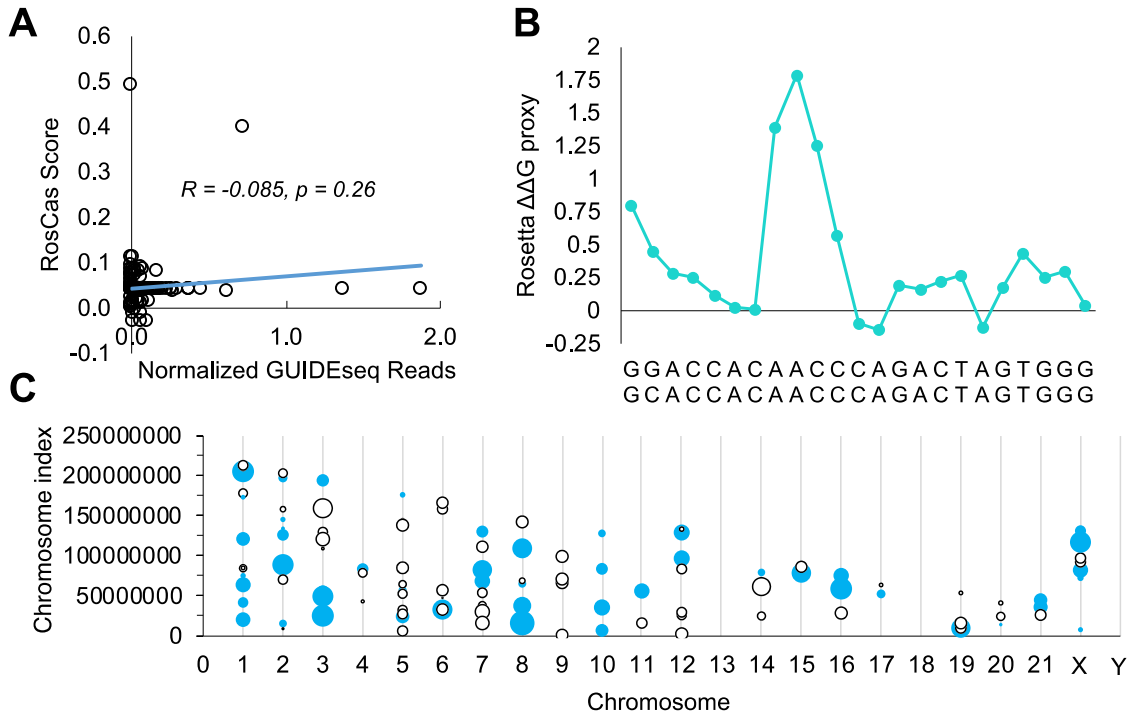


Figure 15 lbCas12 correlation and statistics.

Fitting results from Cas12 structures to experimental data. **(A)** The size and degenerate target site nature of the Kleinstiver et al. dataset[35] creates poor correlation statistics. **(B)** Although a mismatch occurred 2 bases from the PAM site, the thermodynamic effect was delayed by approximately 7 nucleotides, suggesting mismatches can be more than just locally destabilizing. **(C)** By using the simulation scores as a guide, the difference of "targetablility" between different sites across the genome can be elucidated, highlighting this method's ability to interrogate relative loci activity creating a genome-wide profile of activity for future guide-RNA design.

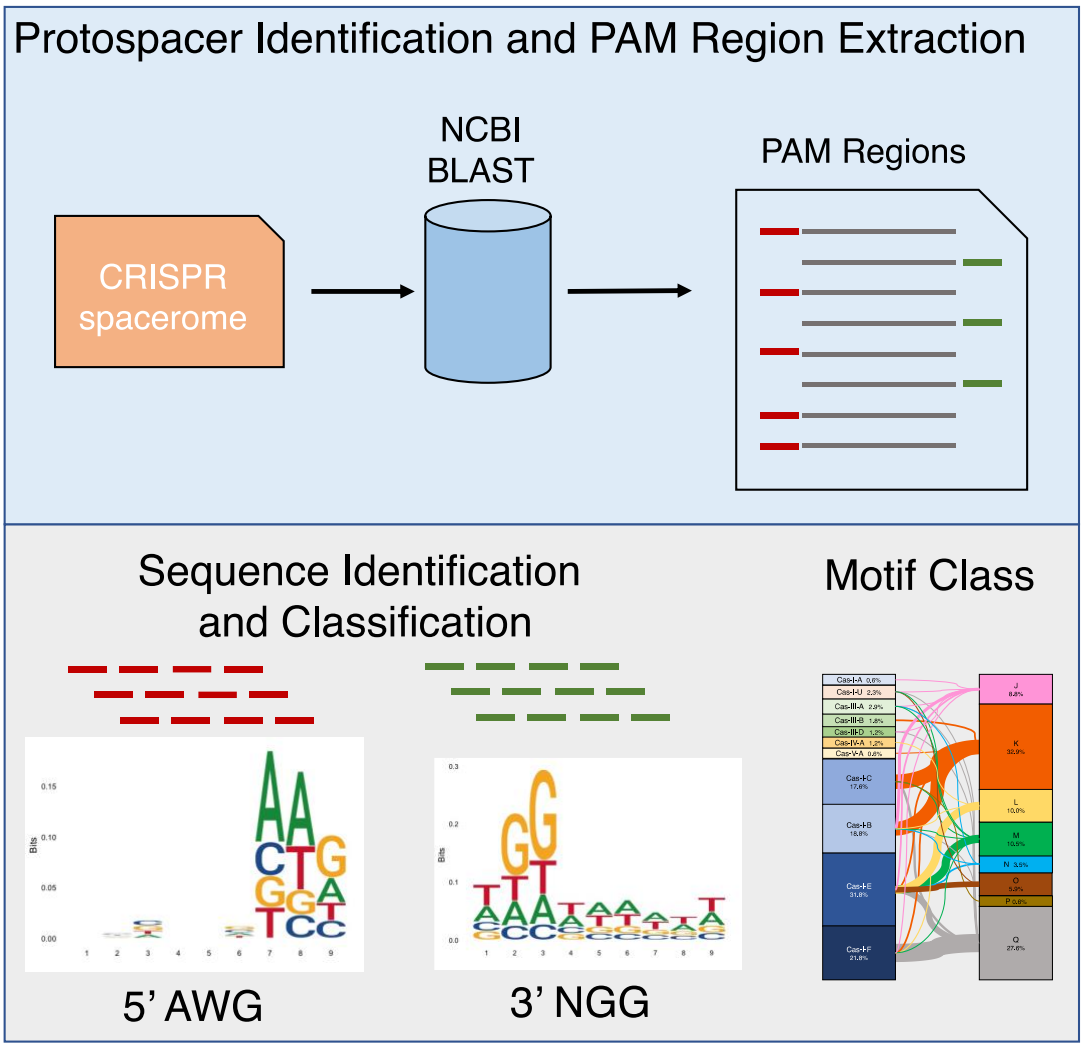


Figure 16 Schematic of CASPERpam workflow.

The previously identified CRISPR spacerome is re-BLASTED to obtain adjacent sequence regions, which then are aligned and classified into motif categories.

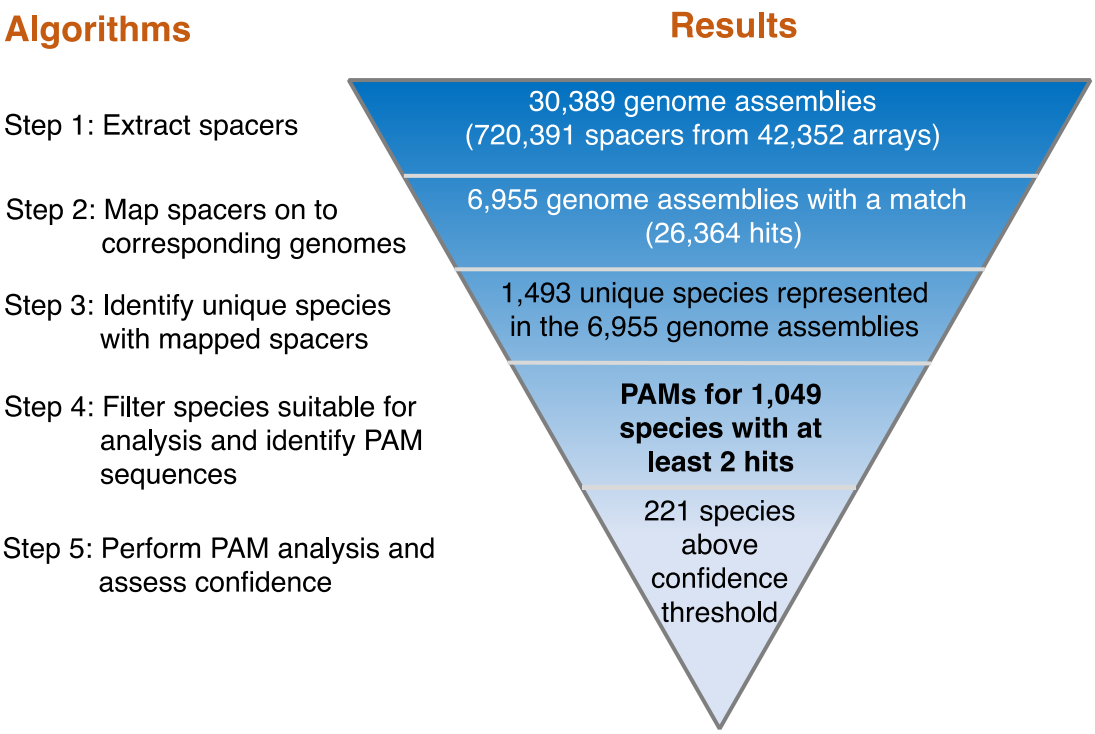


Figure 17 CASPERpam algorithm for identifying PAM sequences.

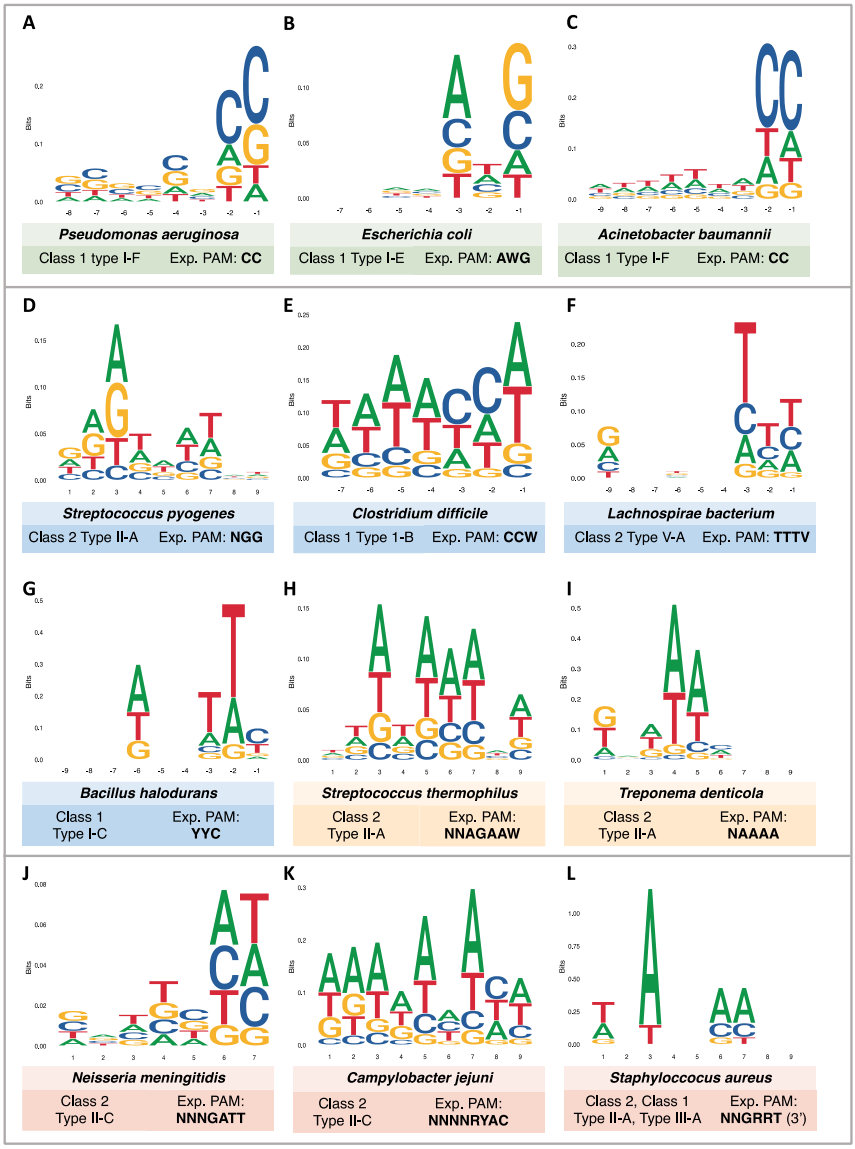
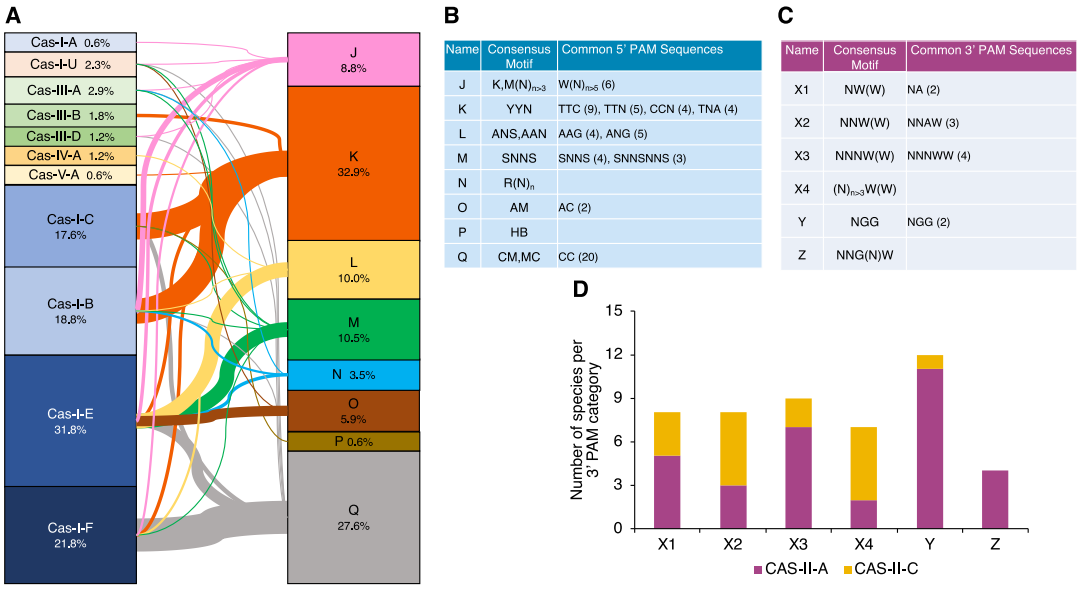


Figure 18 Predicted PAM for validated species.

Predicted PAM sequences for experimentally validated species including **(A)** Pseudomonas aeruginosa, **(B)** Escherichia coli, **(C)** Acinetobacter baumannii, **(D)** Streptococcus pyogenes, **(E)** Clostridium difficile, **(F)** Lachnospiracae bacterium, **(G)** Bacillus halodurans, **(H)** Streptococcus thermophiles, **(I)** Treponema denticola, **(J)** Neisseria meningitides, **(K)** Campylobacter jejuni, and **(L)** Staphyloccocus aureus.



B

Name	Consensus Motif	Common 5' PAM Sequences
J	$K,M(N)_{n>3}$	$W(N)_{n>5}$ (6)
K	YYN	TTC (9), TTN (5), CCN (4), TNA (4)
L	ANS,AAN	AAG (4), ANG (5)
M	SNNS	SNNS (4), SNNSNNS (3)
N	$R(N)_n$	
O	AM	AC (2)
P	HB	
Q	CM,MC	CC (20)

C

Name	Consensus Motif	Common 3' PAM Sequences
X1	NW(W)	NA (2)
X2	NNW(W)	NNAW (3)
X3	NNNW(W)	NNNWW (4)
X4	$(N)_{n>3}W(W)$	
Y	NGG	NGG (2)
Z	NNG(N)W	

Figure 19 Predicted PAM sequences and categorization.

(A) Mapping PAM sequence categories to CRISPR/Cas types. **(B)** Common patterns and assigned 5' PAM sequence categories. **(C)** Common patterns and assigned 3' PAM sequence categories. **(D)** Number of species per 3' PAM category.

Table 2: Top species by confidence score.

Genus	Species	Cas type	PAM	Exp. PAM	Consensus	Relative PamCoS	Qualitative assessment
Pseudomonas	*aeruginosa*	CAS-I-F	5'	CC	cS	100.0000	match
Acinetobacter	*baumannii*	CAS-I-F	5'	CC	CC	59.3357	match
Salmonella	*enterica*	CAS-I-E	5'	-	ANG	30.0805	
Streptococcus	*agalactiae*	CAS-II-A	3'	-	NgG	24.6686	
Escherichia	*coli*	CAS-I-E	5'	AWG	aNS	20.4289	match
Streptococcus	*suis*	CAS-II-A	3'	-	NgR	16.7142	
Salinispora	*arenicola*	CAS-I-E	5'	-	SNNS	16.1926	
Klebsiella	*pneumoniae*	CAS-I-E	5'	-	aWg	14.3394	
Neisseria	*meningitidis*	CAS-II-C	3'	NNNNGATT	NNNNNMH	13.2987	no match
Vibrio	*cholerae*	CAS-I-E	5'	-	MM	9.4119	
Leptospira	*interrogans*	CAS-I-B	5'	-	ATG	8.5312	
Listeria	*monocytogenes*	CAS-I-B	5'	-	ccW	8.3762	
Clostridium	*botulinum*	CAS-III-B	5'	-	TNW	8.2882	
Fusobacterium	*nucleatum*	CAS-I-B	5'	-	YNW	8.2191	
Enterococcus	*faecalis*	CAS-II-A	3'	-	NgRNW	7.2013	
Clostridium	*difficile*	CAS-I-B	5'	CCW	WNNNW	6.5629	strong partial
Salinispora	*pacifica*	CAS-I-E	5'	-	SNNS	6.4865	
Pasteurella	*multocida*	CAS-I-F	5'	-	Cc	6.3487	
Moraxella	*catarrhalis*	CAS-I-F	5'	-	WNNY	5.5702	
Leptospira	*santarosai*	CAS-I-E	5'	-	AaG	5.5368	
Bifidobacterium	*longum*	CAS-II-C	3'	-	SNNS	5.2296	
Serratia	*marcescens*	CAS-I-F	5'	-	aMS	5.0828	
Clostridioides	*difficile*	CAS-I-B	5'	-	MNW	4.4906	
Bifidobacterium	*breve*	CAS-I-C	5'	-	TTC	3.9543	
Streptococcus	*thermophilus*	CAS-II-A	3'	NNNNGWWT	NNWNWa	3.8955	weak partial
Methanosarcina	*mazei*	CAS-I-B	5'	-	Tta	3.5871	
Streptococcus	*pyogenes*	CAS-II-A	3'	NGG	NNR	3.4889	strong partial
Yersinia	*pseudotuberculosis*	CAS-I-F	5'	-	CC	3.4491	
Fusobacterium	*necrophorum*	CAS-II-A	3'	-	NGG	3.3977	
Propionibacterium	*freudenreichii*	CAS-I-U	5'	-	SSaN	3.0403	
Campylobacter	*coli*	CAS-I-B	5'	-	TNNNNNW	2.9907	
Clostridium	*novyi*	CAS-I-B	5'	-	TNW	2.8437	
Corynebacterium	*ulcerans*	CAS-I-E	5'	-	AC	2.8430	
Salinispora	*tropica*	CAS-I-B	5'	-	SNNS	2.8154	
Lactobacillus	*paracasei*	CAS-II-A	3'	-	TNNAA	2.6750	
Aggregatibacter	*actinomycetemcomitans*	CAS-I-C	5'	-	C	2.6473	
Treponema	*pedis*	CAS-II-C	3'	-	NNNWA	2.6391	
Corynebacterium	*striatum*	CAS-I-E	5'	-	AM	2.5033	
Lactobacillus	*rhamnosus*	CAS-II-A	3'	-	NNNWR	2.4688	
Xenorhabdus	*nematophila*	CAS-I-E	5'	-	AAg	2.4337	
Campylobacter	*jejuni*	CAS-II-C	3'	NNNNRYAC	NNNNWNA	1.5839	no
Lachnospiraceae	*bacterium*	CAS-V-A	3'	TTTV	TNH	0.8278	strong partial
Treponema	*denticola*	CAS-II-A	3'	NAAAAN	NNNWW	0.6881	weak partial
Staphylococcus	*aureus*	CAS-III-A	5'	NNGRRT (3')	TNN	0.2079	no
Bacillus	*halodurans*	CAS-I-C	5'	YYC	WNNTTN	0.1347	strong partial

Lower case letter indicates a frequency between 0.25 and 0.5. Degenerate Codes: W = A/T; S = G/C; R = A/G; Y = C/T; K = G/T; M = A/C; B = not A; D = not C; H = not G; V = not T.

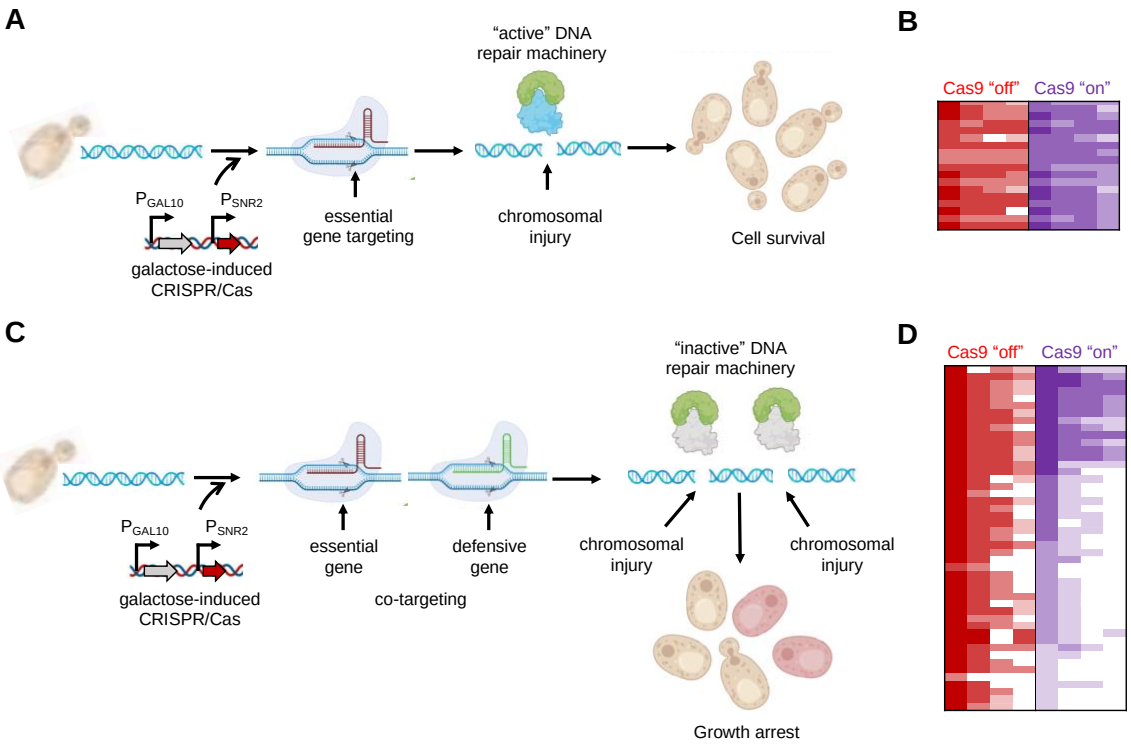


Figure 20 Schematic of co-targeting and agar serial dilution heat maps.

Co-targeting mechanism for increased growth suppression/cell death. **(A)** Targeting a single essential gene locus results in cell survival due to active DNA repair machinery. **(C)** Co-targeting parts of the DNA repair response reduce the ability to repair legions at essential and defensive sites, increasing lethality via essential gene knockout and DSB-induced apoptosis. **(B, D)** Heat maps of serial dilution results on agar plates of essential and defensive gene combinations. Non-induced strains grown on glucose plates are indicated in red while induced Cas9 strains grown on galactose pads are indicated in purple. Larger colony size is represented by darker coloration.

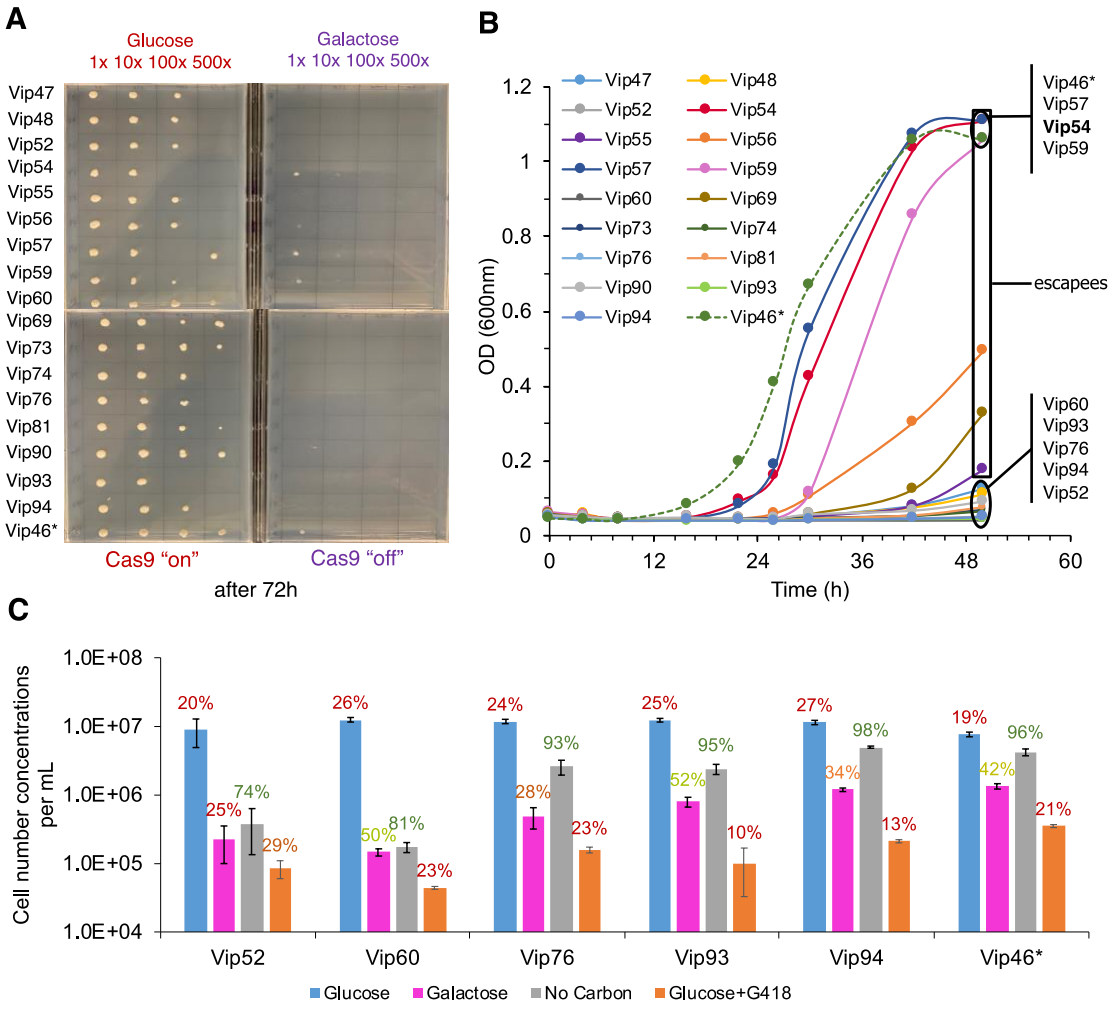


Figure 21 Detailed characterization of 18 strains.

(A) All selected strains exhibit extreme loss of growth on galactose (Cas9-induced) plates. **(B)** In liquid media however, 3 strains exhibit a return to a normal growth phenotype as compared to the strain targeting two defensive genes (Vip46 RAD51/RAD52). Three strains show a growth lag but begin to enter exponential growth after 48 hours outgrowth. The remaining 11 strains show growth inhibition. **(C)** Selected strains exhibit a viability profile in between the no carbon (starvation) and antibiotic profiles as determined by flow cytometry.

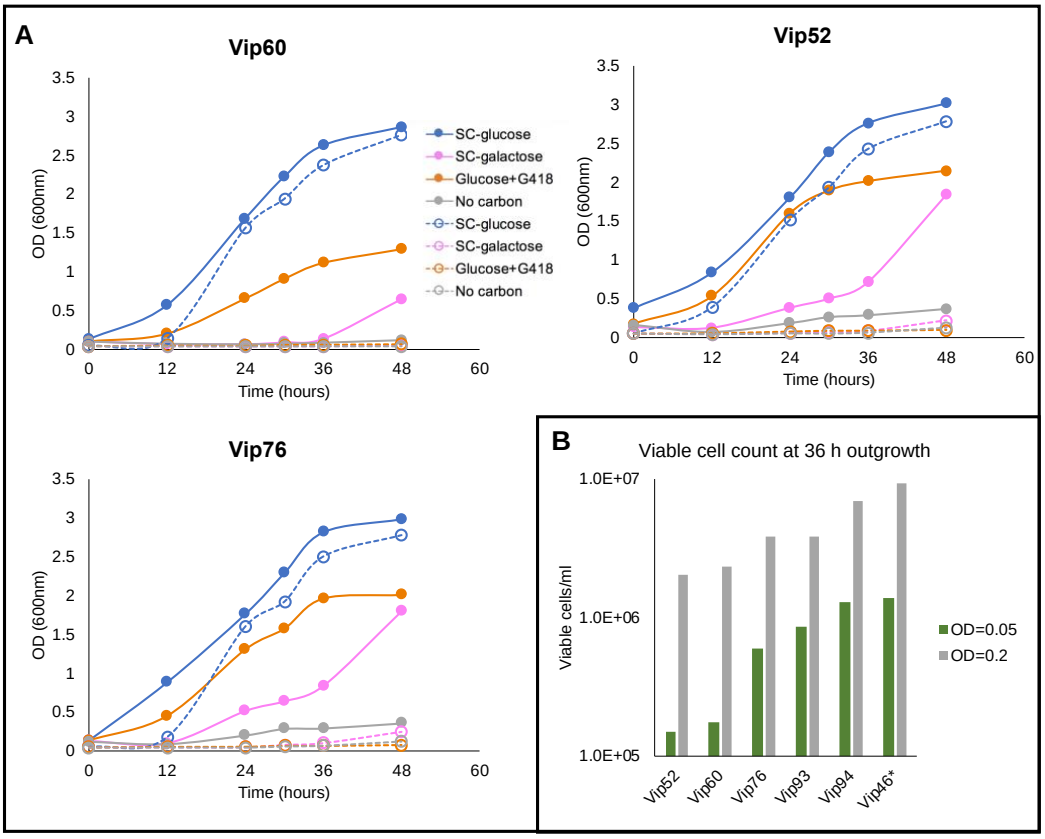


Figure 22 Effect of inoculation concentration on killing efficiency.

(A) Inoculation of three of our high performing strains using either OD = 0.05 or OD = 0.2. Higher concentration of cells at the outset of the experiment can overwhelm the population by decreasing lag time, thereby increasing both antibiotic tolerance and tolerance to growth suppression by our CRISPR-Cas antifungal design. **(B)** Flow cytometry measurements confirm the increase in viable cells in the high inoculation scenario.

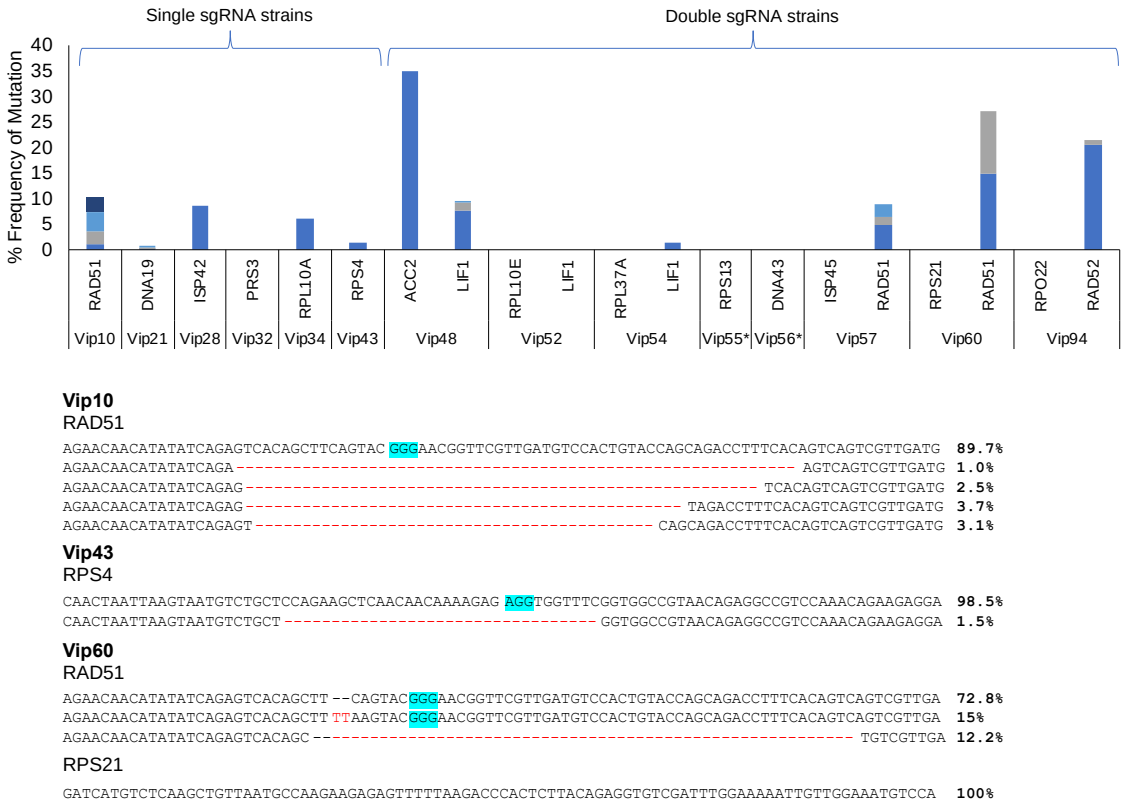


Figure 23 Targeted amplicon sequencing results.

Oxford Nanopore ligation sequencing kit was used to sequence PCR amplification of targeted genes. Single-targeted strains show low-levels of mutation with the exception of the defensive genes, in particular RAD51. As seen in the sequence alignments provided for Vip10, Vip43, and Vip60, mutants are dominated by large (>20bp) deletions, although mutation frequency is relatively low and heterogeneous. Essential genes in double-targeted strains show no significant mutation with the exception of ACC2 in Vip48.

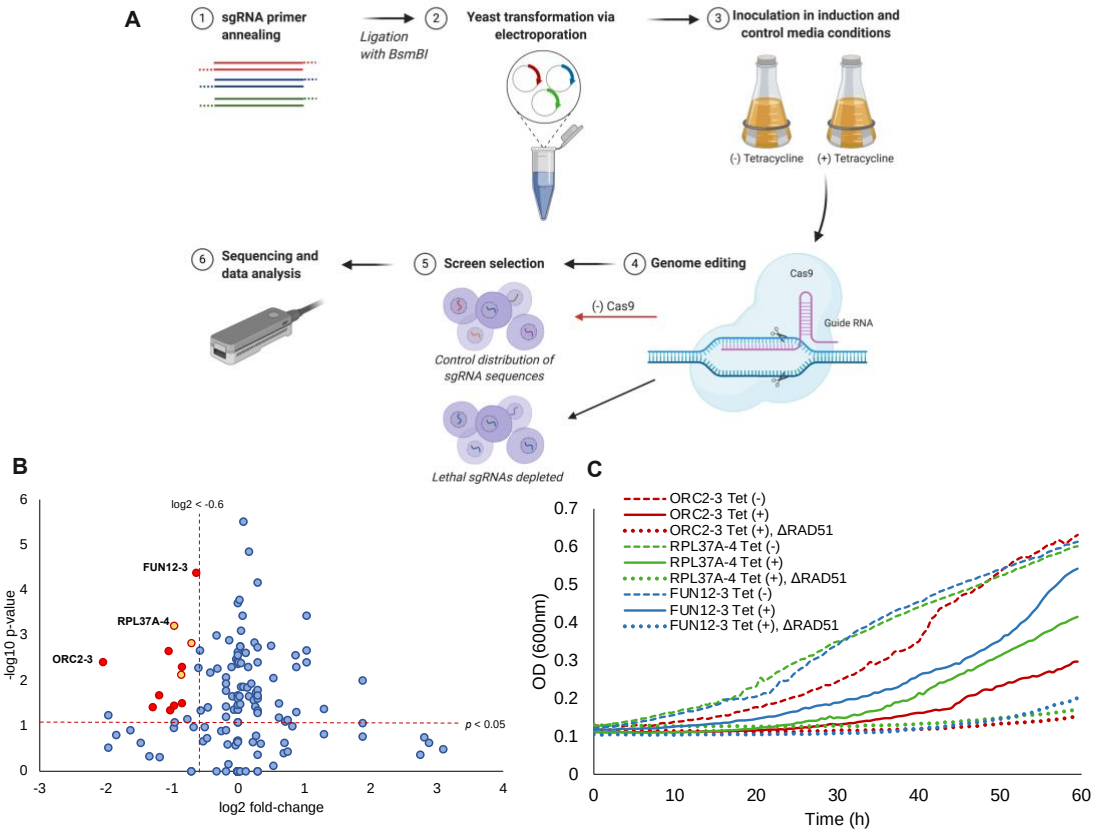


Figure 24 Single plasmid library and characterization.

(A) Workflow for the library depletion assay of 184 sgRNAs. **(B)** Volcano plot of a library depletion/enrichment assay of 185 guide-RNAs spanning 37 essential genes. Library was transformed into *S. cerevisiae* BY4741 and grown for 48h in selective media with and without 5 ug/ml tetracycline (for Cas9 induction). Log2 fold-changes represent the change in read counts between the control culture and induced culture. Fold-changes were averaged over 4 replicates. Significantly depleted (p-value < 0.05 and fold-change < -0.6) sgRNA sequences are highlighted in red. RPL37A (yellow, red outline) was the only gene to appear multiple times in the cutoff region, which agrees with our results in our two-plasmid system screen. **(C)** Growth characterization over 60 hours of three high performing targets identified by the library screen. In all cases co-targeting with RAD51 sgRNA increases growth suppression.

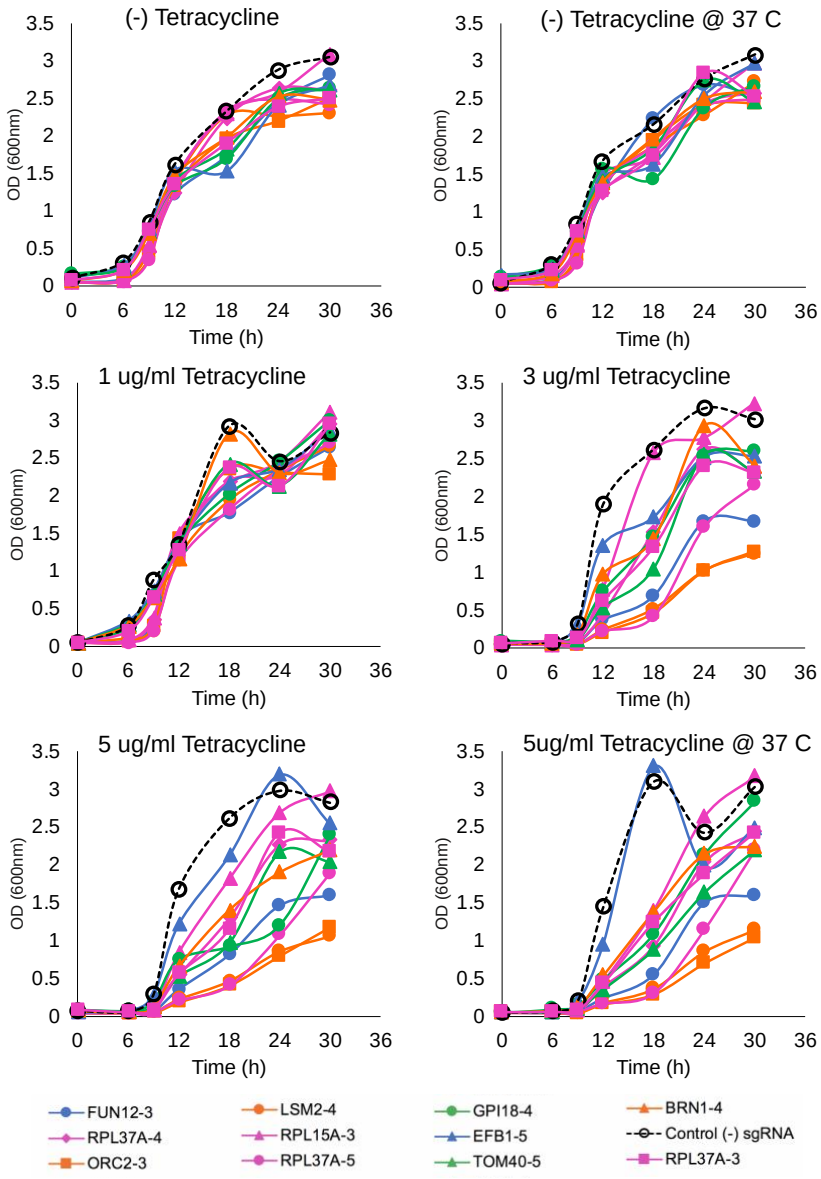


Figure 25 Characterization of library-identified sgRNAs.

Effect of temperature and tetracycline concentration on single essential genes identified from library screen. Increasing dosage of tetracycline increases expression of Cas9 and results in increased growth suppression. The difference of growth suppression between 3 and 5 ug/ml tetracycline is minimal. Increasing temperature to the optimal spCas9 temperature (37ºC) results in only a slight increase in killing efficiency.

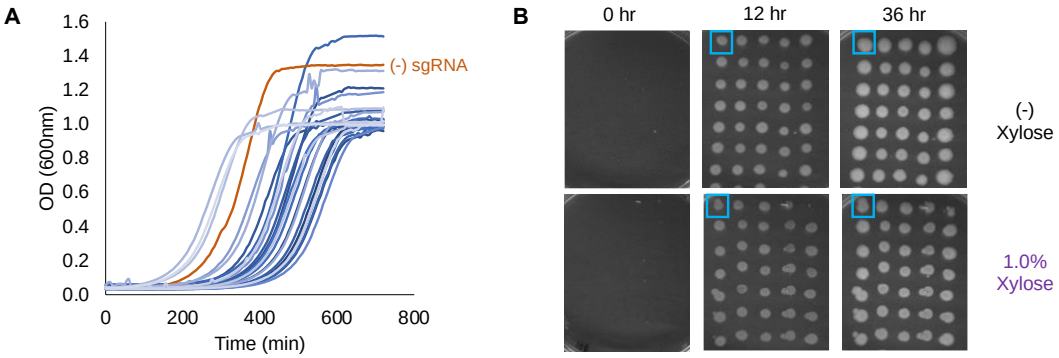
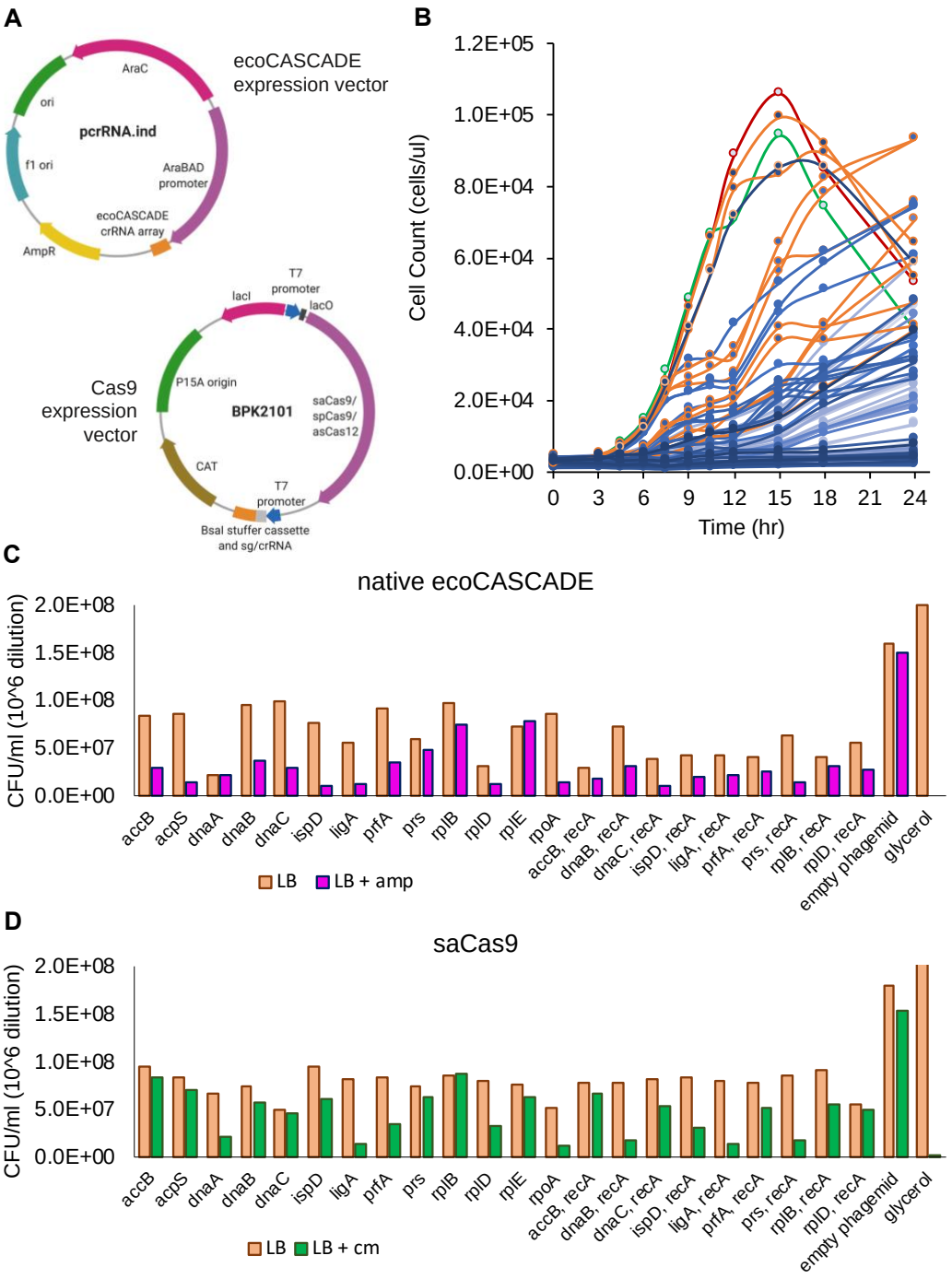


Figure 26 Growth characterization of *B. subtilis*.

Measurements in **(A)** liquid LB media containing 1% xylose using OD **(B)** agar plating show robust escape mechanisms in *B. subtilis* in the presence of Cas9 with sgRNAs targeting essential genes. Cas9 is expressed under the P_{xyl} promoter and strongly induced in the presence of 1% xylose. Increased lag time compared to the empty sgRNA vector and a slight decrease in colony size compared to the media absent xylose suggest some effect of the Cas system on growth.

Figure 27 Characterization of saCas9 and ecoCASCADE.

E. coli was used as a subject to compare the differences in utilizing endogenous or heterologous Cas systems. **(A)** Vectors used for expressing the native crRNA array (pcrRNA.ind$_{65}$) and the Cas protein with its respective sgRNA (BPK2101$_{97}$). **(B)** Liquid growth (LB+amp/cm) of Cas9 (orange) and ecoCASCADE (blue) carrying strains with sgRNAs targeting essential gene loci. Control strains containing the empty plasmid backbones are shown in red (BPK2101) and green (pcrRNA.ind) respectively. **(C)** Colony counts post-12-hour liquid outgrowth from strains infected with M13 phages carrying crRNAs targeting essential genes and co-targeted with a crRNA targeting the recA gene using the ecoCASCADE system. Liquid outgrowth samples were plated on either LB (orange) or LB+amp (purple) plates to assess the effect of removing selective pressure for maintaining the CRISPR-Cas machinery. **(D)** Colony counts post-12-hour liquid outgrowth of strains infected with M13 phages carrying saCas9 machinery targeting essential and essential + recA sgRNAs.


A
ecoCASCADE
expression vector
AraC
ori
pcrRNA.ind
f1 ori
AraBAD
promoter
ecoCASCADE
crRNA array
AmpR
Cas9
expression
vector
T7
promoter
lacI
lacO
P15A origin
BPK2101
saCas9/
spCas9/
asCas12
CAT
T7
promoter
BsaI stuffer cassette
and sg/crRNA
B
Cell Count (cells/ul)
1.2E+05
1.0E+05
8.0E+04
6.0E+04
4.0E+04
2.0E+04
0.0E+00
0
3
6
9
12
15
18
21
24
Time (hr)
C
native ecoCASCADE
CFU/ml (10^6 dilution)
2.0E+08
1.5E+08
1.0E+08
5.0E+07
0.0E+00
accB
acpS
dnaA
dnaB
dnaC
ispD
ligA
prfA
prs
rplB
rplD
rplE
rpoA
accB, recA
dnaB, recA
dnaC, recA
ispD, recA
ligA, recA
prfA, recA
prs, recA
rplB, recA
rplD, recA
empty phagemid
glycerol
LB
LB + amp
D
saCas9
CFU/ml (10^6 dilution)
2.0E+08
1.5E+08
1.0E+08
5.0E+07
0.0E+00
accB
acpS
dnaA
dnaB
dnaC
ispD
ligA
prfA
prs
rplB
rplD
rplE
rpoA
accB, recA
dnaB, recA
dnaC, recA
ispD, recA
ligA, recA
prfA, recA
prs, recA
rplB, recA
rplD, recA
empty phagemid
glycerol
LB
LB + cm

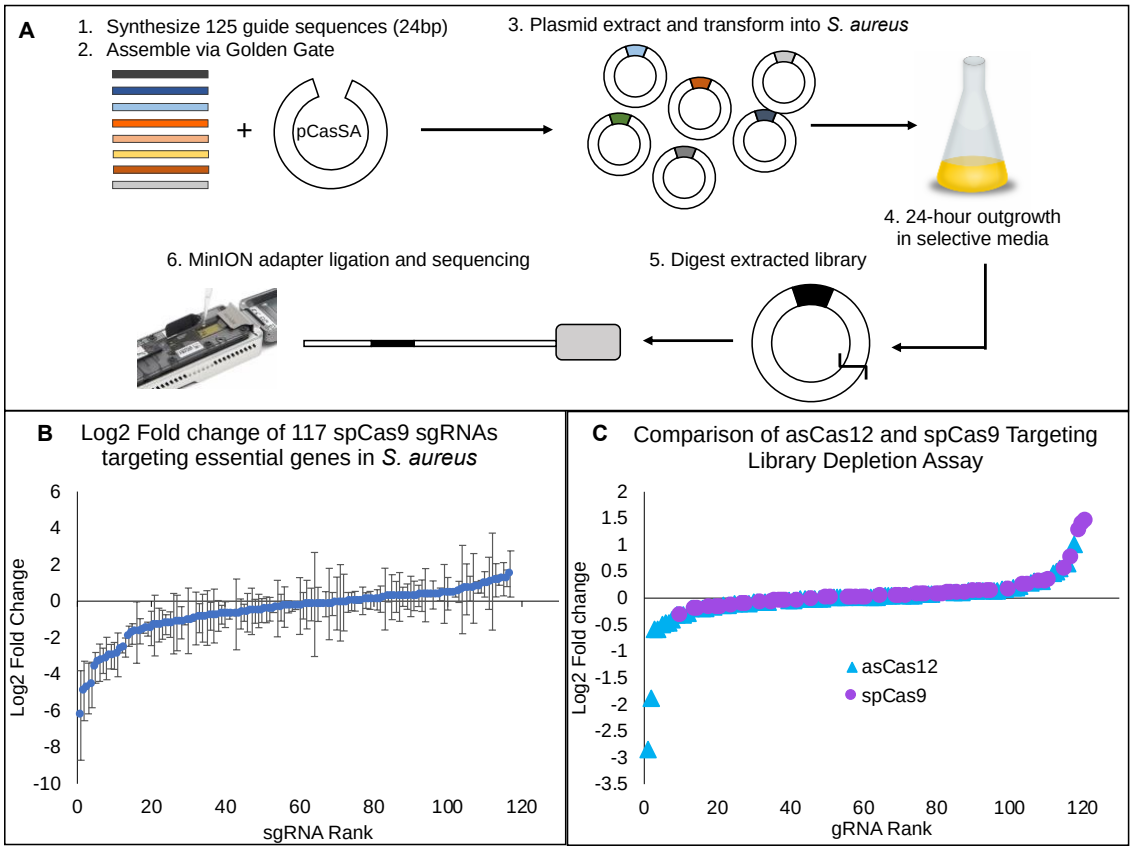


Figure 28 Library depletion assays in *S. aureus.*

(A) Workflow for implementing a library depletion assay in *S. aureus.* **(B)** Library results from spCas9 system show significant log2 fold change (<-2) for a subset of 16 guides after 24-hour outgrowth in selective media (5 ug/ml chloramphenicol). **(C)** asCas12 library of guides significantly outperforms the high-performing spCas9 library subset, suggesting asCas12 is a more potent choice of targeted antimicrobial.

Table 3 Log2 fold changes for most depleted sgRNA sequences.

Gene-sgRNA #	Log2 fold change
dnaJ-5	-6.2596621
teichB-3	-4.9159227
glnA-5	-4.7822436
teichB-2	-4.5885362
gyrB-2	-3.6478624
dnaJ-4	-3.3721395
prsA-3	-3.2666951
fabF-5	-3.1416913
gyrB-5	-3.0404854
trxB-3	-2.9974578
dnaJ-2	-2.9413006
ftsA-5	-2.6357435
gmk-2	-2.5402509

VITA

Brian is a PhD candidate at the University of Tennessee – Knoxville in the lab of Dr. Cong Trinh. His research focuses on the development of CRISPR-Cas tools for manipulating microorganisms in both antimicrobial and microbiome engineering contexts. His work on antimicrobials is funded by a DARPA YFA (Young Faculty Award) and Director's Fellowship. Previous to his doctoral studies he spent two years working on new market strategies for grid-level energy storage solutions. He graduated from Stanford University in 2012 with a B.S. in Chemistry and a B.A. in Classics.