Title

High-Throughput Technologies for Genome Interrogation and Editing

Author

Rishi, Harneet Singh

Publication Date

2019

High-Throughput Technologies for Genome Interrogation and Editing

by

Harneet Singh Rishi

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Biophysics
and the Designated Emphasis
in
Computational and Genomic Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Adam P. Arkin, Chair
Professor Jennifer A. Doudna
Professor Nicholas T. Ingolia
Professor Kathleen R. Ryan

Summer 2019

High-Throughput Technologies for Genome Interrogation and Editing

Abstract

High-Throughput Technologies for Genome Interrogation and Editing

by
Harneet Singh Rishi
Doctor of Philosophy in Biophysics
and the Designated Emphasis
in Computational and Genomic Biology
University of California, Berkeley
Professor Adam P. Arkin, Chair

Microbial organisms are key drivers in processes related to human health, industrial biotechnology, and environmental ecology. In our attempts to predict, control, and design biological outcomes in each of these application areas, we require both an understanding of how microbes encode and perform their innate functions and the tools to implement engineering decisions. Here we present new technologies for interrogating genome function and editing genomes to carry out user-defined functions. First, we describe the development of CRISPRi as a platform for high-throughput genome annotation, focusing on a proof-of-concept in *Escherichia coli* to demonstrate our ability to map the genotype-phenotype-function landscape and uncover new design considerations for improving CRISPRi-based genetic screens. Next, we present the creation of a platform strain of *Escherichia coli* that achieves high rates of multiplex genome editing while retaining a low background mutagenesis rate, a favorable tradeoff that many editing strains have struggled to achieve. Finally, we report the design of novel CRISPR architectures that act not only as promising scaffolds for effector fusions but can also be programmed to sense and respond to endogenous cellular signals. Together, these technologies improve our ability to rapidly understand and engineer microbial behaviors.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Above all, I would like to thank my family for their constant, unconditional love and support: to my parents, who have always pushed me to become a better version of myself and supported me through every phase of life, and to my little brother, who has constantly been a source of joy in my life.

Throughout my academic career, I have been extremely fortunate in having mentors who have strongly contributed to my growth as a scientist and a communicator. In all areas of life, including graduate school, it is important to know the importance of what one is doing – both for oneself and for society – and then be able to effectively communicate and defend that position to others in a logical manner. As a mentor, Adam has imparted this valuable lesson upon me and, for that, I am extremely grateful. I am also grateful to my former mentors – Esteban Toro, Will DeLoache, and John Dueber – for their teachings and advice over the years. I have truly been blessed to have mentors that are not only such great teachers but whom I also consider to be close friends.

I would also like to thank everyone in the Arkin lab for making a research laboratory, which many would normally call a workplace, feel like a home by fostering an environment of friendship, learning through feedback and collaboration, and passion for science. I am grateful to have been surrounded by an exceptionally talented cohort of graduate students – Ben Adler, Kelly Wetmore, Mika Tei, Sean Carim, Judy Savitskaya, Aaron Berliner, and Tuesday Simmons. I have also been fortunate to learn from "seasoned" scientists in the lab – Vivek Mutalik, Rob Egbert, Ophelia Venturelli, Morgan Price, Lauren Lui, Jeff Skerker, and Adam Deutschbauer. Finally, I would like to thank Gwyneth Terry and Kate Chase for providing administrative support that has continually streamlined the scientific process.

I am also grateful for my friendships with a number of other amazing scientists during my time at Berkeley – from the Dueber lab (Zach Russ, Luke Latimer, Parry Grewal, Jen Samson, Tammy Hsu, Ryan Protzko, Mike Lee) and other graduate programs (Gabriel Lopez , John Canty, Chris Eiben, Ben Oakes, Julia Lazzari-Dean, Arjun Ravikumar, Pradeep Ramesh).

I would also like to thank my former K-12 teachers – specifically, Mrs. Caryl Brewbaker, Mr. J.B. Luzar, Mr. Charles Bissell, and Mr. Nate Loskutoff – for fostering my curiosity in science and engineering. I still fondly remember getting “hooked” on science from making salt crystals with pipe cleaners in Mrs. Brewbaker’s fourth grade science club. Furthermore, Mr. Luzar, Mr. Bissell, and Mr. Loskutoff always made learning science and mathematics enjoyable in their own way.

I would also like to thank my thesis committee members – Adam Arkin, Jennifer Doudna, Nicholas Ingolia, Kathleen Ryan – for their support and advice over the years.

Finally, I would be remiss not to thank my coauthors, who enabled me to do amazing science during my time at Berkeley.

- High-throughput CRISPRi as a platform for bacterial functional genomics: Esteban Toro, Honglei Liu, Xiaowo Wang, Lei S. Qi, Adam P. Arkin
- A versatile platform strain for high-fidelity multiplex genome editing: Robert G. Egbert, Benjamin A. Adler, Dylan M. McCormick, Esteban Toro, Ryan T. Gill, Adam P. Arkin
- CRISPR-Cas9 circular permutants as programmable scaffolds for genome modification: Benjamin L. Oakes, Christof Fellman, Kian L. Taylor, Shawn M. Ren, Dana C. Nadler, Rayka Yokoo, Adam P. Arkin, Jennifer A. Doudna, David F. Savage

Chapter 1. Introduction

1.1. Motivation

Microbial organisms are key drivers in processes related to human health, industrial biotechnology, and environmental ecology[1]. As we move forward in an era of engineering biology for effecting outcomes in each of these application areas, we appreciate that understanding how microbes encode and perform their innate functionalities is central to being able to predict, control, and design microbial behaviors.

Moving towards this aim requires an intertwined approach of using mechanistic insights from systems biology studies to make informed engineering implementations using synthetic biology tools (**Figure 1-1**). At the same time, we can use synthetic biology tools, which can themselves be sourced from systems-level studies, to make better technologies for interrogating microbial function (**Figure 1-1**).

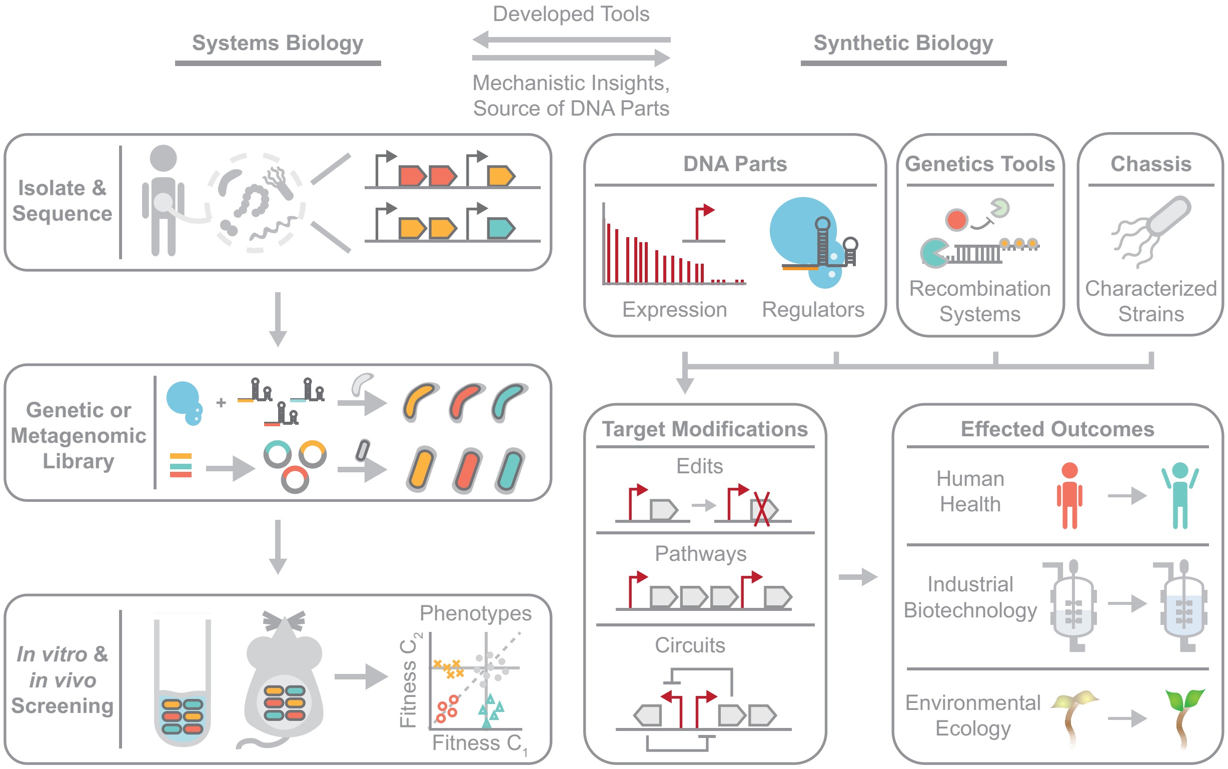


Figure 1-1. Example of feedback between systems and synthetic biology workflows for understanding and engineering microbial behaviors.

For example, a systems biology workflow could entail the isolation and sequencing of a microbe, or microbes, of interest to gain a first-level understanding of genetic potential. We can then create genetic or metagenomic libraries to enable the testing of genetic hypotheses. Finally, we can conduct high-throughput screens to gather process-level functions of genes and pathways (*in vitro* screening*)* along with their importance in an ecologically-relevant setting (*in vivo* screening), thus creating a map between genotype, phenotype, function, and traits. On the synthetic biology end, we can take characterized DNA parts, regulators to control those parts, genetic tools to put the parts together, and a chassis strain into which to put compiled DNA constructs to implement a variety of modifications ranging from simple edits to entire pathways and circuits. These engineered strains can then be deployed to effect outcomes in desired application spaces. At the same time, we can use synthetic biology tools to make better technologies for interrogating microbes of interest.

The use of each half of this interface has already been realized with high-throughput screens being used to characterize microbial genome function in ecologically-relevant contexts[2-4], and engineered microbial strains have been developed to effect outcomes *in vivo*[5,6]. As an example of combining the two halves, we can consider interventions for human health. High-throughput genetic screens can be used to identify fitness determinants of antibiotic-resistant pathogenic bacteria (e.g. genes required for the establishment of a metabolic niche, adhesion to intestinal receptors, extracellular signaling pathways). Given an understanding of how these processes work, we can then use an engineered microbe to interfere with such essential functions (e.g. compete for a metabolic niche by engineering competitive resource utilization, disrupt adhesion by secreting biomolecules to competitively inhibit intestinal binding receptors, interference with signaling molecules by converting or degrading the molecules). Moving forward, this workflow holds promise for numerous engineering applications.

1.2. Organization

In the following chapters (each representing a manuscript), we discuss (1) the development of a systems biology platform for genome interrogation in bacteria using CRISPRi and (2) synthetic biology tools for genome editing in the form of (i) a highly-efficient platform strain for implementing engineered functionalities and (ii) novel architectures of CRISPR Cas9 that function as promising scaffolds for DNA effector fusions and that can be programmed as highly versatile regulators. A graphical summary is provided in **Figure 1-2** and is followed by an abstract for each of the remaining chapters.

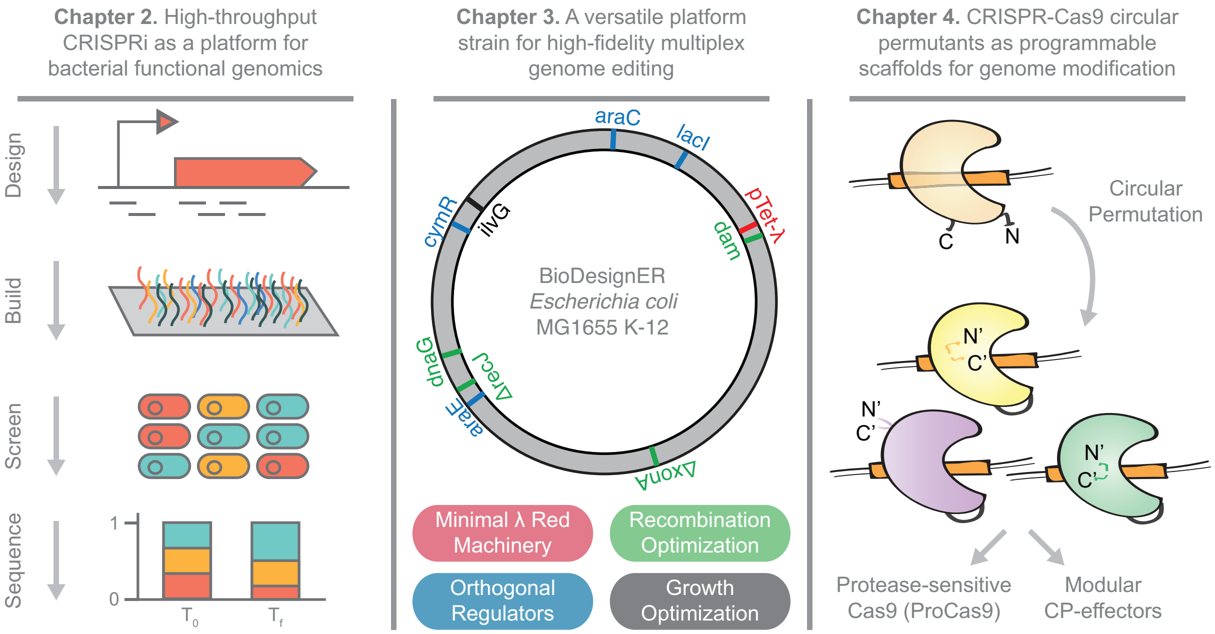

Chapter 2. High-throughput CRISPRi as a platform for bacterial functional genomics
Design
Build
Screen
Sequence
1
0
T0
Tf
Chapter 3. A versatile platform strain for high-fidelity multiplex genome editing
araC
lacI
pTet-λ
dam
ilvG
cymR
dnaG
araE
ΔrecJ
ΔxonA
BioDesignER
Escherichia coli
MG1655 K-12
Minimal λ Red Machinery
Recombination Optimization
Orthogonal Regulators
Growth Optimization
Chapter 4. CRISPR-Cas9 circular permutants as programmable scaffolds for genome modification
C
N
Circular Permutation
N'
C'
Protease-sensitive Cas9 (ProCas9)
Modular CP-effectors

Figure 1-2. Graphical summary of thesis.

Chapter 2: CRISPRi screens have enabled the high-throughput identification and characterization of essential genes in a number of bacteria. The programmability of CRISPRi targeting also enables the precise interrogation of smaller non-coding genomic features such as non-coding RNAs (ncRNAs), promoters, and transcription factor binding sites (TFBSs), yet these feature types have been underexplored in bacterial CRISPRi screens despite their critical roles in determining cellular physiology. Here we use a genome-wide CRISPRi library in *Escherichia coli* MG1655 K-12 targeting ~13,000 genomic features (protein-coding genes, ncRNAs, promoters, and TFBSs) to extend the functionality of bacterial CRISPRi screens. We first demonstrate that our CRISPRi library enriches for biologically significant features by showing that we can successfully knockdown 90% of known *E. coli* essential genes in a pooled screen. We next queried feature essentiality across several biochemical conditions and showed that certain genes regarded as essential by previous high-throughput efforts were only conditionally essential. Through this survey, we also found conditional phenotypes for small RNAs and detailed polar operon effects of CRISPRi. We also used time-series measurements to show that different essential genes exhibit distinct, transient responses when knocked down. We found this response to be correlated to gene function with genes involved in translation exhibiting among the strongest responses. Finally, we screened non-genic features to add phenotypic confidence to promoter annotations, show that gene-targeting more effectively perturbs gene expression than promoter-targeting, and find that targeting the non-template strand of the promoter closest to the target gene was more effective in knocking down gene expression than other promoter targeting orientations. Overall, this work (1) demonstrates the power of CRISPRi screens by revealing novel phenotypes for essential genes and small RNAs, (2) further characterizes the CRISPRi technology by elucidating transient differences in physiological response upon CRISPRi induction, comparing gene and promoter targeting CRISPRi, and highlighting new design rules for promoter CRISPRi, and (3) assesses the limitations of CRISPRi screening in associating phenotypes to specific genomic features through an analysis of polar operon effects and the targeting of regulatory elements in non-coding DNA.

Chapter 3: Precision genome editing accelerates the discovery of the genetic determinants of phenotype and the engineering of novel behaviors in organisms. Advances in DNA synthesis and recombineering have enabled high-throughput engineering of genetic circuits and biosynthetic pathways via directed mutagenesis of bacterial chromosomes. However, the highest recombination efficiencies have to date been reported in persistent mutator strains, which suffer from reduced genomic fidelity. The absence of inducible transcriptional regulators in these strains also prevents concurrent control of genome engineering tools and engineered functions. In this published work[7], we introduce a new recombineering platform strain, BioDesignER, which incorporates (i) a refactored λ -Red recombination system that reduces toxicity and accelerates multi-cycle recombination, (ii) genetic modifications that boost recombination efficiency, and (iii) four independent inducible regulators to control engineered functions. These modifications resulted in single-cycle recombineering efficiencies of up to 25% with a 7-fold increase in recombineering fidelity compared to the widely used recombineering strain EcNR2. To facilitate genome engineering in BioDesignER, we have curated eight context--neutral genomic loci, termed Safe Sites, for stable gene expression and consistent recombination efficiency. BioDesignER is a platform to develop and optimize engineered cellular functions and can serve as a model to implement comparable recombination and regulatory systems in other bacteria.

Chapter 4: The ability to engineer natural proteins is pivotal to a future, pragmatic biology. CRISPR proteins have revolutionized genome modification, yet the CRISPR-Cas9 scaffold is not ideal for fusions or activation by cellular triggers. In this published work[8], we show that a topological rearrangement of Cas9 using circular permutation provides an advanced platform for RNA-guided genome modification and protection. Through systematic interrogation, we find that protein termini can be positioned adjacent to bound DNA, offering a straightforward mechanism for strategically fusing functional domains. Additionally, circular permutation enabled protease-sensing Cas9s (ProCas9s), a unique class of single-molecule effectors possessing programmable inputs and outputs. ProCas9s can sense a wide range of proteases, and we demonstrate that ProCas9 can orchestrate a cellular response to pathogen-associated protease activity. Together, these results provide a toolkit of safer and more efficient genome-modifying enzymes and molecular recorders for the advancement of precision genome engineering in research, agriculture, and biomedicine.

Chapter 2. High-Throughput CRISPRi as a Platform for Bacterial Functional Genomics

2.1. Author Contributions

This chapter represents a manuscript with contributions from Harneet S. Rishi (H.S.R.), Esteban Toro (E.T.), Honglei Liu (H.L.), Xiaowo Wang (X.W.), Lei S. Qi (L.S.Q.), and Adam P. Arkin (A.P.A.). Given the collaborative nature of this work, it is important to acknowledge the contributions of all authors: H.S.R. led the experimental work and computational analyses. H.S.R, E.T., and A.P.A designed experiments. E.T. cloned the CRISPRi library and performed initial experiments. H.L. and X.W. designed the sgRNA library. A.P.A. supervised the research. H.S.R and A.P.A. wrote the manuscript. L.S.Q. and A.P.A. conceived of the research.

2.2. Introduction

Genome sequencing, catalyzed by advances in next-generation sequencing (NGS), has become the standardized first step to discerning the functional potential of microbes; however, characterizing the function of the 1000s of genes and non-genic features for each new microbial genome remains challenging. Computational pipelines attempting to bridge this gap can be limited in the scope of their inference models[9-11], indicating the need for complementary experimental approaches. To this end, genome-wide genetic screens have been utilized to infer gene function by generating large libraries of genetically perturbed gene mutants and profiling a phenotypic response (e.g. growth) to a gene perturbation (e.g. deletion) across a variety of biochemical conditions[12-15]. Several approaches have been developed for making such genetic perturbations at genome-scale via targeted modifications using λ-Red recombination[16-20] or random insertions using transposon elements[21-24].

In addition, Cas9 has been developed as a powerful tool for programmable gene repression[25], and the ability to induce genetic perturbation at a user-defined time – a feature not available in conventional gene disruption or deletion techniques – has enabled the CRISPRi-mediated characterization of essential genes in a number of bacteria[26-31]. The programmability of CRISPRi targeting also enables the interrogation of smaller non-coding DNA (ncDNA) features such as non-coding RNA (ncRNA) genes, promoters, and transcription factor binding sites (TFBSs). ncDNA features, which represent ~12 percent of the *E. coli* genome, play important roles in the regulation of gene expression in a condition-dependent manner. For example, small RNAs (sRNAs) have been implicated in transient regulatory processes involving membrane biogenesis, metabolism, and the synthesis of key transcription factors[32] while ncDNA regulatory elements drive key physiological decisions such as complex metabolism[33], pathogenicity[34], and gene expression diversification[35]. However, ncDNA features have been difficult to perturb using traditional methods due to the random targeting of transposons and disruption of local genomic context by insertions, making their interrogation via CRISPRi highly valuable.

Despite this potential value-add, previous bacterial CRISPRi screening studies have been limited in their study of RNA genes beyond simple cases (e.g. tRNA, rRNA genes) and have rarely addressed non-coding genomic features such as promoters and TFBSs. In comparison, CRISPRi screens in eukaryotic systems have been routinely employed to find new regulatory sites in enhancer regions[36-38] and functionally profile lncRNAs[39-41], indicating the untapped potential of CRISPRi for the functional characterization of bacterial genomes. In addition, most CRISPRi screens measure phenotypes using end-point fitness measurements by calculating the change in strain abundance between the beginning and end of a screen, which ignores dynamic outcomes that may occur over the course of an experiment. However, the physiological response resulting from CRISPRi-mediated gene repression could vary between different genes, arising from differences in protein and mRNA decay rates, feedback regulation, interaction network structure, and the physiological relevance of the targeted gene itself.

Here we leverage the programmable nature of CRISPRi to target approximately 13,000 *E. coli* MG1655 K-12 genomic features (protein-coding genes, non-coding RNAs (ncRNAs), promoters, and TFBSs) using a compact, designed oligoarray library of 32,992 sgRNAs. We first validated our technology by showing that we could knock down 90% of essential genes (as annotated by the Profiling of *E. coli* Chromosome - PEC - database[42,43]) in a pooled screen with the entire library. Through this process, we showed that a designed, compact library with ~4 guides/gene is sufficient for probing gene essentiality, which represents a considerable reduction in comparison to a previous designed *E. coli* screening study using 15 guides/gene[28]. Given that gene essentiality is context dependent, we expected that querying essentiality under a variety of biochemical conditions would allow us to delineate between a core set of essential genes and an accessory set of conditionally-essential genes. We thus leveraged the inducible nature of CRISPRi to propagate strains targeting essential genomic features and assay the library in several conditions to find condition-dependent phenotypes for essential genes and also ncRNAs. Next, we sought to investigate how different genomic features might respond upon CRISPRi induction. We used time-series measurements to track the dynamic response of genes in our library to CRISPRi perturbation and showed that essential genes exhibited distinct profiles that were correlated with their physiological function – a phenomenon not reported from previous CRISPRi screens due to their use of only endpoint measurements of fitness.

Finally, we studied the physiological effects of perturbing DNA regulatory elements such as promoters and TFBSs as these features have been understudied in previous bacterial CRISPRi screens. We showed that targeting promoters of essential genes could knock down gene expression and used this phenotypic outcome to add annotation strength to RegulonDB promoters. We also showed that perturbing gene expression was more successful when inhibiting transcription elongation (gene targeting CRISPRi) as opposed to inhibiting transcription initiation (promoter targeting CRISPRi) in our library through a comparison of guides targeting the promoter and gene sequences of known essential genes. By analyzing differences in sgRNA design features and the genomic context of targeted promoters, we found targeting the non-template strand of the promoter closest to the target gene was more effective in knocking down gene expression than other promoter targeting orientations, indicating a new design considering for promoter CRISPRi. Finally, we looked at the effect of dCas9 targeting to TFBSs to see if TFBS-targeted CRISPRi could perturb gene expression. We analyzed TFBSs regulating promoters of essential genes; however, due to the proximity or overlap of targeted TFBSs with promoters we were largely unable to associate phenotypes to specific TFBS features in most cases – finding only one case of a condition-dependent phenotype for a TFBS cluster regulating expression of a conditionally-essential aerobic respiration gene. Together, this work represents an extension and characterization of bacterial CRISPRi screens as well as a framework for the design, construction, pooled screening, and analysis of CRISPRi libraries for the high-throughput functional annotation of bacterial genomic features.

2.3. Results

2.3.1. Design and construction of CRISPRi library

We designed a CRISPRi library consisting of 32,992 unique sgRNAs to target 4457 genes (including 130 small RNAs; sRNAs) and gene-like elements (e.g. insertion elements / prophages), 7442 promoters and transcription start sites (TSSs), and 1060 transcription factor binding sites (TFBS) across the *E. coli* K-12 MG1655 genome using bioinformatic and biophysical design constraints (**Figure 2-1A**, see **Materials and Methods** for design details, **Extended Data-1A** for sequences). In brief, guides were designed to target proximal to a PAM site (NGG for *S. pyogenes* dCas9 used in this work), target a unique genomic sequence, maintain secondary structure of the sgRNA, and avoid extreme GC content. Gene-targeting guides were designed to target the non-template strand and target close to beginning of the gene. When possible, multiple guides were designed for each feature. Agilent Technologies synthesized the designed sgRNAs as an oligo pool (**Extended Data-1B**). To allow for the screening of smaller, more focused libraries the terminal 3′ end of each oligo was designed with a category code that allows for the amplification of subsets from the oligo library (**Figure 2-1B**, **Extended Data-1C**). To construct the genome-wide library, sgRNAs were PCR amplified from the oligo pool and then cloned into an expression vector using a golden gate assembly strategy (**Materials and Methods**). This expression vector (ColE1 origin) maintains the guides under arabinose-inducible control using a pBad promoter. The sgRNA library assembly was transformed into a strain harboring a genomically-encoded *dCas9* under aTc-inducible control using a pTet promoter.

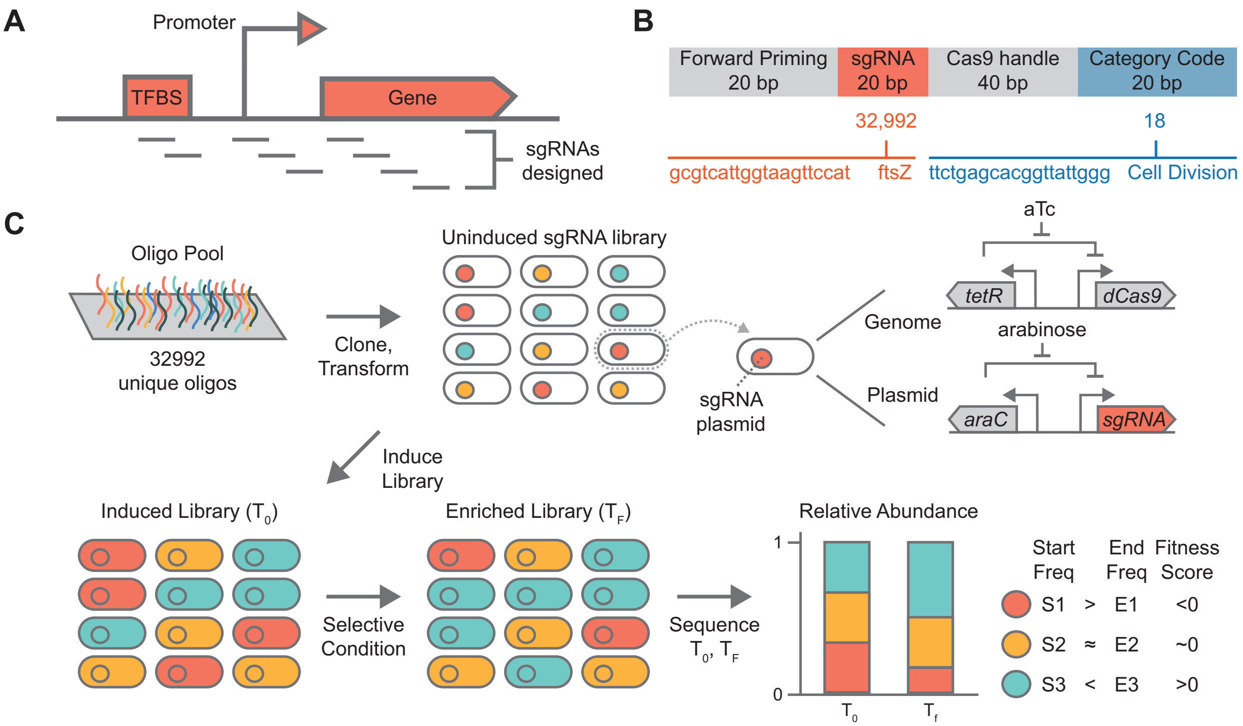


Figure 2-1. Overview of CRISPRi screening platform.

(**A**) Guide sequences were designed to target three feature types on the E. coli genome: (i) gene sequences (ii) promoters (iii) transcription factor binding sites (TFBSs). Multiple guides were designed for each feature where possible (Methods). (**B**) Guide sequences were synthesized as oligos and ordered via Agilent Technologies as a pool. Category codes (short DNA barcodes) were included in designed oligos to enable amplification of subpools from the library. (**C**) Guides were first cloned into a receiver vector and transformed into a strain containing chromosomally integrated dCas9. At the beginning of an experiment the library is induced and an initial time-point (T_0) is taken. After growth in a selective condition for a period of time a final time-point (T_F) is taken. The initial and final samples are sequenced and the fitness of each library member is calculated.

The identity of each knockdown strain in the library is determined solely by the sgRNA plasmid it harbors, specifically the 20 base pair variable region of the sgRNA that directs dCas9 targeting and encodes a DNA barcode for the strain. The relative abundance of every sgRNA, and by extension every strain, can be measured by amplicon sequencing of the variable sgRNA region from a plasmid DNA extraction of the sgRNA library. To perform a pooled functional screen the library is induced to express dCas9 and the sgRNAs and grown under selection for a short period of time (e.g. 24 population doublings) in a user-defined experimental condition (**Figure 2-1C**, **Materials and Methods**). During this competition, strains that carry an sgRNA targeting a feature important for growth will decrease in abundance in the pool. This phenotypic outcome can be quantified by measuring the starting and ending frequency of each strain and calculating a fitness score, which is defined as the normalized log2 ratio of the relative abundance of the guide-strain after the experiment to before the experiment (**Materials and Methods**). For gene targeting guides, we also define a composite gene fitness score as the median of fitness scores for all guides targeting a gene.

2.3.2. Technology validation of genome-wide CRISPRi gene knockdowns

To assess the ability of the library to yield biologically meaningful results, we profiled the phenotypic effect of knockdown for all genes in the library via a fitness experiment in LB Lennox rich media (LB). We found that CRISPRi was highly reproducible (Pearson $r_{\text{biological}} = 0.90$, $p < 0.05$, permutation test; Pearson $r_{\text{technical}} = 0.96$, $p < 0.05$, permutation test) (**Figure 2-2**). Furthermore, we observed that sgRNAs targeting known essential genes were severely depleted (i.e. strains harboring these guides exhibited a strong growth defect) over the course of an experiment when compared with sgRNAs targeting non-essential genes (**Figure 2-3A**). We compared the fitness results with the Profiling of E. coli Chromosome (PEC) database, which reports 304 *E. coli* K-12 MG1655 genes for which a knockout could not be generated, implying that these genes were essential for growth in LB rich medium under aerobic conditions (i.e. the condition of library construction)[42,43]. sgRNAs targeting 274 of 303 (~90%) essential genes were severely depleted (composite gene fitness ≤ -2) over the course of CRISPRi fitness experiments in the same condition, yielding 90 percent agreement with the PEC database. This also included proper depletion of all essential E. coli ncRNAs assayed in the experiment as well. Of the remaining 29 essential genes, 15 had at least one sgRNA with fitness ≤ -2 and an additional six had at least one sgRNA with fitness ≤ -1 (**Extended Data-2**). Overall, we found that 289 of 303 essential genes (~95%) could be knocked down by at least one designed sgRNA with fitness ≤ -2, indicating high activity of the CRISPRi library. We also tested the library in M9 minimal medium (M9) under aerobic conditions and found that 385 out of 415 (93%) minimal media essential genes had a gene fitness score ≤ -2 when knocked down (**Figure 2-4**).

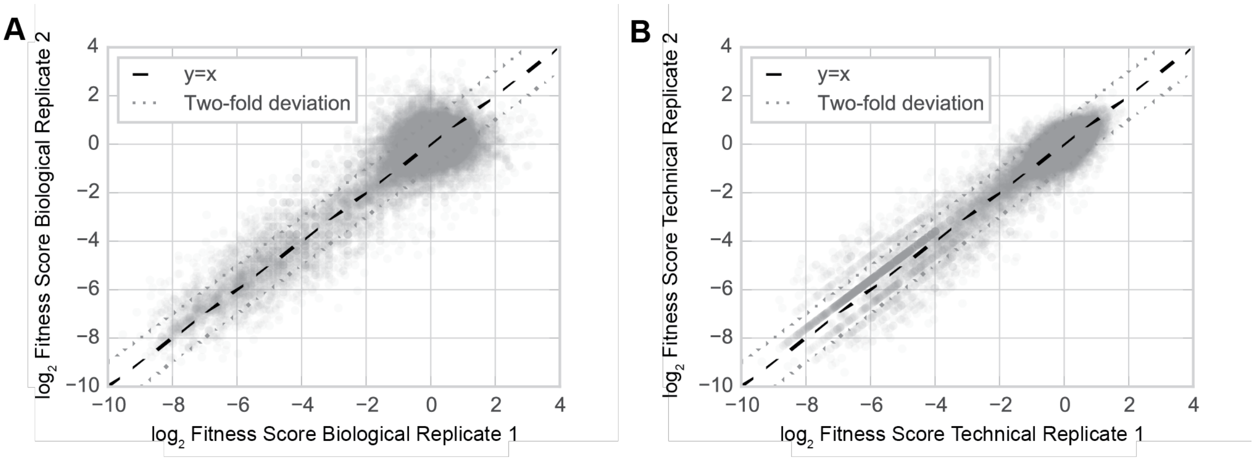


Figure 2-2. CRISPRi library replicability.

(**A**) Two biological replicates of a CRISPRi experiment where the library was grown in LB rich media. Each dot represents an sgRNA. A biological replicate represents a distinct library aliquot. (**B**) Two technical replicates of a CRISPRi experiment where the library was grown in LB rich media. Each dot represents an sgRNA. A technical replicate represents an aliquot of the library that was split prior to the start of the experiment.

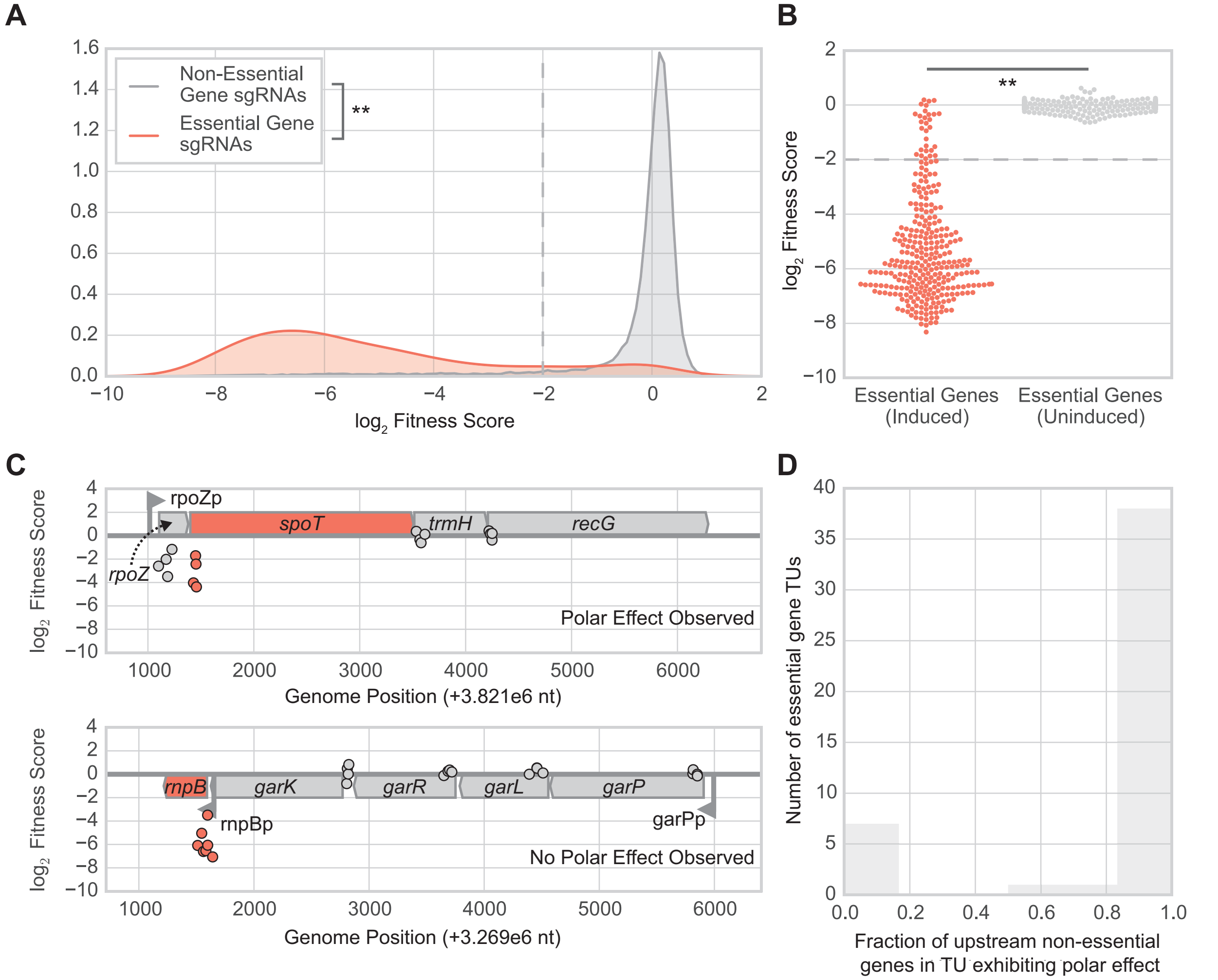


Figure 2-3. Technology validation of CRISPRi screening platform.

(**A**) Depletion of essential gene targeting sgRNAs compared to non-essential gene targeting sgRNAs over the course of a pooled fitness experiment with the CRISPRi library in LB rich media (with CRISPRi system induced) under aerobic growth conditions for 24 population doublings. $^{**}p < 0.001$ (Mann-Whitney U-test); Cohen's d = 3.7. (**B**) Demonstration of tight, inducible control of sgRNA library via comparison of essential gene fitness scores from pooled fitness experiments where the CRISPRi library was either induced (left) or uninduced (right). In the induced condition, the library was induced with aTc and arabinose to express dCas9 and sgRNA and then grown in LB media for 24 doublings as in a regular fitness experiment (**Materials and Methods**). In the uninduced condition, the library was also grown in similar culturing conditions (e.g. LB media for 24 doublings); however, neither aTc nor arabinose were added. $^{**}p < 0.001$ (Mann-Whitney U-test); Cohen's d = 3.7. (**C**) Example of CRISPRi-mediated polar operon effects where targeting a non-essential gene (*rpoZ*) upstream of an essential gene (*spoT*) in the same transcriptional unit (*rpoZ-spoT-trmH-recG*) produces a fitness defect (top panel). In the presence of an intra-operonic promoter (e.g. rnpBp), knockdown of upstream non-essential genes (*garK*, *garR*, *garL*, *garP*) in the same transcriptional unit (*garP-garL-garR-*

garK-rnpB) does not produce a fitness defect because essential gene expression can be rescued by the intra-operonic promoter (bottom panel). Targeting the intra-operonic promoter (rnpBp) or essential gene (*rnpB*) itself does produce a fitness defect. Each dot represents an sgRNA (centered at midpoint of chromosomal target) targeting either an essential (red-orange) or non-essential (gray) gene. (**D**) Fraction of non-essential genes upstream of an essential gene within the same transcriptional unit (TU) that also show a fitness defect when knocked down, likely indicating a CRISPRi-mediated polar operon effect.

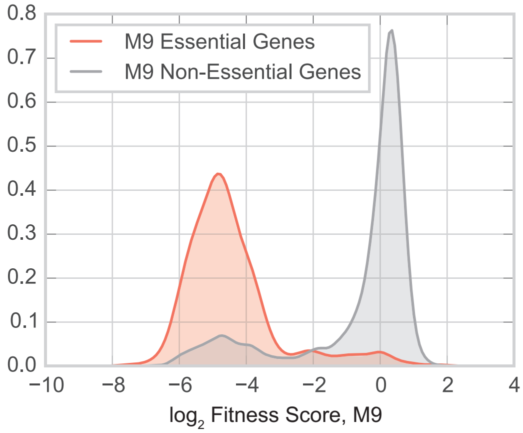


Figure 2-4. CRISPRi library minimal media experiment.

Depletion of minimal media (M9) essential gene targeting sgRNAs compared to non-essential gene targeting sgRNAs over the course of a pooled fitness experiment with the HT-CRISPRi library in M9 minimal media (with CRISPRi system induced) under aerobic growth conditions for 24 population doublings.

We also measured the tightness of inducible control for the CRISPRi library by growing it with no inducer (i.e. no aTc or arabinose added to turn on expression of *dCas9* and sgRNA) for the same period of time as a regular fitness experiment (24 population doublings). Strains with essential gene-targeting sgRNAs exhibited a negligible growth defect in this uninduced condition (with gene fitness scores near 0), and the fitness defect of essential gene strains was significantly different between this uninduced case and an induced case ($p < 0.001$, Mann-Whitney U-test; Cohen's d effect size = 3.7) (**Figure 2-3B**). This suggested that library strains with sgRNAs targeting essential genomic features can be maintained when the library is propagated in an uninduced state. We also checked if fitness was biased by factors such as position of targeting relative to chromosomal origin, GC content of the sgRNA, or chromosomal strand of the targeted gene and found no significant correlation (**Figure 2-5**). In agreement with prior reports of CRISPRi in bacteria[26,28-30], we found CRISPRi-mediated polar operon effects where knockdown of an upstream nonessential gene in an essential gene containing operon produced a growth defect similar to the essential gene itself, indicating that CRISPRi can knockdown entire operons (**Figure 2-3C**). Out of 160 operons containing at least one essential gene targeted in our library, we focused on 47 operons where the essential gene was not the first gene in the operon to assess the prevalence of polar operon effects. We found operon effects to be highly prevalent, with every non-essential gene (based on PEC database) upstream of the essential gene in 38 out of the 47 operons exhibiting a growth defect when targeted with dCas9 (**Figure 2-3D**).

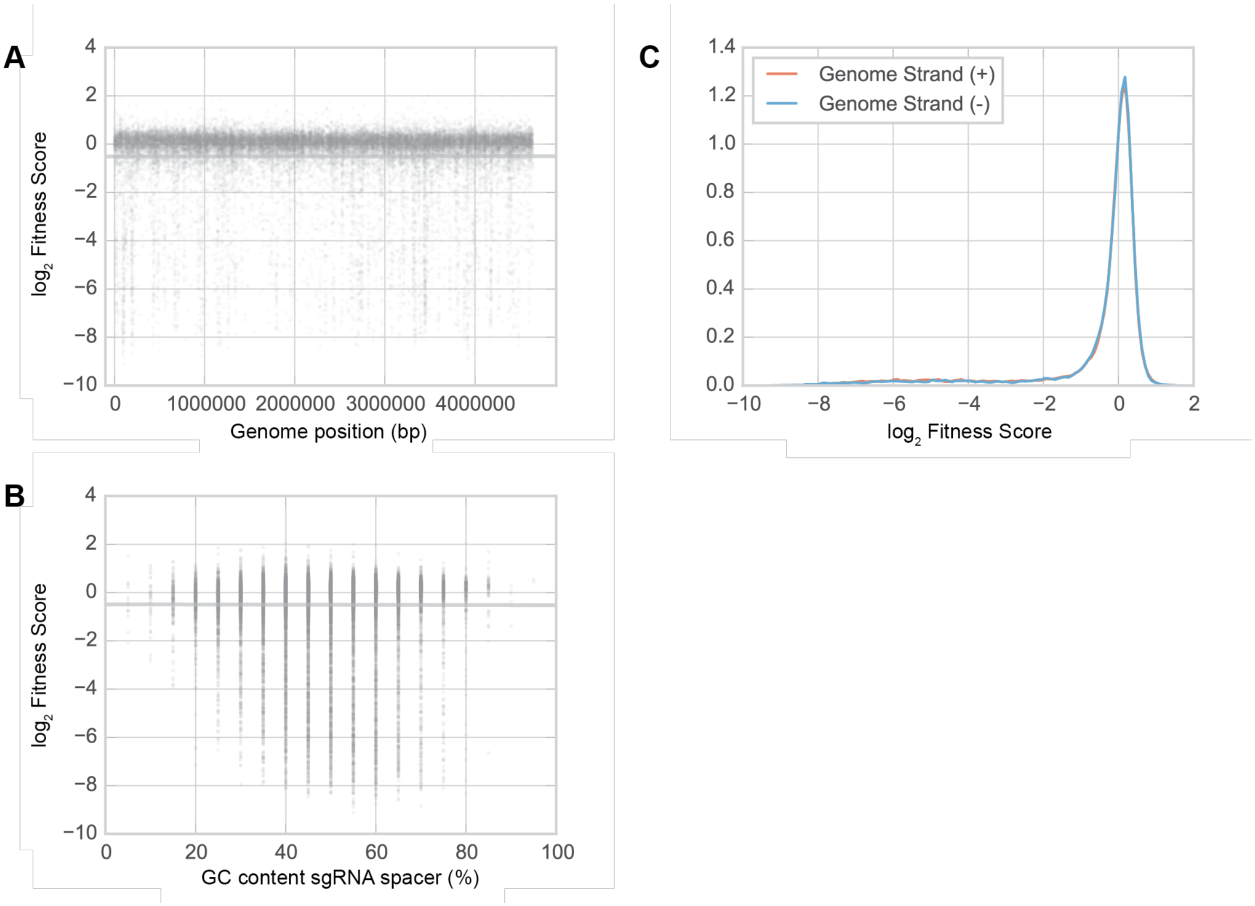


Figure 2-5. Investigation of bias in CRISPRi library.

(**A**) Genome position of library sgRNAs plotted against fitness of respective sgRNAs from a pooled experiment in LB media under aerobic conditions. Gray line represents linear relationship between fitness and genome position with 95% confidence interval. (**B**) GC content of sgRNA variable region for library sgRNAs plotted against fitness of respective sgRNAs from a pooled experiment in LB media under aerobic conditions. Gray line represents linear relationship between fitness and GC content of sgRNA spacer with 95% confidence interval. (**C**) Distribution of fitness scores for sgRNAs targeting features on the + or - strand of the genome.

2.3.3. Conditional screening of knockdown mutants enables discovery of specific phenotypes

To evaluate whether CRISPRi could assess feature fitness in a condition-specific manner, we compared feature enrichment in the library by varying two physiologically relevant parameters – nutrient availability and oxygen availability. In the case of nutrient availability, we profiled the CRISPRi library in M9 media, M9 media supplemented with casamino acids (M9Ca), and LB media under aerobic growth conditions. In the case of oxygen availability, we profiled the CRISPRi library in LB media under aerobic and anaerobic growth conditions.

We first compared enrichment between varied nutrient availability conditions (LB, M9Ca, M9). As previously discussed, we saw a strong depletion of sgRNAs targeting known essential genes (based on knockout studies) in LB and M9 media. We next analyzed non-essential genes that should exhibit condition-dependent phenotypes between these conditions by comparing the enrichment of known amino acid metabolism genes for expected auxotrophic phenotypes. We found a strong depletion of guides targeting genes involved in amino acid biosynthesis in the amino acid deficient medium (M9) but not the supplemented medium (M9Ca), indicating that CRISPRi can enrich for conditionally essential genes (**Figure 2-6**).

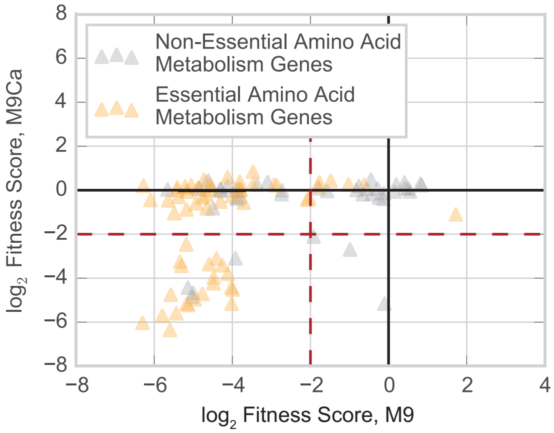


Figure 2-6. CRISPRi library amino acid auxotrophy experiment.

Depletion of amino acid biosynthetic gene targeting sgRNAs over the course of a pooled fitness experiment in either M9 minimal media (x-axis - M9) or M9 minimal media supplemented with casamino acids (y-axis - M9Ca) under aerobic growth conditions for 24 population doublings. Essential amino acid metabolism genes (yellow triangles) refer to genes classified as essential in Joyce et al *J Bacteriol* 2006 via screening of the Keio essential gene deletion collection on glycerol minimal medium.

Finally, we looked beyond phenotypes for protein-coding genes and analyzed sRNA feature enrichment. Out of the 130 sRNAs with designed guides in the library, we had fitness data for 114 in each condition (some sRNAs did not have data due to low read depth in one or more conditions). Of these 114 sRNAs, we found novel phenotypes for the *hok*/*sok* Type I toxin-antitoxin (TA) system, which has been implicated in bacterial persistence through the stringent response[44,45]. Specifically, under stress or amino acid starvation, (p)pGpp and Obg induce (via an unknown mechanism) expression of the *hokB* toxin gene, which leads to membrane depolarization and persistence[46]. In our CRISPRi screens, a knockdown of the *sokB* antitoxin sRNA gene resulted in a successively stronger growth defect in LB, M9Ca, and M9 media (**Figure 2-7**), likely due to its inability to inactivate the *hokB* toxin gene product under conditions where it is expressed. The related *hokC*-*sokC* system exhibited a similar, yet even stronger, response to the knockdown of antitoxin *sokC*. Previous literature has suggested that *hokC* is likely inactive due to an insertion element located 22 bp downstream of the *hokC* reading frame[47]. However, the *sokC* antitoxin sRNA exhibits a strong deleterious phenotype when knocked down, implying that *hokC* may still be functional. We hypothesize that this phenotype was not seen earlier because the *hokC*-*sokC* system had only been investigated in nutrient-rich conditions (e.g. LB); however, here we are able to uncover this phenotype by combining the programmability of CRISPRi targeting to investigate this small 55 bp feature with the ability to assess feature fitness across conditions.

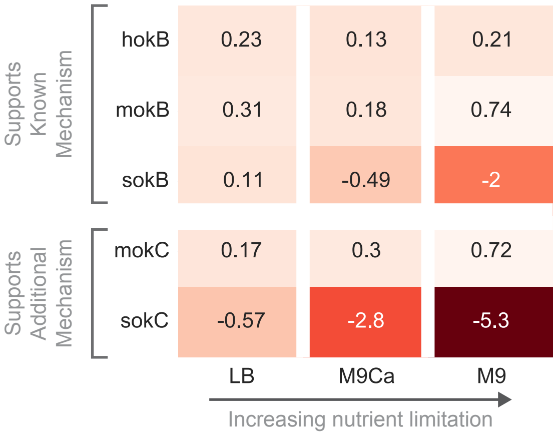


Figure 2-7. Conditional phenotypes for *hok-sok* toxin-antitoxin system.

Gene fitness scores for genes in the *hok-sok* toxin-antitoxin systems (B & C) showing increasing defect as a result of *sokB* and *sokC* knockdown under conditions of increasing nutrient limitation with *sokC* depicting a stronger phenotypic response than *sokB*. Mechanism for *hokB-sokB* is reported in Verstraeten et al *Molecular Cell* 2015. Nutrient conditions: LB (rich media), M9Ca (M9 minimal media supplemented with casamino acids), M9 (M9 minimal media). Gene fitness scores are averaged from a minimum of three replicates. Data from pooled fitness experiment with library grown for 24 population doublings under induction in stated condition.

We next compared enrichment between the aerobically varied conditions, expecting to find condition-specific phenotypes for genes involved in aerobic or anaerobic growth processes. Many strains with guides targeting genes involved in aerobic respiration (e.g. pyruvate conversion genes, heme biosynthetic genes, ubiquinol biosynthetic genes, cytochrome *bd*-I terminal oxidase subunits, ATP synthase F_1 synthase complex subunits) were depleted in the aerobic condition but dispensable under anaerobic growth (**Figure 2-8A**). NADH:quinone oxidoreductase I (*nuoABCEFGHIJKLMN*; NDH-1) and NADH:quinone oxidoreductase II (*ndh*; NDH-2) showed a previously unreported phenotype (**Figure 2-9**). NDH-1 only exhibited a defect in aerobic minimal media conditions (M9Ca, M9) while NDH-2 only exhibited a defect in the aerobic rich media condition (LB), implying that NDH-1 may be the dominant oxidoreductase in nutrient limited conditions and NDH-2 may be dominant in nutrient rich conditions. We noted that seven genes (*hemB, hemC, hemD, hemH, ispB, nrdA, nrdB*) previously characterized as essential according to the Keio database of essential genes in *E. coli* K-12 BW25113[16] and the PEC database of essential genes in *E. coli* K-12 MG1655 were dispensable for growth under anaerobic conditions (**Figure 2-8A**). These genes are involved in heme biosynthesis (*hemB, hemC, hemD, hemH*) and ubiquinol biosynthesis (*ispB*), which play critical roles in the aerobic electron transport chain. The essential genes *nrdA* and *nrdB*, which are involved in aerobic nucleotide metabolism[48,49], were also dispensable under anaerobic growth. We clonally verified the conditional essentiality of *nrdA* and *hemB* by showing that we could generate viable strains with deletions of these genes under anaerobic conditions and that these deletion strains were not viable under aerobic conditions (**Figure 2-8B-C**, **Table 2-1**). By demonstrating that these “essential” genes are only conditionally essential, we show that they are not part of the core, essential genome but instead part of the growth-supporting, conditionally-essential genome. We also noted that of the genes with conditional phenotypes in **Figure 2-8A**, 20 were genes (genes with double asterisks in **Figure 2-8A**) for which a gene disruption mutant was not generated during a high-throughput transposon insertion screen using Rb-TnSeq due to the attempted construction of the mutants under a condition where the underlying genes were essential. We clonally verified one of these genes, *ubiD*, by showing that we could generate a viable deletion strain under the condition determined as permissive via the CRISPRi screen (**Figure 2-8B-C**, **Table 2-1**). This analysis presents a proof of concept for the use of two intertwined capabilities of CRISPRi screening – the ability to induce CRISPRi to interrogate features traditionally regarded as essential and the ability to probe feature essentiality across conditions – to delineate between the core, essential and accessory, conditionally-essential genome.

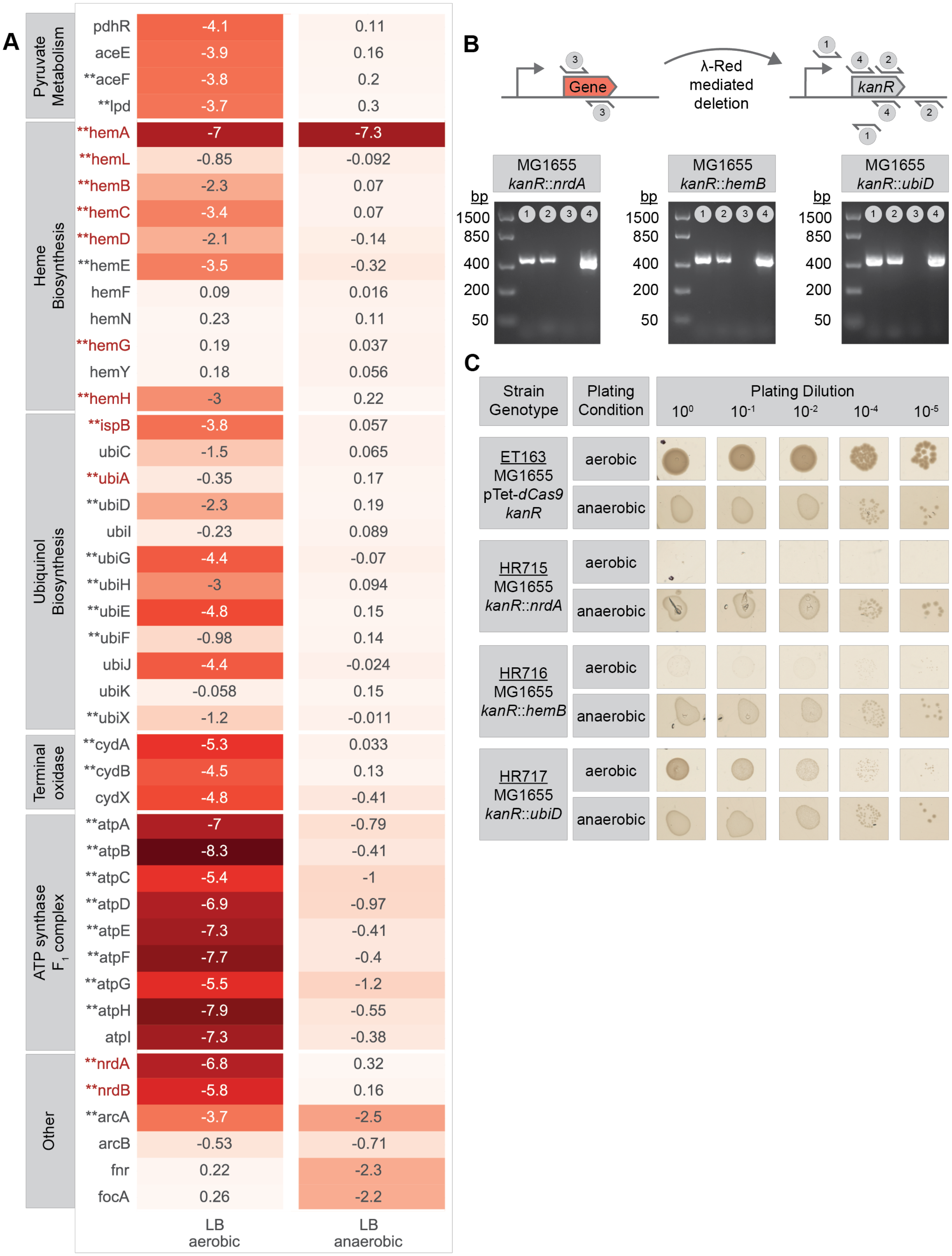


Figure 2-8. Conditional phenotypes from CRISPRi screening.

(**A**) Comparison of CRISPRi phenotypes (gene fitness scores) between aerobic and anaerobic conditions in LB. Gene names in maroon represent genes classified as essential by the Keio collection (*E. coli* K-12 BW25113) and PEC database of essential genes in *E. coli* K-12 MG1655. Gene names with a preceding “**” superscript represent genes for which a mutant could not be generated using RbTnSeq during a high-throughput screen in E. coli K-12 BW25113. Gene fitness scores are averaged from a minimum of three replicates. (**B**) λ-Red recombineering mediated deletion of select aerobic essential genes from Keio collection/PEC database (*nrdA*, *hemB*) or sick genes from Rb-TnSeq (*ubiD*) under permissive condition (anaerobic) as discovered via the CRISPRi screen. Gel images with reactions validating in-frame deletion of each essential gene via PCRs showing successful integration of kanR resistance cassette and removal of essential gene at native gene locus. (**C**) Confirmation that anaerobically generated knockouts of selected genes are non-viable under aerobic condition (non-permissive condition). An MG1655 strain with kanR cassette integrated on the chromosome is provided as a WT-like reference (ET163).

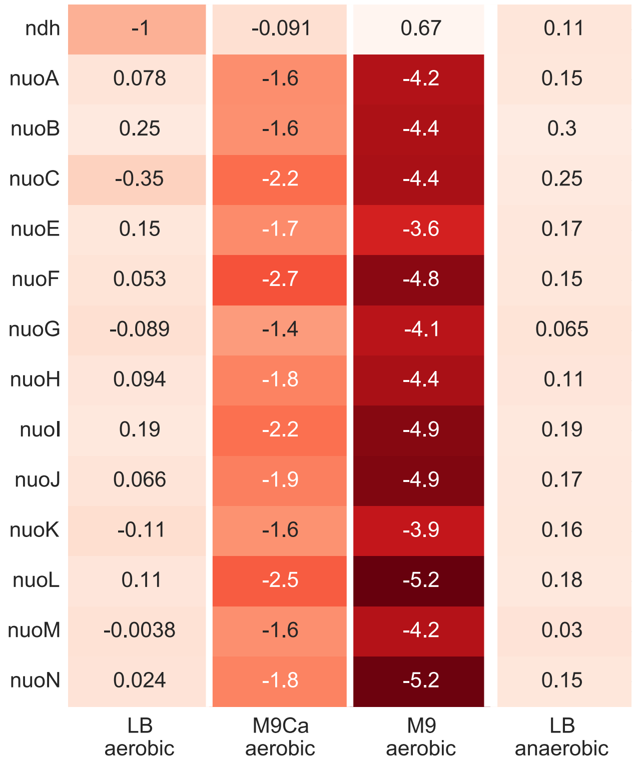


Figure 2-9. Conditional phenotypes for NADH:quinone oxidoreductases.

Comparison of CRISPRi phenotypes (gene fitness scores) between aerobic conditions in LB, M9Ca, and M9 media against anaerobic condition in LB for NADH:quinone oxidoreductase I (NDH-1; *nuo* genes) and NADH:quinone oxidoreductase 2 (NDH-II; *ndh*). Gene fitness scores are averaged from a minimum of three replicates. Data from pooled fitness experiment with library grown for 24 population doublings under induction in stated condition.

Table 2-1. List of essential gene knockout validation strains.

Strain Number	Plasmid Name	Description	Host	Resistance	Link to Modified Sequence
HR715	n/a	*kanR::nrdA*	MG1655 K-12	Kan	https://benchling.com/s/seq-LQ6uGkCQr09iU68FCnzh
HR716	n/a	*kanR::hemB*	MG1655 K-12	Kan	https://benchling.com/s/seq-Wu1pfiZX4M5JlaIFtora
HR717	n/a	*kanR::ubiD*	MG1655 K-12	Kan	https://benchling.com/s/seq-XOLhFwSV5CbGRdp5i1Yt

2.3.4. Time-series measurements elucidate dynamic knockdown response of essential genes

We next leveraged the ability to induce CRISPRi perturbations on-demand to probe the dynamic response to knockdown for the library, focusing on essential genes. Specifically, we grew the induced library and sequenced samples at regular intervals over a period of 18 population doublings in LB rich media (**Figure 2-10**). We examined the fitness of strains harboring guides targeting essential genes across the timepoints and found that these strains exhibited successively stronger growth defects over progressive time points (**Figure 2-11A**). We next clustered the essential gene time-series data (**Materials and Methods**) and found that essential genes could be classified into one of three groups (Early, Mid, Late) based on their temporal growth trajectory (see **Figure 2-11B** for examples and **Figure 2-11C** for groupings). For example, some essential genes showed a fitness defect soon after the first few population doublings while other genes did not show a defect until several population doublings had occurred. Of the 287 essential genes analyzed, 78 were in the Early group, 114 in the Mid group, and 95 in the Late group (**Extended Data-3A**).

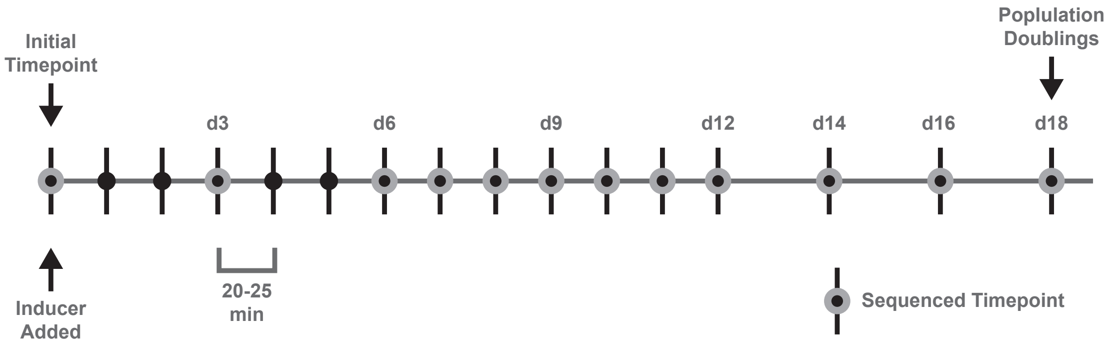


Figure 2-10. Workflow of CRISPRi time-series experiment.

The library was induced and an initial timepoint was taken. Samples of the library were taken every population doubling for the first 12 doublings and then every other doubling until population doubling 18. Timepoints with gray circles were sequenced.

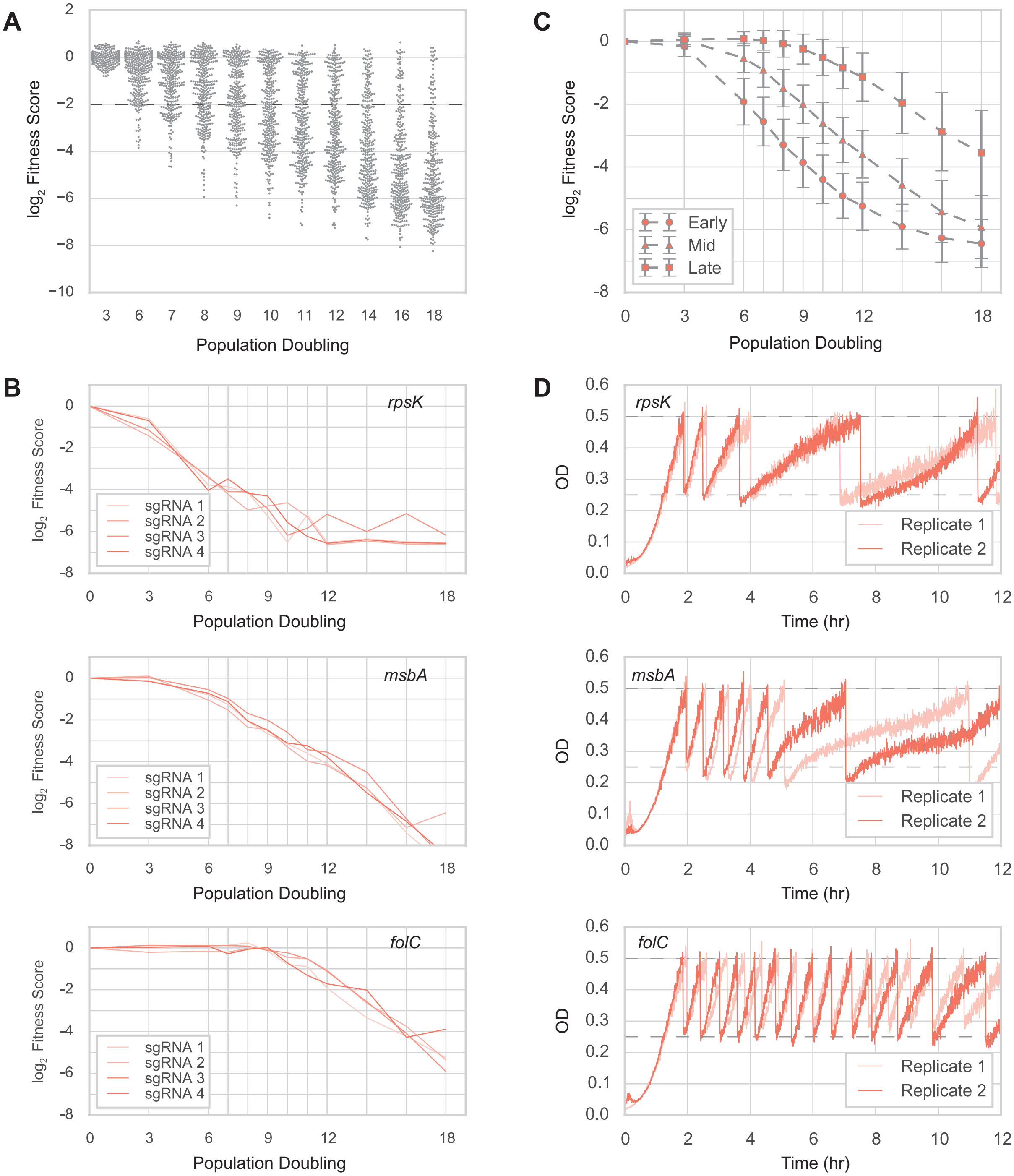


Figure 2-11. Temporal knockdown profiling of CRISPRi library.

(**A**) Gene fitness scores for PEC essential genes (n=304) from pooled CRISPRi experiment calculated at progressive timepoints (e.g. population doubling 3, 6, 7...) relative to initial timepoint (T_0). (**B**) Example temporal trajectories constructed from pooled CRISPRi experiment depicting one of three characteristic profiles observed for essential genes from K-means clustering. Each line represents an sgRNA for annotated gene (Early - *rpsK*, Mid - *msbA*, Late - *folC*). (**C**) Grouping of essential genes into classes (Early, Mid, Late) from K-means clustering and depiction of resulting composite growth curves. Each curve corresponds to an essential gene class with each solid marker (circle, triangle, square) denoting the mean fitness score of genes (averaged across two replicates) with that essential gene class at a given population doubling ($n_{Early} = 78$, $n_{Mid} = 114$, $n_{Late} = 95$; error bars represent ±1 standard deviation). (**D**) Growth curves of CRISPRi strains for candidate genes from each essential gene class as measured on eVOLVER, an automated turbidostat. For each selected essential gene, an sgRNA targeting that gene was selected from the CRISPRi library and cloned into a strain expressing dCas9. An uninduced culture of each strain was inoculated into the eVOLVER and grown until OD 0.50 in LB + antibiotics (carb/kan) media without inducers. Upon reaching this setpoint, each strain was diluted to OD 0.25 with LB + antibiotics (carb/kan) + inducers (aTc, arabinose) media and then allowed to grow between OD 0.25 and 0.50 with fresh inducer media being used for subsequent dilutions. Two replicates were grown for each CRISPRi gene strain.

We performed a gene ontology enrichment analysis to see if these classes were enriched for specific biological functions (**Extended Data-3B**, **Materials and Methods**). An analysis with TIGR Role ontologies[50] revealed that essential genes in the Early group were significantly enriched for genes involved in ribosomal protein synthesis and modification ($p < 0.001$, p-value from Hypergeometric test followed by FDR correction) with 32 out of 41 essential genes with this TIGR Role present in the Early group. Resource allocation studies in *E. coli* have shown that in rapidly dividing cells ribosomes are most abundant and important for growth[51] and haploinsufficiency studies in yeast have shown that ribosomal genes exhibit strong dose responses to gene expression perturbation in rich media[52]. This would support our finding of ribosomal protein synthesis and modification genes exhibiting a faster physiological response to expression knockdown (via growth defect) relative to other essential genes queried. An analysis of the Mid group revealed a strong enrichment in genes involved in tRNA aminoacylation ($p < 0.001$, Hypergeometric test with FDR correction) with 19 out of 22 essential genes with this TIGR Role present in the Mid group. The presence of tRNA aminoacylation genes in the Mid class also agrees with previous resource allocation studies, which report that the dosage effects observed under exponential growth are present, but less strong, for tRNA genes[53,54]. Finally, an analysis of the Late group revealed an enrichment of all eight essential genes involved in the 2-*C*-methyl-D-erythritol 4-phosphate/1-deoxy-D-xylulose 5-phosphate (MEP/DOXP) pathway ($p < 0.05$, Hypergeometric test with FDR correction). The MEP/DOXP pathway[55] represents the mevalonate-independent pathway for producing the isoprenoid precursors isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP), and its presence in a later, albeit still essential class, in comparison to translation-related genes indicates that the abundance of certain pathway metabolites may not be as rate-limiting to growth in rich media as genes related to translation.

We next analyzed all genes targeted in the library to see whether genes classified as non-essential also exhibited varied responses (**Materials and Methods**). We observed three categories after clustering, two of which contained genes exhibiting a growth defect via the knockdown of both essential and non-essential genes and a third category of genes that did not exhibit a growth defect (**Figure 2-12**, **Extended Data-3C**). Across the two categories of genes exhibiting a defect we saw an enrichment of a number of processes including translation, transcription, aerobic respiration, and fatty acid metabolism (**Extended Data-3D**).

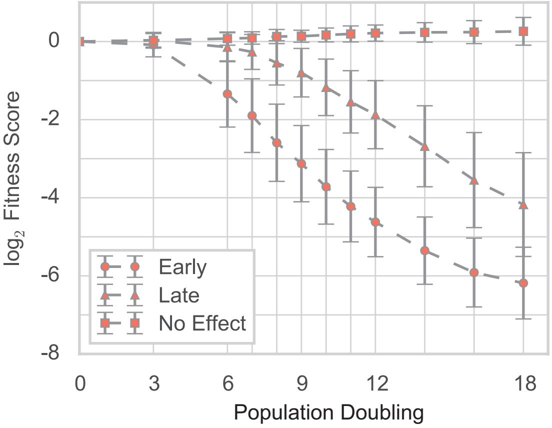


Figure 2-12. Time-series classification of all genes in CRISPRi library.

Grouping of all genes targeted in CRISPRi library into classes (Early, Late, No Effect) from K-means clustering and depiction of resulting composite growth curves. Each curve represents a gene class with each solid marker (circle, triangle, square) denoting the mean fitness score of genes (averaged across two replicates) with that gene class at a given population doubling (n_{Early} = 188, n_{Late} = 218, $n_{\text{No Effect}}$ = 4046; error bars represent ±1 standard deviation).

The composite nature of the analyzed growth curves meant that the apparent decline in abundance of a given strain could be the result of the slower growth of that strain, the faster growth of another strain, or a combination of the two cases. To distinguish between these cases and validate the trends among essential gene classes, we chose a representative essential gene from each class (Early, Mid, Late), generated individual strains with dCas9 and sgRNAs to separately target these essential genes (**Table 2-2**), and used the eVOLVER[56], an automated cell culture system, to monitor the temporal knockdown response. We also generated a strain expressing dCas9 along with an sgRNA that did not target any genomic locus to serve as a reference control. We used the eVOLVER as a turbidostat by programming it to keep cells between two optical density (OD) ranges, which allowed us to track changes in doubling time in response to CRISPRi induction. The control sgRNA strain exhibited no change in doubling time after CRISPRi induction (**Figure 2-13A**). In comparing the essential gene-targeting validation strains, we found that *rpsK* (Early gene) was the first to show an increase in doubling time upon induction of CRISPRi, followed by *msbA* (Mid gene) and *folC* (Late gene), thus confirming our observations from the pooled screen (**Figure 2-11D**). We also found that even within a gene class, different genes could have different profiles. For example, *msbA* showed a progressive increase in doubling time while *ftsZ* (another Mid gene) consistently showed a halt in cell growth after a set number of doublings (**Figure 2-13B**). Together, these results demonstrate that while CRISPRi knockdown of an essential gene eventually leads to a fitness defect, different genes can exhibit varied dynamic responses to perturbation, potentially indicating the functional importance of the genes and their biological roles as well as highlighting target considerations for CRISPRi applications where transient dynamics are important (e.g. CRISPRi-based genetic circuits).

Table 2-2. List of strains used for eVOLVER CRISPRi experiment.

Strain Number	Plasmid Name	Description	Host	Resistance	Link to Plasmid Sequence
ET169	pT169	Pbad:control sgRNA	ET163	Amp, Kan	https://benchling.com/s/seq-aLVjhEiBggyDQgeltKfg
ET170	pT170	Pbad:*ftsZ* sgRNA	ET163	Amp, Kan	https://benchling.com/s/seq-Zq5ApvVfBGbslqFt2RLW
HR664	pHR664	Pbad:*rpsK* sgRNA	ET163	Amp, Kan	https://benchling.com/s/seq-s88pSK0iEeEyRJ7xwr6G
HR665	pHR665	Pbad:*msbA* sgRNA	ET163	Amp, Kan	https://benchling.com/s/seq-JZWG9whqXzBqjFkrt4iA
HR666	pHR666	Pbad:*folC* sgRNA	ET163	Amp, Kan	https://benchling.com/s/seq-SqywQzhOAWZCmbDTgYHT

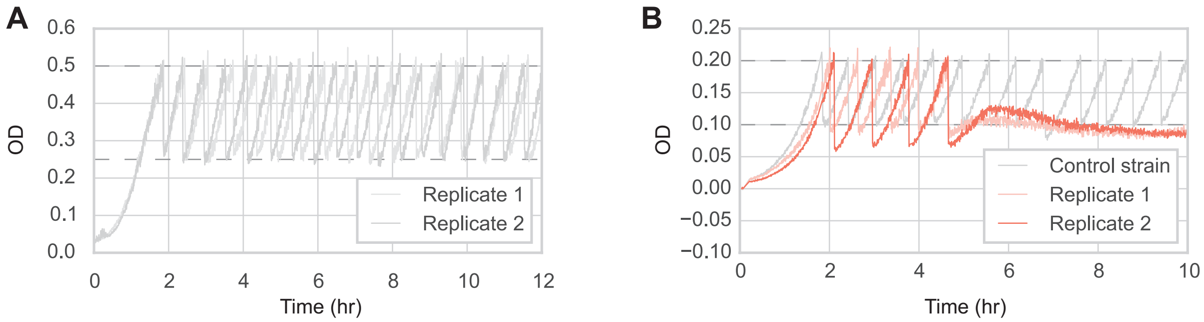


Figure 2-13. eVOLVER profiling of control and *ftsZ* CRISPRi strains.

(**A**) eVOLVER growth curves of two replicates of a CRISPRi strain expressing dCas9 and a control sgRNA that does not target any locus on the chromosome. An uninduced culture of the strain was inoculated into the eVOLVER and grown until OD 0.50 in LB + antibiotics (carb/kan) media without inducers, after which each strain was diluted down to OD 0.25 with LB + antibiotics (carb/kan) + inducers (aTc, arabinose) media and then allowed to grow between OD 0.25 and 0.50. (**B**) eVOLVER growth curves of replicate *ftsZ*-targeting CRISPRi strains. An sgRNA targeting *ftsZ* was selected from the CRISPRi library and cloned into a strain expressing dCas9. An sgRNA designed to not target any locus in the *E. coli* genome was also cloned into a strain expressing dCas9 and used as a reference control strain. An uninduced culture of each strain was separately inoculated into the eVOLVER and grown until OD 0.20 in LB + antibiotics (carb/kan) media without inducers, after which each strain was diluted down to OD 0.10 with LB + antibiotics (carb/kan) + inducers (aTc, arabinose) media and then allowed to grow between OD 0.10 and 0.20 for multiple generations until ~10 hours.

2.3.5. HT-CRISPRi uncovers design considerations for non-genic targeting

Promoter interference

The CRISPRi library contains 14,188 sgRNAs targeting 3,237 promoters and 4205 transcription start sites (TSSs) from RegulonDB (**Extended Data-4A**). To measure the efficacy of CRISPRi targeting for promoters on a genome scale we assessed whether knockdowns of promoters regulating essential genes produced a growth defect (**Figure 2-14A**). An analysis of 1,102 sgRNAs targeting 337 essential gene promoters across experiments in rich and minimal media (**Extended Data-4B**) revealed that (i) for 74% of essential gene promoters at least 1 sgRNA produced a mild knockdown phenotype (e.g. Fitness ≤ -1), and (b) for 51% of essential gene promoters, all sgRNAs produced a mild knockdown phenotype. Through this survey, we collected additional experimental phenotypes (i.e. collection of fitness scores) for 141 known promoter annotations from RegulonDB, which primarily uses RNA-seq as the primary source of experimental characterization for promoters (**Extended Data-5**). We also found, to the best of our knowledge, the first phenotype-based experimental evidence for four computationally predicted promoters of essential genes (**Extended Data-5**), highlighting the utility of CRISPRi to improve the annotation strength of non-genic genomic features. We compared the fitness effect of targeting essential gene sequences to that of targeting promoter sequences of essential genes and found that targeting promoters to knockdown gene expression was less efficient that targeting the gene sequence itself (**Figure 2-14B**). However, we did find cases where promoter-targeting produced a knockdown phenotype similar to gene-targeting knockdowns and where promoter-targeting yielded better knockdown performance than the gene knockdown (**Figure 2-15**), indicating the potential of promoter CRISPRi as an alternative to gene CRISPRi for control over gene expression. We also revisited the time-series data to analyze how promoter CRISPRi compared to gene CRISPRi following a perturbation. To avoid the confounding effects of multiple genes within the same transcriptional unit (TU) and multiple promoters driving the same TU, we focused on 27 monocistronic essential gene TUs regulated by a single promoter (**Materials and Methods**). We found a strong overlap between the trajectories of the two knockdown implementations (**Figure 2-16**), which further indicated the potential of promoter CRISPRi in the presence of well-designed sgRNAs. To elucidate factors contributing to better promoter guide design, we analyzed cases where promoter CRISPRi failed. We noted that 91% of essential promoters targeted by the 334 guides that did not produce a growth defect (Fitness > -1) either were part of a promoter array (i.e. two or more promoters in tandem regulating the same TU) or displayed a strong strand-dependency with respect to knockdown efficiency.

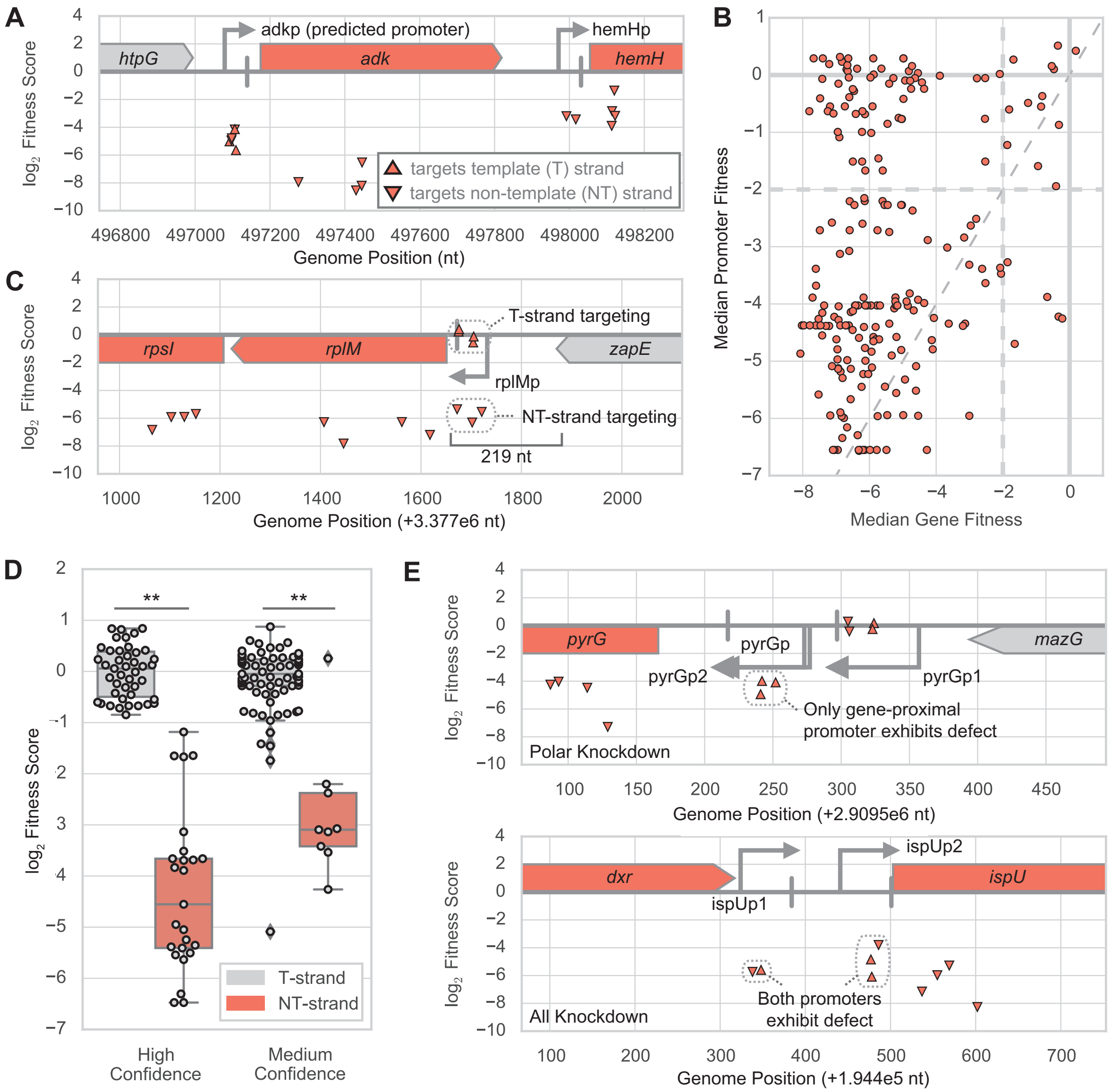


Figure 2-14. Non-genic phenotypes from CRISPRi library.

(**A**) Demonstration of how CRISPRi fitness data for promoter knockdowns can add experimental confidence to predicted promoters (e.g. adkp) and known promoters (e.g. hemHp) by confirming that targeting the promoter produces a similar phenotype (i.e. fitness outcome) in comparison to targeting its regulated gene (e.g. adkp - *adk*; hemHp - *hemH*). (**B**) Comparison of efficacy of gene-targeting CRISPRi against promoter-targeting CRISPRi for gene expression knockdown. For all essential genes for which guides targeting both gene and promoter sequences were present in the library, the median of fitness scores for sgRNAs targeting the gene sequence (x-axis) is plotted against the median of fitness scores for sgRNAs targeting the promoter sequence (y-axis). Note that the thin diagonal dashed line represents y = x. (**C**) Depiction of strand-dependency of CRISPRi-mediated promoter knockdown for rplMp driving expression of the *rplM-rpsI* operon. Only sgRNAs targeting the NT-strand of the promoter (relative to the gene) produce a fitness defect, while T-strand targeting sgRNAs do not. (**D**) Boxplots (with data points overlaid) showing strand dependent promoter CRISPRi for 12 high-confidence cases and 26 medium-confidence cases. Each case represents a TU and all of the promoters regulating it (**Materials and Methods**). **p < 0.001 (Mann-Whitney U-test); Cohen's d = 4.3 (left), 3.2 (right). (**E**) Phenotypic profiles of tandem promoter arrays where only knockdown of an essential-gene proximal promoter yields a CRISPRi-mediated growth defect (top) or where a knockdown of any promoter regulating the essential gene can yield a growth defect (bottom).

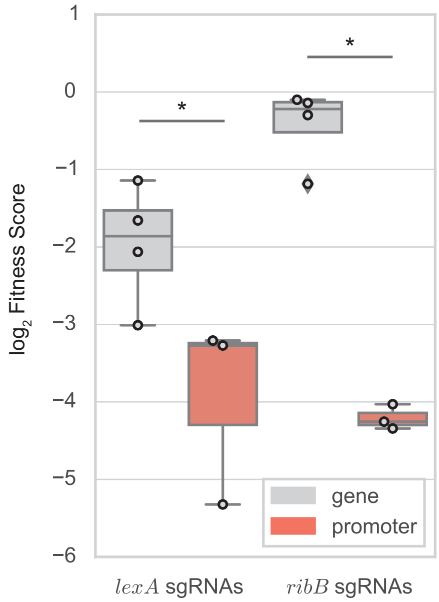


Figure 2-15. Examples of promoter-targeting guides more effective than gene-targeting guides.

Example case where promoter-targeting sgRNAs provide better knockdown of a known essential gene than functional gene-targeting sgRNAs (left - *lexA*) and gene-targeting sgRNAs that were unable to produce a fitness defect (right - *ribB*). *p < 0.05 (Mann-Whitney U-test); Cohen's d = 2.4 (left), 10.7 (right).

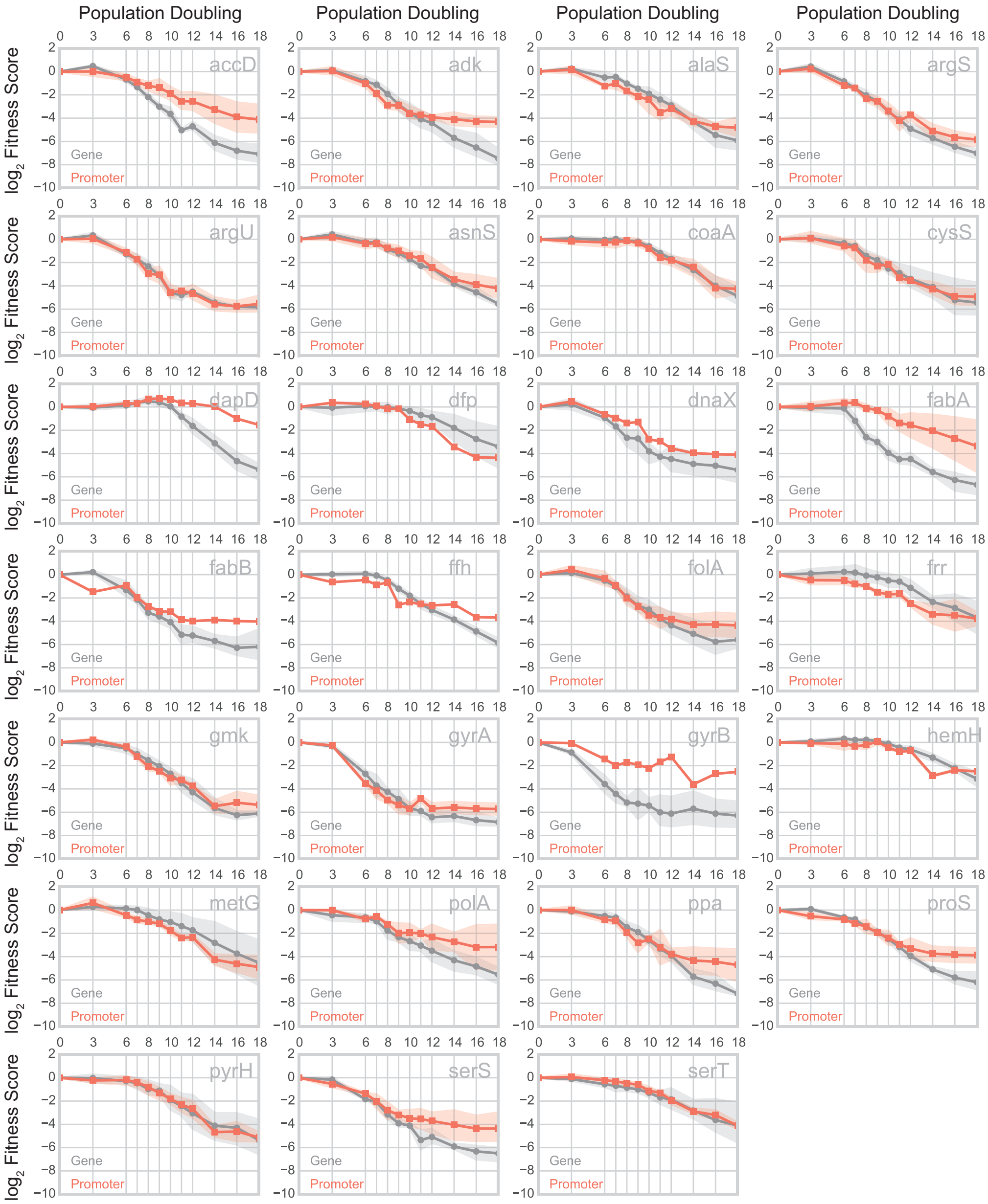


Figure 2-16. Comparison of promoter- and gene-targeting CRISPRi time series.

Composite fitness curves of promoter- and gene-targeting sgRNAs with Fitness ≤ -1 for monocistronic essential gene transcriptional units regulated by a single promoter (see **Materials and Methods** for details). Each curve represents the mean fitness of gene- (gray; circle marker) or promoter- (red-orange; square marker) targeting sgRNAs (averaged across two replicates) for each measured time point with corresponding shaded regions representing 95% confidence intervals.

We hypothesized that for cases where effective promoter knockdown was strongly dependent on the targeted strand, the sgRNAs could be targeting more effective positions within the promoter to interfere with transcriptional initiation or that the local genetic context was influencing knockdown efficacy. In the latter scenario, we hypothesized a model of "transcriptional coupling" where CRISPRi targeting of a promoter on the template strand failed to produce a fitness defect (while targeting the non-template strand could produce a defect) due to its inability to block RNAP readthrough from an upstream transcriptional event. We systematically identified 11 high-confidence cases where targeting the non-template strand produced a growth defect while targeting the template strand did not (**Figure 2-14C-D**, see **Extended Data-5** for cases and scoring metrics). One explanation for this result could be the transcriptional overlap of intra-operonic promoters in operons containing multiple TUs (e.g. one TU within a larger TU). Recent reports have also suggested that the transcription boundaries of operons are not as static as previously thought with one study using long read sequencing (SMRT-Cappable-seq) to demonstrate that 34% of RegulonDB operons can be extended by at least one gene and that 40% of transcription termination sites have read-through that alters operon content[57]. Indeed, of the 11 high-confidence cases, five were TUs contained within larger operons and the remaining six TUs were a part of an extended RegulonDB operon in the SMRT-Cappable seq study (**Extended Data-6**). We also found an additional 26 cases of medium-confidence (**Extended Data-5**) that are candidates for this transcriptional coupling that we could not fully confirm either due to an insufficient number of guides available in both targeting orientations to test our strand hypothesis or due to cases where most, but not all, guides produced phenotypes matching the strand hypothesis (**Figure 2-14D**, **Extended Data-5**). Of these 26 cases, 15 were TUs that were part of larger operons and seven were part of extended RegulonDB operons (**Extended Data-6**). Overall, our results suggest that targeting the non-template promoter strand can lead to a higher likelihood of successful CRISPRi knockdown for promoters in certain operonic contexts.

We also found that targeting CRISPRi in promoter arrays can yield distinct phenotypic profiles. Out of 59 tandem promoter arrays analyzed in the essential gene promoter data set, we found 40 tandem promoter arrays where we observed one of two distinct phenotypic profiles: (1) all promoters in the array produced a knockdown phenotype or (2) only the downstream promoter produced a fitness defect (**Figure 2-14E**, **Extended Data-5**). In the case where all promoters produced a deleterious knockdown phenotype, we hypothesized that either the most upstream promoter was the primary driver of expression or that all promoters in the array were required for appropriate expression. In the case where only the downstream most promoter showed a deleterious knockdown phenotype, we hypothesized that either the downstream most promoter was the primary expression driver or that all promoters in the array are required for appropriate expression. The remaining 19 tandem promoter arrays analyzed either had an insufficient number of guides to draw any conclusions or were inconsistent with the aforementioned phenotypic profiles (**Extended Data-5**). Overall, our results showed that the promoter closest to the target gene is more likely to yield a knockdown phenotype and thus should be targeted when attempting to knockdown expression of a gene regulated by multiple promoters via promoter CRISPRi.

TFBS interference

Finally, we analyzed a set of 1810 sgRNAs in the library that were designed to target 1060 TFBSs on the chromosome (**Extended Data-7**). We first focused on a subset of 175 sgRNAs that targeted 102 TFBSs regulating an individual promoter controlling expression of at least one rich media (based on PEC database) or minimal media (based on Joyce et al *J Bacteriol* 2006) essential gene. We found that most TFBS knockdowns that yielded a deleterious knockdown phenotype were present within the RNAP footprint for promoter binding, which we conservatively defined as between -60 to +20 nt relative to the transcription start site (TSS) associated with the promoter (**Figure 2-17**). Due to this overlap, we were unable to specifically associate such phenotypic outcomes to the TFBS alone as they could also be (and likely were) a result of promoter knockdown. Ultimately, we found that it was challenging to parse the phenotypic contribution of TFBSs due to their presence in promoters or binding site arrays with multiple diverse transcription factors.

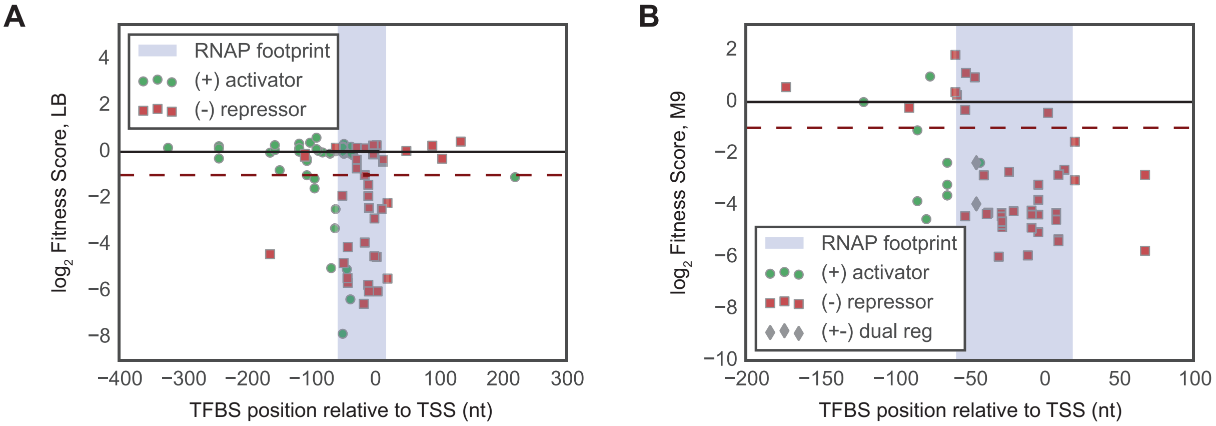


Figure 2-17. CRISPRi knockdown of TFBSs regulating single essential gene promoters.

(**A**) Fitness scores for sgRNAs targeting TFBSs regulating single promoters of transcription units containing at least one LB essential gene (as determined by PEC database). The RNAP footprint is defined as the window between -60 to +20 nt relative to the transcription start site (TSS) of the regulated promoter. Each object in the scatter plot represents the fitness of an sgRNA (y-axis) targeting a TFBS at a given distance from the TSS of the promoter it regulates (x-axis). A given TFBS can have a positive effect on gene expression (green circles), negative effect on gene expression (red squares), or dual effect on gene expression (gray diamonds) as determined by RegulonDB annotations. (**B**) Fitness scores for sgRNAs targeting TFBSs regulating single promoters of transcription units containing at least one M9 essential gene (as determined by Joyce et al *J Bacteriol* 2006).

We next looked at all TFBSs that exhibited a growth defect when targeted across all conditions in which the library was assayed. The activating NarL TFBS regulating the *cydDC* promoter, cydDp, exhibited a mild condition-dependent phenotype between aerobic and anaerobic conditions in LB (**Figure 2-18**). sgRNAs targeting *cydD*, which plays a role in respiration, and cydDp exhibited a growth defect in an aerobic fitness assay in LB medium but displayed no such defect under anaerobic conditions where no terminal electron acceptor was added and thus no respiration was active. Similarly, an sgRNA targeting the NarL TFBS, which has a positive effect on gene expression for *cydDC* and is situated -126 nt from the cydDp TSS, exhibited a mild growth defect as well (Fitness ~ -1.5) in the aerobic condition and a negligible growth defect (Fitness ~ 0) in the anaerobic condition.

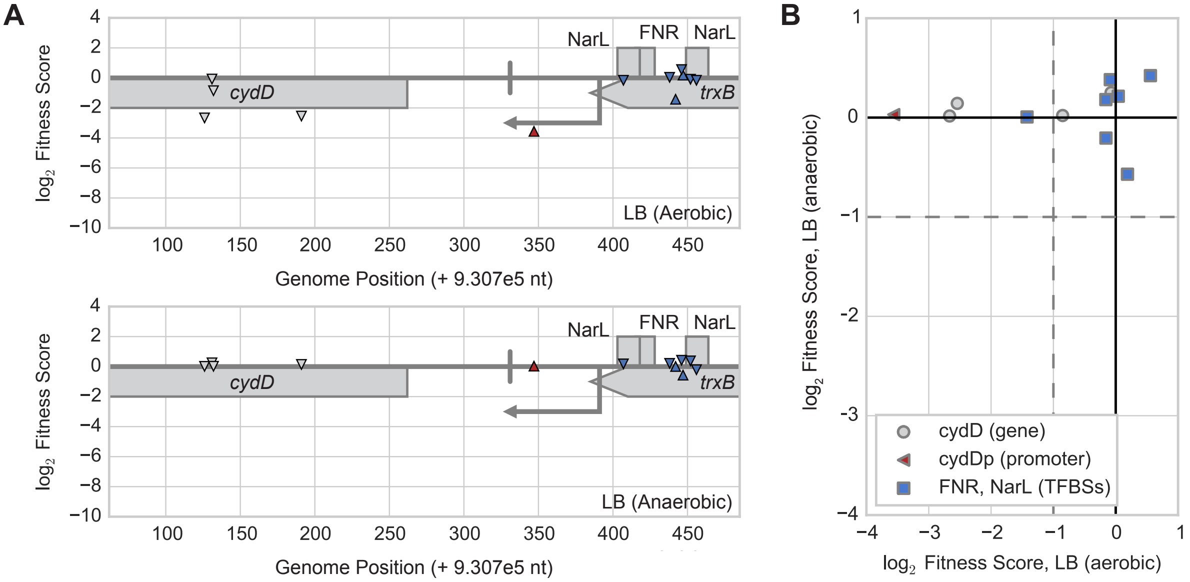


Figure 2-18. Feature cofitness of *cydD* gene, promoter, and TFBS-targeting sgRNAs.

(**A**) Fitness data for *cydD* gene, its corresponding promoter (cydDp), and TFBSs (NarL - gene expression activator, FNR - gene expression activator) regulating its promoter from fitness assays in LB media between aerobic (top panel) and anaerobic (bottom panel) conditions. Each triangle represents an sgRNA (centered at midpoint of chromosomal target) targeting either the chromosomal strand corresponding to the non-template (downward facing triangle) or template (upward facing triangle) strand of the *cydD* gene. (**B**) Scatter plot comparing conditional phenotypes for sgRNAs targeting *cydD* (gray circles), *cydD* promoter (red triangle), and *cydD* TFBSs (blue squares) between aerobic and anaerobic conditions.

2.4. Discussion

Here we used CRISPRi as a platform for the high-throughput phenotypic interrogation of the *E. coli* genome. During the preparation of this manuscript, two other studies reported the use of genome-wide CRISPRi libraries to identify essential genes and genes involved in phage-host interactions in *E. coli*[28,58]. Our work here presents a complementary and extended demonstration of the power of CRISPRi-based approaches to interrogate microbial genomes with the discovery of novel phenotypes for essential genes using a more compact library, application of time-series measurements to track and elucidate phenotypic changes arising after CRISPRi induction, presentation of refined rules for CRISPRi targeting of promoters, and investigation of CRISPRi targeting of TFBSs.

We leveraged the inducible nature of CRISPRi to propagate strains with sgRNAs targeting essential genomic features and query them in a number of biochemical contexts, a task unfeasible using conventional gene disruption or knockout approaches. This enabled us to generate 100s of essential gene strains not covered by conventional knockout or Tn-Seq approaches in *E. coli*. Furthermore, we showed that a number of genes classified as essential genes according to classical aerobically generated *E. coli* knockout collections or unable to be assayed using Tn-Seq approaches were actually dispensable under anaerobic conditions, representing a more comprehensive annotation of these genes. We validated the dispensability of three of these genes by showing that we could generate strains with deletions of these genes under the condition they were predicted to be dispensable from the CRISPRi screen. We also utilized the inducible nature of CRISPRi to track the effect of knockdown on essential genes post induction of the CRISPRi machinery. Using time series measurements, we found that different essential gene strains displayed growth defects at distinctly different times, and our results enabled us to classify essential genes into specific categories based on how quickly a given gene's knockdown yielded a measurable fitness defect. The genes in the most essential category had a remarkable overlap with genes discovered to be most essential in other systems biology studies of *E. coli* in the same condition and also matched gene dosage studies in yeast.

The programmable nature of CRISPRi targeting also allowed us to interrogate promoters and TFBSs. Specifically, we were able to compare gene-targeted CRISPRi (inhibit transcription elongation) to promoter-targeted CRISPRi (inhibit transcription initiation), finding that gene-targeting CRISPRi largely outperformed promoter-targeting CRISPRi. We also attributed phenotypic evidence to 141 known RegulonDB-annotated promoters and associated, to our knowledge, the first experimental evidence to four predicted promoters from RegulonDB. Finally, we explored phenotypic profiles associated with tandem promoter arrays and promoters that displayed strand-dependent knockdown success to conclude that targeting the NT-strand of the promoter closest to the target gene can yield more successful CRISPRi knockdowns in comparison to other promoter-mediated orientations for certain genomic contexts.

While we demonstrated a high utility for microbial genome interrogation via CRISPRi-based screens in this work, CRISPRi still has a number of limitations. First, targeting in operons yields polar effects, thus limiting the analysis of essentiality to transcriptional units and assigning specific phenotypic confidence to only the last gene in the transcriptional unit. As such, CRISPRi should serve as a complementary method to transposon insertion and recombineering-based approaches, which are less prone to polar operon effects. Second, the compact organization of bacterial genomes yields architectures with overlapping or tightly spaced TFBS and promoter features. This makes it especially challenging to precisely attribute phenotypes to a specific TFBS (due to its proximity or overlap with other TFBSs and promoters). Precise genome editing methods such as MAGE and CREATE are likely more suitable for such cases. Regardless, the programmability of CRISPRi targeting can be used to uncover intergenic regions of phenotypic importance through tiled screens, which can be combined with TFBS and promoter predictions along with high-throughput measurements (e.g. protein-DNA interactions, RNA-seq) to add annotation confidence for newly-sequenced microbes. Overall, the CRISPRi library developed here presents a resource of curated and phenotype-linked sgRNAs for use in *E. coli* and the workflow developed here for interrogating genic and non-genic chromosomal features provides the basis for high-throughput CRISPRi studies in other bacteria.

2.5. Materials and Methods

2.5.1. Chemicals, reagents, and media

LB Lennox Medium (EZMix™ powder microbial growth medium, Sigma Aldrich) was used to culture strains for experiments in rich media. M9 Minimal Medium (1X M9 salts, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 0.4% glycerol) was used to culture strains for experiments in minimal media. Anhydrotetracycline (aTc; CAS 13803-65-1, Sigma-Aldrich) was used at 200 ng/mL to induce dCas9 expression. Arabinose was used at 0.1% to induce sgRNAs. Antibiotic concentrations used were 100μg/mL for carbenicillin and 30μg/mL for kanamycin. Glucose was used at 0.2% in media for outgrowth of the library from a freezer aliquot. Casamino acids (0.2%) were also used in M9 Minimal Medium for select assays.

2.5.2. CRISPRi library design

We designed the sgRNA library following rules described in prior work [59]:

Selection of sgRNAs for oligo pool:

1. We first identified all 5′- XXXXX XXXXX XXXXX XXXXX NGG-3′ sequences by searching both the sense and anti-sense strand of the genome, to generate the original pool of the potential sgRNA binding sites.
2. To avoid potential off-target effects, we mapped all 5′-XX XXXXX XXXXX-NGG-3′ from step #1 back to the genome using the short reads mapping program Seqmap (http://www-personal.umich.edu/~jianghui/seqmap/) with parameter setting "1 /output_all_matches", and filtered out the sequences with multiple mappings.
3. We required that the designed sgRNAs should be able to fold properly. To check that this was true, we linked the 42nt scaffold sequence 5′-GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCG-3′ to the 3′ end of the 20 nt specific target binding sequence and checked the folding structure of this 62nt sequence by RNA secondary structure prediction using RNAfold (http://www.tbi.univie.ac.at/~ronny/RNA/) with default

parameters. We only kept the ones that the scaffold region could fold to the hair-pin structure as reported previously[60].

4. Finally, we filtered out any sgRNA sequences containing the BsaI restriction site (GGTCTC), which we used for cloning purposes.

The sequences that passed these four steps composed our pool of potential sgRNAs. Next, we chose sgRNAs from the sgRNA pool to target all (1) annotated genes, (2) promoters and (3) TFBSs, according to RegulonDB.

1. sgRNAs that target coding sequences:

   We tried to collect 4 sgRNAs for each annotated gene in the *E. coli* genome. We implemented a recursive approach to select sgRNAs as close to the ATG as possible and on the non-template strand for each gene. We first looked at the first 50%. Next, we looked at the annotated 5′ UTR regions, and the ones close to the start codon where selected with higher priority. Finally, we looked at the last half of the CDS sequence and chose the sites closer to the start codon with higher priority. By using this approach, 4281 genes could be targeted with 4 sgRNAs, 193 additional genes could be targeted by 1-4 sgRNAs, and 158 genes could not be targeted by any sgRNA.

   We further looked at the 158 genes that could not be targeted by the previous pipeline. We noticed that 39 of them were located in operons where an upstream gene in that operon had properly selected sgRNAs.

   For the rest 109 genes, we found many of them had closely related homologs on the genome, which caused the sgRNAs targeting these regions to be not unique on the genome and could target both of the homologs. So we compared the sequences of all the annotated genes, and defined a homolog gene set by performing a megablast search with parameter setting of "-F F -D 3 -e 1e-10". We searched for the potential sgRNA target sites that locate in both the homolog genes but not any other sites on the genome. 48 genes could be targeted in this way.

   Finally, there are 71 genes could not be targeted by our sgRNA design procedure. Most of them are small RNAs that don't have any PAM site.

Finally, we designed 17622 sgRNAs, which could target 4561 genes (4522 directly, 39 indirectly) on the *E. coli* genome.

2. sgRNA target promoters:

   For the promoters that did not overlap with any annotated UTR or CDS regions, we selected the sgRNA from both the sense and anti-sense strand in the region from upstream 60 bp to downstream 10 bp relative to the transcription start site.

   For the promoters located within a gene body, we only designed sgRNAs that binds to the template strand of that region. 14257 sgRNAs were selected to target 7404 Promoters.

3. sgRNA target TFBS sites:

   We designed all the sgRNAs that could target the TFBSs annotated in the RegulonDB database. An sgRNA is selected if it could cover at least one-third of the annotated TFBS. If the TFBS is shorter than 15 bp, we required that the overlap should be at least 5bp. 1867 sgRNAs were selected to target 1264 TFBS sites.

4. sgRNA for subcategories:
   a. We designed sgRNAs for 21 genes subcategories (e.g. cell division, small RNAs, central intermediary metabolism). These sgRNAs are encoded with an additional category code in the 3′ end of each library oligo to enable amplification of subpools of the library. Categories and their corresponding category codes for amplification can be found in **Extended Data-1C**.

We used the following external files as annotations for our sgRNA design:

- Genome sequence: *Escherichia coli* str. K-12 substr. MG1655, complete genome, NCBI Reference Sequence NC_000913.2 (http://www.ncbi.nlm.nih.gov/nuccore/NC_000913.2)
- Genome annotations from RegulonDB v8.1: (http://regulondb.ccg.unam.mx/download/Data_Sets.jsp)
  - Gene coordinate: Gene_sequence.txt
  - Promoter annotation: PromoterSet.txt
  - UTR annotation: UTR_5_3_sequence.txt
  - Transcription factor binding sites: BindingSiteSet.txt

Note: To keep with genome annotation updates, sgRNAs were remapped to promoter and TFBS features using more recent RegulonDB annotations:

- Promoter annotation: PromoterSet.txt (RegulonDB v9.4; release date 05-08-2017)
- TFBS annotation: BindingSiteSet.txt (RegulonDB v10.5; release date 09-13-2018)

See **Extended Data-1** for sgRNA feature annotations, sequence-level details, and a summary of category codes.

2.5.3. CRISPRi library construction

To clone the sgRNA library, sgRNAs were amplified from the OLS oligo pool using primers 282 (5′ CACATCCAGGTCTCTCCAT 3′) and 284 (5′ cacatccaggtctctCGGACTAGCCTTATTTTAACTTG 3′) using Phusion II HS and the following protocol: 98°C for 10 sec and 15 cycles of 98°C for 10 sec, 60°C for 30 sec, and 72°C for 20 sec followed by a final extension of 72°C for 5 min. The PCR reaction was purified using a Zymo DNA Clean & Concentrator kit and eluted in water. The purified library was cloned into the library receiver plasmid, pT154 (https://benchling.com/s/seq-YGEVpcmWzQjGfRrP8oDc), via a goldengate reaction using BsaI and T7 DNA ligase. The goldengate reaction product was purified using a Zymo DNA Clean & Concentrator kit, following the kit parameters for a plasmid cleanup. A derivative of *Escherichia coli* K-12 MG1655 (ET163: MG1655 FRT-*kanR*-FRT *tetR*-pTet-*dCas9*; https://benchling.com/s/seq-Gxu6IV96FF6y8jycpTrU) was used as the recipient strain for the sgRNA library. The purified library was electroporated into a competent cell preparation of ET163 and maintained under carbenicillin (plasmid marker) and kanamycin (strain marker) selection. Aliquots of the resulting library were stored at -80°C.

2.5.4. CRISPRi fitness experiments

An aliquot of the library was taken from storage at -80°C and thawed at room temperature. The aliquot was used to inoculate a 5 mL culture of LB Lennox media (LB) with carbenicillin, kanamycin, and glucose (multiple aliquots were used to inoculate distinct cultures for experiments with biological replicates). The culture was grown at 37°C until it reached OD600 0.5. A 4 mL aliquot was taken as an initial timepoint for the library (t0 sample); this sample was centrifuged (Eppendorf 5810R) at 4000 RPM (3202xg) and stored at -80°C. The remaining 1 mL of culture was centrifuged (Eppendorf 5417R) at 8000xg and washed twice with 1 mL of LB media. 156 uL of this washed sample was added to 10 mL of LB media (~1:64 dilution) with arabinose (0.1%), aTc (200 ng/mL), carbenicillin (100µg/mL), and kanamycin (30µg/mL). Technical replicates were generated by dividing this initial culture into 5 mL cultures. Cultures were grown at 37°C until they reached OD600 ~0.5, indicating 6 population doublings of the library. The library was again diluted 1:64 into 5 mL of LB media with arabinose, aTc, carbenicillin, and kanamycin and grown at 37°C until the culture reached OD600 ~0.5. This process was repeated until the library had undergone a total of 24 population doublings under induction. After 24 population doublings, all of the sample was centrifuged (Eppendorf 5810R) at 4000 RPM (3202xg) and stored at -80°C.

For experiments in minimal media, the original freezer aliquot of the library was inoculated in M9 media with glycerol (0.4%), glucose (0.2%), carbenicillin, and kanamycin. For induction of the CRISPRi system, the library was cultured in M9 media with glycerol, arabinose, aTc, carbenicillin, and kanamycin. Casamino acids (0.2%) were added depending on the assay condition.

For time-series experiments, samples were collected every doubling after the t0 sample was taken for the first 12 doublings, after which samples were collected every two doublings until the library had undergone a total of 18 doublings. During the experiment, the library was maintained between OD600 ~0.25 and ~0.50.

2.5.5. CRISPRi sequencing library preparation

Frozen, centrifuged samples from fitness experiments were taken from storage at -80°C and thawed at room temperature. The HT-CRISPRi sgRNA library was isolated using a QIAprep® Spin Miniprep Kit. 10-20 ng of DNA from each sample was used for a PCR reaction to generate NGS-ready sequencing samples in a 50 uL reaction using Phusion polymerase and two primers to add one of two sets of indexed Illumina adaptors. The first set contained a constant reverse primer and a variable forward primer with sample-specific 8 nucleotide barcodes that were sequenced "in-line" during an Illumina sequencing read. The second primer set contained a constant forward primer and a variable reverse primer with sample-specific indices that could be sequenced during an indexing read (**Extended Data-1D**). Both primer sets yielded comparable sequencing results; however, we eventually shifted to using the second primer set as the data could be readily demultiplexed using Illumina software.

Each reaction was performed using the following protocol: 98°C for 30 sec and 21 cycles of 98°C for 10 sec, 67°C for 15 sec, and 72°C for 10 sec followed by a final extension of 72°C for 5 min. 5 uL of each PCR sample was pooled and purified using a Zymo DNA Clean & Concentrator kit. The purified sample was quantified using the Qubit dsDNA HS assay kit and product size was confirmed using a Bioanalyzer 2100 automated electrophoresis system (DNA 1000 Kit). Final samples were run on either an Illumina Miseq or HiSeq instrument (2000/2500; Vincent J. Coates Genomics Sequencing Laboratory, UC Berkeley). All relevant sequencing data has been deposited in the National Institutes of Health (NIH) Sequencing Read Archive (SRA) at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA559958 under Accession code PRJNA559958.

2.5.6. CRISPRi sequencing data analysis

Sequencing runs were demultiplexed using standard Illumina software for samples using the second primer set or a custom python script (demultiplex_fastq.py) for samples using the first primer set. Demultiplexed reads were processed using the following set of custom python scripts: trim_sgRNA_reads.py to trim and filter reads according to quality thresholds; bwa_samtools.py to map the trimmed sgRNA reads to a BWA index of the sgRNA library; parse_bam.py to convert mapped reads to a table of counts that represent the abundance of each sgRNA in the sample. Custom scripts for analysis are available at https://github.com/rishih91/Thesis/Scripts.

2.5.7. CRISPRi fitness score calculation

A small constant (i.e. pseudocount, usually 1) was added to the raw read counts to avoid errors in calculating fold-change in subsequent fitness calculations due to division by 0. These adjusted read counts for each sample were normalized by the median abundance for that sample, thus generating relative abundance (RA) values for each sgRNA library member and enabling comparisons between different samples. The fitness score was calculated as the $\log_2$ ratio of the RA of a guide strain in a test condition relative to its RA in a control condition. In this framework, the test condition was a sample of the library after being subjected to grown over the course of an experiment, and the control condition was the t0 sample. The fitness scores from each sample were normalized such that the median fitness score for the sample was 0. In practice, library members with t0 raw read counts < 10 were filtered out to limit variability due to low read depth. Significance values for each sgRNA fitness score were calculated via the edgeR package using raw read counts as the input[61,62].

We also created a gene fitness score, which we calculated as the median of fitness values for all sgRNAs targeting a given gene. This provided a more stringent metric for quantifying strong fitness scores. For example, for a given gene with four sgRNAs, at least two guides would have to yield a strong fitness score in order for the median to be lower than -2. Fitness scores for all relevant experimental samples are listed in **Extended Data-8**.

2.5.8. Analysis of time-series data

The fitness of each sgRNA strain was calculated at each sequenced time point relative to the initial timepoint of the experiment. This constructed a time-series fitness curve for each sgRNA in the library.

Time-series Analysis 1 – Clustering of Essential Genes:

1. Calculate gene fitness scores for each gene annotated as essential in the PEC database
2. Filter out any genes that did not have a gene fitness score ≤ -1 (i.e. keep only essential genes that showed a knockdown phenotype)
3. Keep only timepoints with a Pearson correlation ≥ 0.8 across two replicates
4. Average the remaining timepoints across replicates
5. Performed a min-max scaling of each timepoint (i.e. i.e. fitness values at each timepoint were scaled to between 0 and 1) from Step 4 to ensure that all timepoints were treated equally
6. Used the Elbow method to track the variation of the within-cluster-sum-of-squares (WCSS) with the number of clusters (k – ranging from 1 to 14) and found k = 3 to be the optimal number of clusters for K-means based on visual inspection.
7. Performed K-means clustering with selected k from Step 6 to classify essential gene curves

8. Visualize K-means clusters (Early / Mid / Late)

Time-series Analysis 2 – Clustering of All Genes:

1. Calculate gene fitness scores for each gene targeted in the CRISPRi library
2. Keep only timepoints with a Pearson correlation ≥ 0.8 across two replicates
3. Average the remaining timepoints across replicates
4. Performed a min-max scaling of each timepoint (i.e. i.e. fitness values at each timepoint were scaled to between 0 and 1) from Step 3 to ensure that all timepoints were treated equally
5. Used the Elbow method to track the variation of the within-cluster-sum-of-squares (WCSS) with the number of clusters (k – ranging from 1 to 14) and found k = 3 to be the optimal number of clusters for K-means based on visual inspection.
6. Performed K-means clustering with selected k from Step 5 to classify essential gene curves
7. Visualize K-means clusters (Early / Late / No Effect)

Gene Ontology Enrichment for Analysis 1 and 2:
For either time-series analysis, each gene was associated with its annotated TIGR Role. A hypergeometric test was carried out for each TIGR Role in each gene class (for analysis 1 – Early / Mid / Late; for analysis 2 – Early / Late / No Effect) with parameters: N = #total essential genes in data set, K = #total genes in class, n = #total genes with TIGR Role in data set, k = #genes with TIGR Role in class. The Benjamini-Hochberg correction was applied to the resulting p-values using the multitest function (parameter: "fdr_bh") in the statsmodels python module (http://www.statsmodels.org/stable/index.html). The threshold of $p_{FDR\text{-}adjusted} \leq 0.05$ was used as the significance threshold.

Time-series Analysis 3 – Comparison of gene-targeting and promoter-targeting CRISPRi:

1. Select all essential genes for which guides targeting the corresponding promoter and the gene itself were designed in the library
2. Of these promoter-gene pairs, select all essential genes that are the first and only gene in their respective transcription unit (TU). This enables association of a specific promoter knockdown or gene phenotype to the specific gene itself.

3. Of the remaining promoter-gene pairs, select cases where the gene only has one promoter
4. Keep only sgRNAs that had t0 counts ≥ 10 and had a fitness score ≤ -1 by the final timepoint (i.e. timepoint 15)
5. Plot time-series using lineplot function from seaborn plotting library (v0.9.0) with the parameter setting "ci = 95" to generate 95% confidence intervals via bootstrapping.
    a. Lineplot function: https://seaborn.pydata.org/generated/seaborn.lineplot.html
6. For each gene, compare the overlap of the 95% confidence intervals between population doublings 6 and 12 (these timepoints were selected because they are both highly correlated across replicates and because after doubling 12 we start to see fitness scores leveling out due to limitations in sequencing read depth)

2.6. Extended Data

Due to the large-scale nature of genomic data, certain data tables cannot be practically included in a document like this. As such, these data tables have been made available online at https://github.com/rishih91/Thesis/ExtendedData. A list of data tables available online is listed below:

- **Extended Data 1.** CRISPRi library design details
- **Extended Data 2.** List of genes with median gene fitness score > -2
- **Extended Data 3.** Gene classification and ontological enrichment from time-series analyses
- **Extended Data 4.** Annotations for sgRNAs targeting promoters
- **Extended Data 5.** Results from analysis of essential gene promoters
- **Extended Data 6.** Comparison of transcription readthrough results with SMRT-Cappable Seq study
- **Extended Data 7.** Annotations for sgRNAs targeting TFBSs
- **Extended Data 8.** Fitness scores for relevant experimental samples

Chapter 3. A Versatile Platform Strain for High-Fidelity Multiplex Genome Editing

3.1. Author Contributions

This chapter represents a manuscript with contributions from Harneet S. Rishi (H.S.R.), Robert G. Egbert (R.G.E.), Benjamin A. Adler (B.A.A.), Dylan M. McCormick (D.M.M.), Esteban Toro (E.T.), Ryan T. Gill (R.T.G.), and Adam P. Arkin (A.P.A). Given the collaborative nature of this work, it is important to acknowledge the contributions of all authors: H.S.R. and R.G.E. led the work. H.S.R., R.G.E., B.A.A., D.M.M., and E.T. conducted experiments, and all authors interpreted results. H.S.R., R.G.E., B.A.A., and A.P.A. wrote the manuscript with input from all authors. A.P.A. supervised the research. R.T.G. and A.P.A. conceived of the research.

3.2. Introduction

The design-build-test (DBT) cycle is a common paradigm used in engineering disciplines. Within the context of synthetic biology, it is employed to engineer user-defined cellular functions for applications such as metabolic engineering, biosensing, and therapeutics[63,64]. The rapid prototyping of engineered functions has been facilitated by advances in *in vitro* DNA assembly, and plasmids have traditionally been used to implement designs *in vivo* given their ease-of-assembly and portability. However, for deployment in contexts beyond the laboratory such as large-scale industrial bioprocesses or among complex microbial communities, plasmid-based circuits suffer from multiple limitations: high intercellular variation in gene expression, genetic instability from random partitioning of plasmids during cell division, and plasmid loss in environments for which antibiotic use could disrupt native microbial communities or is economically infeasible[65,66]. These shortcomings can be ameliorated once a design is transferred from a plasmid to the host genome, which offers improved genetic stability and lower expression variation[67] along with reduced metabolic load[68]. However, behaviors optimized for plasmid contexts often do not map predictably to the genome. As such, building and testing designs directly on the genome can reduce the DBT cycle time and facilitate engineering cellular programs for complex environments.

Expanding synthetic biology efforts to genome-scale engineering has historically been limited by factors such as low endogenous rates of recombination, lack of optimized workflows for recombination, and uncertainty due to locus-dependent expression variability[69,70]. The advent of recombination-based genetic engineering (recombineering), which relies on homologous recombination proteins - often *exo, bet,* and *gam* from bacteriophage λ - in conjunction with linear donor DNA containing target homology and the desired mutations, has enabled genomic deletions, insertions, and point mutations at user-defined loci[71-75]. Recombineering has enabled generation of genomic discovery resources such as the *Escherichia coli* K-12 in-frame, single-gene deletion collection of non-essential genes (Keio collection) [16] and technologies such as trackable multiplex recombineering (TRMR), which enables genome-scale mapping of genetic modifications to traits of interest[18,19]. In addition, pooled library recombineering approaches such as CRISPR-enabled trackable genome engineering (CREATE) have combined CRISPR-Cas9 gene editing schemes with barcode tracking to enable high-throughput mutational profiling at single-nucleotide resolution on a genome-wide scale[20].

Meanwhile, techniques such as multiplex automated genome engineering (MAGE) have been developed to generate complex mutagenesis libraries by extending recombineering to simultaneously modify multiple genetic loci through iterative cycles of single-strand DNA (ssDNA) oligonucleotide recombination[17]. MAGE has enabled several genome-scale recombineering efforts such as the recoding of all 321 occurrences of TAG stop codons with synonymous TAA codons in a single *E. coli* strain[76,77], the removal of all instances of 13 rare codons from 42 highly expressed essential genes to study genome design constraints[78], the insertion of multiple T7 promoters across 12 genomic operons to optimize metabolite production[79], and the His-tagging of 38 essential genes that encode the entire translation machinery over 110 MAGE cycles for subsequent *in vitro* enzyme studies[80]. In addition, methods such as tracking combinatorial engineered libraries (TRACE) have been developed to facilitate the rapid, high-throughput mapping of multiplex engineered modifications from such genomic explorations to phenotypes of interest[81,82].

To achieve the high levels of recombination necessary to carry out large-scale, multiplexed genome editing, many of these studies required the use of mutagenic strains. Specifically, the endogenous methyl-directed mismatch repair (MMR) system, which acts to revert newly made recombineering modifications when active, was removed to more effectively retain targeted modifications in the standard MAGE strain EcNR2. While deactivation of MMR dramatically enhances recombination efficiency, it also increases the rate of background mutagenesis by 100–1000 fold[83,84]. Indeed, in converting all 321 occurrences of TAG stop codons to TAA stop codons, Lajoie *et al.* noted the addition of 355 unintended (i.e. off-target) mutations after the final strain construction[77].

Several approaches have been proposed to circumvent the use of MMR-deficient strains and thus avoid their high basal rates of off-target mutagenesis. Designs utilizing mismatches that are poorly repaired or that introduce silent mismatches near the desired mutation can be used to evade MMR, which only recognizes short mismatches[85]. Furthermore, oligos containing chemically modified bases can be used to evade MMR correction and increase allelic-replacement efficiency[86]. While these approaches boost recombination rates without increasing basal mutagenesis rates, they either limit the range of mutations that can be implemented or significantly increase oligonucleotide costs.

More recent efforts have focused on approaches to create a transient mutagenesis state. Specifically, cells are cycled between phases of elevated mutation rates, during which editing can take place efficiently, and phases of wild type-like mutation rates, during which cells can be propagated without incurring a significant number of background mutations. Nyerges *et al.* reported the use of a temperature-controlled mismatch repair deficient strain (*E. coli* tMMR) in which the MMR machinery can be transiently inactivated by shifting cells to a non-permissive temperature (36°C) during oligonucleotide incorporation and cell recovery and then reactivated by returning cells to the permissive temperature (32°C) for propagation[87]. This approach reduced the number of off-target mutations by 85%. In another work, Nyerges *et al.* developed pORTMAGE, a genome editing workflow that uses a dominant-negative *mutL* allele cloned on a broad-host range vector to transiently inactivate host MMR response using a similar temperature-dependent protocol[88]. The pORTMAGE system demonstrated high on-target mutation rates coupled with no reported off-tiarget mutations. The plasmid-based nature of the system enabled transfer to other bacteria such as *Salmonella enterica* and *Citrobacter freundii*. While both of these approaches reduce the off-target mutation rate, they restrict cell growth to 30–32°C and hence increase the time between recombineering cycles. In contrast, during preparation of this manuscript Bubnov *et al.* reported the construction of a plasmid system with a dominant-negative *mutS* allele and a novel conditionally replicating origin that enabled plasmid propagation and genome editing at higher temperatures (e.g. 37°C)[89]. Finally, Lennen *et al.* developed a plasmid-based MAGE system, Transient Mutator Multiplex Automated Genome Engineering (TM-MAGE). In TM-MAGE, *E. coli* Dam methylase is inducibly overexpressed to transiently limit MMR and thus enable high allelic replacement efficiencies with a 12–33 fold lower off-target mutation rate than strains with fully disabled MMR[90].

To date, recombineering has advanced genome engineering in what can be categorized as two thematic research areas: developing genetic tools for model microbial hosts and multi-stage genome editing for metabolic engineering and synthetic biology. These research areas have divergent requirements. In the former case, the portability of the recombineering vector is key to easily generate mutations across bacterial hosts and allow removal of the recombineering cassette following mutagenesis. Recombineering plasmids function as effective and convenient vectors for a portable genome engineering solution. In the latter case, a chromosomally integrated recombineering cassette has proven effective for multi-cycle recombineering up to dozens of rounds to simultaneously target many genomic loci related to an engineered function. Unfortunately, there is no chromosomally integrated recombineering system reported to date that achieves high-efficiency ssDNA recombination without the side-effect of a significantly elevated global mutagenesis rate. Hence, researchers still face a trade-off between genome editing efficiency and genome stability.

Here we present a rational genome engineering approach to develop a high-fidelity recombineering platform strain, called BioDesignER, with enhanced recombineering efficiency, low off-target mutagenesis rates, and short editing cycle times. We refactored the λ-Red machinery in *E. coli* K-12 MG1655-derived EcNR1 to decrease cycle time and reduce toxicity, stacked genetic modifications shown to increase recombination rates, and characterized gene expression across the chromosome at curated integration loci, herein referred to as Safe Sites. We also introduced genomic modifications to independently control four transcriptional regulators of gene expression and characterized the induction regime for each regulator. We profiled the growth and ssDNA recombination rates of BioDesignER with a dual-fluorescent reporter cassette integrated at each Safe Site and also demonstrated the retention of double-stranded DNA (dsDNA) recombination capabilities in the strain. We performed a comparative study of background mutagenesis rates of our strain and alternative platform strains using a fluorescent reporter-based fluctuation assay and found that

BioDesignER exhibited a 4.2-fold lower mutagenesis rate compared to the widely used recombineering strain EcNR2. Finally, we compared the multi-cycle accumulation of targeted mutations for BioDesignER and other high-efficiency recombineering strains and found that BioDesignER exhibited similar multiplex editing efficiencies to EcNR2.nuc5-, a persistent mutator strain with the highest reported ssDNA recombination efficiency. BioDesignER is a high-fidelity genome engineering strain that uniquely enables high-efficiency recombineering while retaining a low basal mutagenesis rate.

3.3. Results

3.3.1. Rational Strain Design

We introduced multiple targeted modifications to an MG1655-derivative strain to decrease recombination cycle time, reduce toxicity of the recombination machinery, and introduce a transient hypermutation phenotype via hypermethylation (**Figure 3-1AB**; **Table 3-1**). Using EcNR1[17] as the host, we refactored the λ-Red recombination machinery, which consists of the genes *exo*, *bet*, and *gam*, and serves as the basis for mediating homology-directed recombination of ssDNA and dsDNA products. To reduce recombineering cycle times, we replaced the temperature-inducible regulation of the λ-Red locus with a TetR-regulated design (**Figure 3-1C**). This allowed us to propagate cells at 37°C instead of 30–32°C during all phases of a recombineering workflow: competent cell prep, λ induction, cell recovery, and selection. We also minimized the λ prophage by deleting the λ-*kil* gene, which has been reported to be responsible for the cell death phenotype observed under λ-Red expression[91], and other dispensable phage genes. Finally, we introduced DNA adenine methyltransferase (*dam*) to the λ-Red operon of our strain. Co-induction of *dam* with the λ-Red recombination genes results in transient hypermutation via hypermethylation, which has been reported to enable incorporated mutations to evade MMR[90].

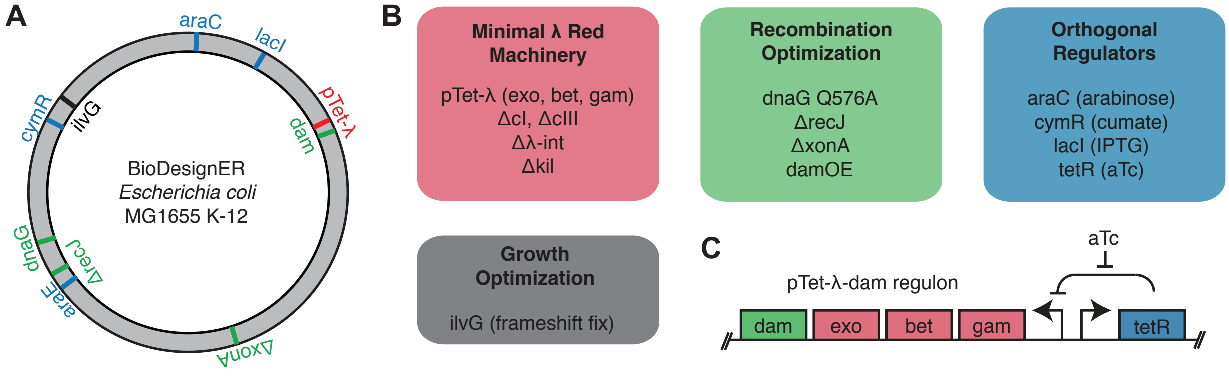


Figure 3-1. Overview of genetic modifications in BioDesignER.

(**A**) Chromosome map of the BioDesignER strain (derived from E. coli MG1655 K-12) with modifications made in the platform strain mapped to corresponding positions on the genome. (**B**) Functional grouping of genomic modifications based on purpose in platform strain (e.g. minimization of λ-Red machinery, optimization of recombination efficiency, implementation of multiple orthogonal regulators, or optimization of growth). (**C**) Genetic architecture of refactored λ-Red machinery and dam over-expression construct regulated by TetR.

Table 3-1. Genotypes of abbreviated strains.

Strain	Lab ID	Parent strain	Genetic modifications	Reference
EcNR1	RE002	MG1655	λ-Red(*ampR*)::*bioA*/*bioB*	Wang *et al.* 2009
EcNR2	ET046	EcNR1	*cmR*::*mutS*	Wang *et al.* 2009
EcNR2.nuc5-	ET003	EcNR2	*dnaG*.Q576A ∆*recJ* ∆*xonA* ∆*xseA* ∆*exoX* ∆*red-α*	Mosberg *et al.*2012
pTet-*λ*	RE574	MG1655	pTet2-*gam-bet-exo*/*tetR*/*ampR*::*bioA*/*B ilvG+*	This study
damOE	RE824	MG1655	pTet2-*gam-bet-exo-dam*/*tetR*/*ampR*::*bioA*/*B ilvG+*	This study
dnaG.Q	RE626	pTet-*λ*	*dnaG*.Q576A	This study
exo1	RE628	pTet-*λ*	*dnaG*.Q576A ∆*recJ*	This study
exo2	HR146	pTet-*λ*	*dnaG*.Q576A ∆*recJ* ∆*xonA*	This study
BioDesignER	RE630	damOE	*dnaG*.Q576A *lacIQ1* Pcp8-*araE* ∆*araBAD* pConst-*araC* ∆*recJ*∆*xonA*	This study

To remove a valine-sensitive growth defect present in *E. coli* K-12, we restored expression of *ilvG*. K-12 contains three acetohydroxy acid synthases (*ilvB*, *ilvG*, and *ilvH*) that are involved in branch-chained amino acid biosynthesis. K-12 does not express *ilvG* due to a natural frameshift mutation and thus exhibits a growth defect in the presence of exogenous valine and the absence of isoleucine[92,93]. This valine-sensitive growth phenotype is alleviated by restoration of *ilvG*[94]. Using oligo-mediated recombination (**Materials and Methods**), we removed the frameshift mutation in the endogenous *ilvG* gene, which has been reported to enable faster growth in minimal media. We called this strain pTet-*λ*.

We next incorporated genomic modifications shown to improve recombination efficiency. Using a scar-free genome engineering workflow that utilizes a novel *thyA* selection/counter-selection cassette containing a fluorescent marker (**Figure 3-2**, **Materials and Methods**), we iteratively generated multiple beneficial mutations. For example, genetic variants of DNA primase (*dnaG*) enhance recombination efficiency by increasing the length of Okazaki fragments, thus exposing longer stretches of the lagging strand of the replication fork to ssDNA recombination[95]. We incorporated into our strain the *dnaG*.Q576A variant, which was shown to boost recombination efficiency more than other *dnaG* mutants in EcNR2.

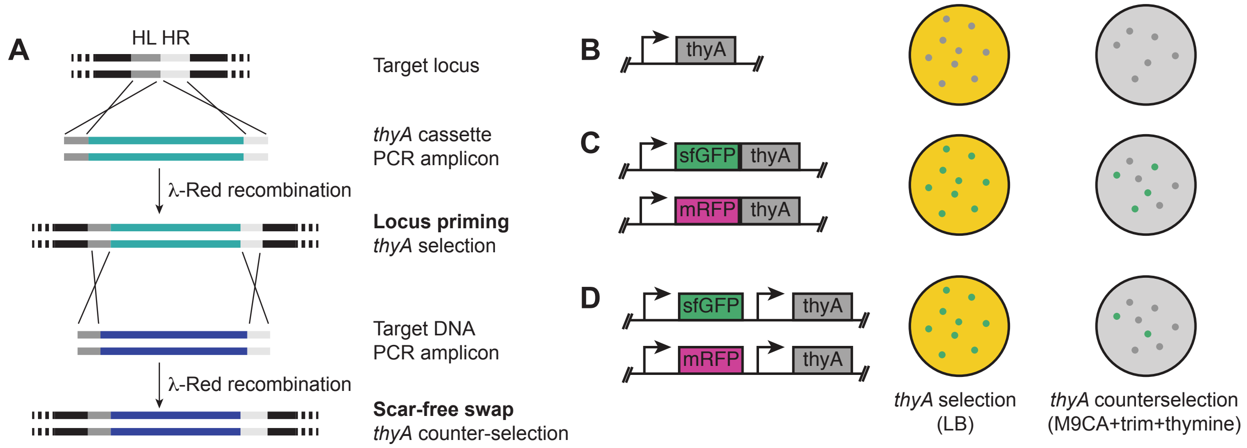


Figure 3-2. Methods and genetic cassettes for *thyA*-based selection and counter-selection.

(**A**) Overview for scar-free, two-step genome integration with thyA selection and counter-selection. Starting from a thyA deletion strain, a linear, double-strand DNA cassette including the thyA gene is amplified with 35-50 base pair homology (HL, HR) for the target genomic locus. The cassette is integrated on the genome using standard l-Red recombination with selection on LB agar. The primed genomic locus is swapped to the target sequence by amplifying the target DNA cassette with the same homology and integrating the cassette with selection on M9 minimal media agar supplemented with casamino acids, trimethoprim, and thymine. (**B**) The selection-counterselection cassette with thyA includes selection on LB and counterselection on M9 minimal media supplemented with casamino acids, trimethoprim and selection-counterselection. Successful transformants at each integration stage are screened via colony PCR and Sanger sequencing. (**C**) Selection-counterselection cassettes with translational fusions of a fluorescent reporter (sfGFP or mRFP1) to thyA allow rapid screening of integration via fluorescence phenotype. Fluorescence screening reduces the number of colonies that must be screened to identify expected clones. (**D**) Selection-counterselection cassettes with tandem promoters for a fluorescent reporter and thyA increase counterselection screening efficiency by decoupling the expression of fluorescence and counterselection markers.

Endogenous nucleases can degrade exogenous DNA used in recombineering workflows. The removal of a set of five nuclease genes (*endA, exoX, recJ, xonA,* and *xseA*) has been shown to improve ssDNA recombination efficiency[96]. However, while this exonuclease knockout strain, EcNR2.nuc5-, exhibited increased recombination efficiency, it also resulted in a lower post-electroporation growth rate compared to EcNR2. This suggested that deletion of the entire set of nucleases introduces to the strain an undesirable physiological defect. To avoid such growth defects, which are compounded for workflows requiring multiple recombineering cycles, we looked to systematically combine exonuclease knockouts that distinctly improve recombination rates. We constructed individual knockouts of each of the five exonucleases in the pTet-λ *dnaG*.Q576A strain and measured the recombineering efficiency of the resulting strains. We assayed recombination efficiency for each exonuclease knockout using oligo mediated recombination at a genomically-encoded *sfGFP* reporter. In this assay, recombination of an oligo designed to introduce a premature stop codon into *sfGFP* results in a loss of fluorescence that can be quantified using flow cytometry. Deletions of *xonA* (4.5 ± 0.3%) (mean ± 1 standard deviation) and *recJ* (2.7 ± 0.1%) showed the greatest efficiencies, while the remaining exonuclease deletions yielded nominal efficiencies (<2%) (**Table 3-2**). Based on these results, we deleted only two of the five exonucleases (*recJ*, *xonA*) in the next step of strain construction. While deletion of the λ-Red exonuclease (*exo*) can also promote stability of exogenous ssDNA[96], we opted to retain it due to its role in dsDNA recombination. The culmination of these genetic modifications in addition to the inducible regulator modifications described below resulted in the BioDesignER strain.

Table 3-2. Recombination efficiencies for individual exonuclease deletions.

Strain	Exonuclease Deletion Strain Genotype	Single Cycle Conversion Rate (mean ± stdev)
HR130	MG1655 l-Red(*sfGFP-kanR*, pTet2:*gam-bet-exo-dam*(fs), pN25:*tetR*, *ampR)*::bioA/bioB ilvG$^+$ *ΔthyA spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE ΔaraBAD* pConst-*araC* RT2P::*endA*	1.6 ± 0.8
HR131	MG1655 l-Red(*sfGFP-kanR*, pTet2:*gam-bet-exo-dam*(fs), pN25:*tetR*, *ampR)*::bioA/bioB ilvG$^+$ *ΔthyA spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE ΔaraBAD* pConst-*araC* RT2P::*exoX*	1.4 ± 0.6
HR132	MG1655 l-Red(*sfGFP-kanR*, pTet2:*gam-bet-exo-dam*(fs), pN25:*tetR*, *ampR)*::bioA/bioB ilvG$^+$ *ΔthyA spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE ΔaraBAD* pConst-*araC* RT2P::*recJ*	2.7 ± 0.1
HR133	MG1655 l-Red(*sfGFP-kanR*, pTet2:*gam-bet-exo-dam*(fs), pN25:*tetR*, *ampR)*::bioA/bioB ilvG$^+$ *ΔthyA spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE ΔaraBAD* pConst-*araC* RT2P::*xonA*	4.5 ± 0.3
HR134	MG1655 l-Red(*sfGFP-kanR*, pTet2:*gam-bet-exo-dam*(fs), pN25:*tetR*, *ampR)*::bioA/bioB ilvG$^+$ *ΔthyA spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE ΔaraBAD* pConst-*araC* RT2P::*xseA*	1.5 ± 0.7

To assess the effect of BioDesignER modifications on strain fitness, we measured the growth rates of key strains in the modification lineage in LB rich media (**Figure 3-3A**). We noted that, in general, doubling times decreased as additional modifications were made. Additionally, in contrast to cell death reported for extended co-expression of λ-*kil* with the recombination machinery[91], we observed only a slight increase in doubling time when expressing the refactored λ-Red cassette.

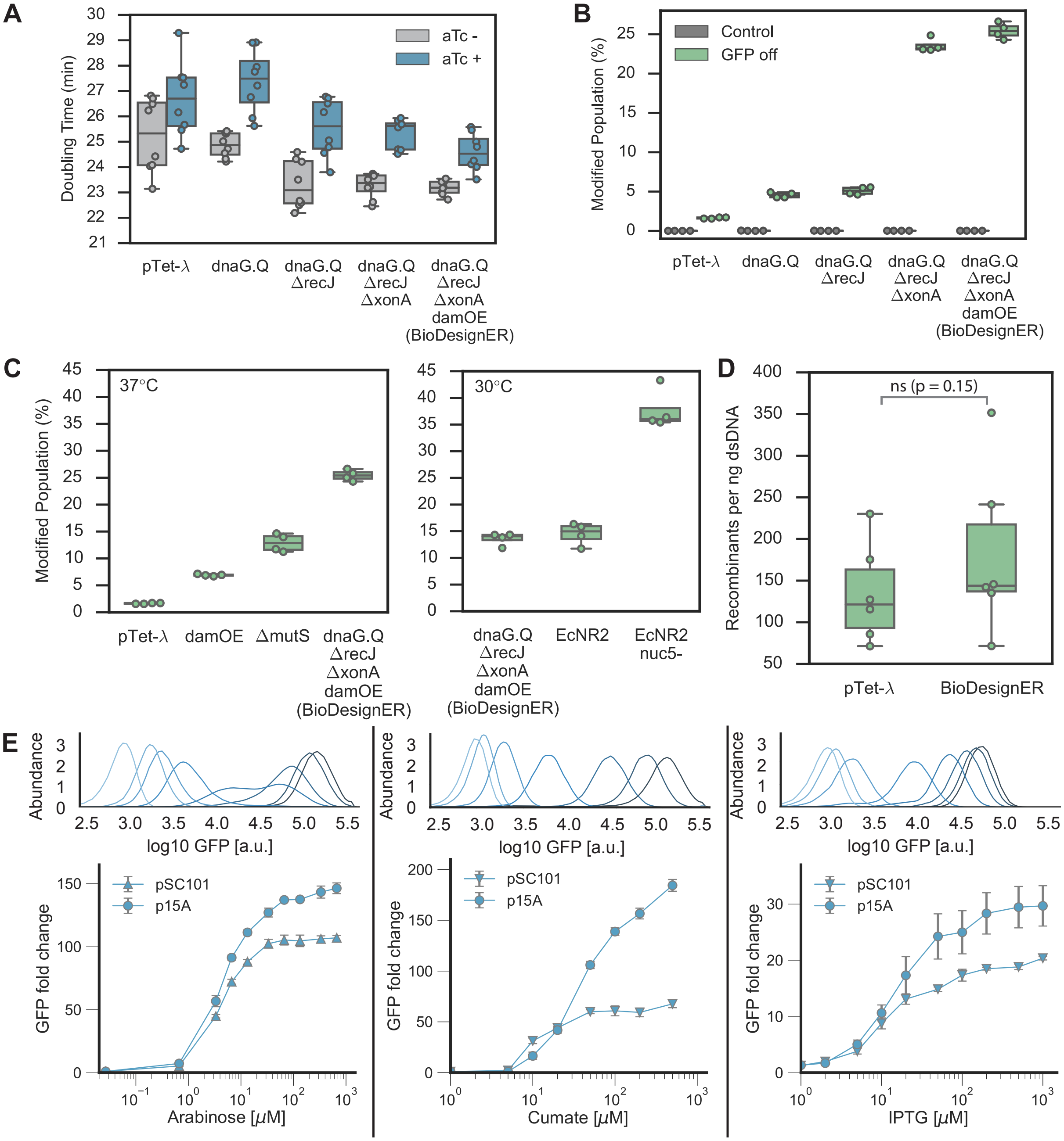


Figure 3-3. Strain characterization for BioDesignER.

(**A**) Doubling times of strains grown at 37°C for selected strains of BioDesignER lineage starting with pTet-λ. Additional modifications shown moving to the right. Doubling times reported for strains grown with (blue) and without (gray) aTc induction to show the effect of λ-Red expression on growth. Data represented as box plots overlaid with corresponding data points. (**B**) ssDNA recombination efficiency enhancements for the strain lineage quantified via inactivation frequency of an sfGFP reporter measured via flow cytometry. (**C**) The recombination efficiency of BioDesignER compared to pTet-λ harboring modifications that interfere with mismatch repair (damOE, ΔmutS) (left, 37°C) and to canonical recombineering strains such as EcNR2 and EcNR2.nuc5- (right, 30°C). (**D**) Transformation efficiency of BioDesignER compared to pTet-λ (control) to show retention of dsDNA recombination efficiency. P-value from Mann–Whitney U-test; ns, not significant. (**E**) Flow cytometry traces (top) with corresponding fold-change response curves (bottom) for each inducible, orthogonal regulator. Inducer concentrations used for flow cytometry traces are: 0, 0.33, 0.67, 1.3, 3.3, 6.7, 33, and 130 μM (arabinose); 0, 2, 5, 10, 20, 50, and 100 μM (cumate); 0, 1, 2, 5, 10, 20, 50, and 500 μM (IPTG). Note that darker colors in flow cytometry traces correspond to increasing inducer concentrations in the provided ranges.

3.3.2. BioDesignER recombineering enhancements

ssDNA recombination enhancement

To quantify recombineering enhancements of key BioDesignER modifications, we measured ssDNA recombination rates for several strain intermediates. We integrated a dual fluorescent reporter cassette expressing both *sfGFP* and *mKate2* at a common genomic locus for each strain of the lineage and quantified ssDNA recombination efficiency. For each strain we transformed an oligo to inactivate *sfGFP* via incorporation of a premature stop codon. We also performed a control reaction in each case using water in place of oligo. After recovery and outgrowth, we measured the fluorescence profiles of each strain using flow cytometry (**Figure 3-3B**). We observed increases in recombination efficiency at each modification stage with single cycle conversion rates improving from 1.6 ± 0.1% in pTet-λ to 25.4 ± 1.0% in BioDesignER.

To investigate the efficacy of mismatch repair evasion on recombination efficiency, we compared BioDesignER against pTet-λ derivative strains containing mismatch repair modifications and against two standard Δ*mutS* recombineering variants, EcNR2 and EcNR2.nuc5-. BioDesignER (25.4 ± 1.0%) exhibits much higher recombination efficiency than pTet-λ with *dam* over-expression (damOE, 6.91 ± 0.19%) or Δ*mutS* (12.9 ± 1.7%) as hypermutagenesis strategies (**Figure 3-3C**, left panel). Performing the same recombineering experiments at 30°C and comparing to EcNR2 and EcNR2.nuc5-, which are constrained to growth at 30°C, we found that BioDesignER (13.6 ± 1.2%) exhibited recombination rates comparable to EcNR2 (14.5 ± 2.1%), yet approximately 3-fold lower than EcNR2.nuc5- (37.7 ± 3.8%) (**Figure 3-3C**, right panel). We were surprised to find that the recombineering efficiency of BioDesignER decreased by nearly 2-fold when grown at a lower temperature.

dsDNA recombination enhancement

Knocking out endogenous exonucleases has been reported to significantly reduce or abolish dsDNA recombination efficiency[96]. We measured the efficiency of dsDNA recombination in pTet-λ and BioDesignER and found no significant reduction in recombination efficiency (**Figure 3-3D**). This result suggests that λ-Exo is sufficient to process dsDNA recombination templates in the absence of RecJ and ExoI ssDNA exonucleases. A previous study reported that dsDNA recombination is at least an order of magnitude less efficient in a four-nuclease deficient genotype (Δ*exoX*, Δ*recJ*, Δ*xseA*, and Δ*xonA*) with abolished dsDNA recombination activity in a three-nuclease (Δ*recJ*, Δ*xseA* and Δ*xonA*) knockout[96]. We note here that we were successful in generating dsDNA recombinants in EcNR2.nuc5- at a similar efficiency to EcNR2 with no alteration to the recombineering protocol, suggesting that another nuclease is aiding dsDNA recombination in *E. coli* or that recombination can occur through an exonuclease-independent mechanism.

3.3.3. Control of multiple independent regulators

BioDesignER expresses transcriptional regulators that utilize four independent small-molecule inducers to allow multi-input control of synthetic circuits, biosynthetic pathways, or gene editing tools. The strain produces the repressors TetR, LacI, and CymR as well as the activator AraC. TetR is expressed from the λ prophage element native to EcNR1. We incorporated the transcriptional over-expression allele *lacI*^{Q1} to boost LacI production, which allows efficient regulation of multi-copy plasmids[97]. We also introduced the tight and titratable regulator CymR[98], which is inactivated by the small molecule cumate. To improve gene regulation by arabinose, we replaced the arabinose-sensitive promoter of the *araE* transporter gene with a constitutive promoter to eliminate all-or-none expression and allow titratable induction[99]. In conjunction with this modification, we introduced a constitutive promoter to drive expression of AraC and deleted the *araBAD* operon to eliminate arabinose degradation via catabolism.

To characterize the induction profiles of each regulator, we quantified the fluorescence levels and growth rates of cells transformed with multi-copy plasmids. We constructed a set of GFP expression plasmids with promoters responsive to each regulator (**Figure 3-3E**, see **Figure 3-4** for sequence-level details) and transformed each plasmid into BioDesignER. Gene expression profiles were characterized by measuring single-cell fluorescence and bulk growth and fluorescence.

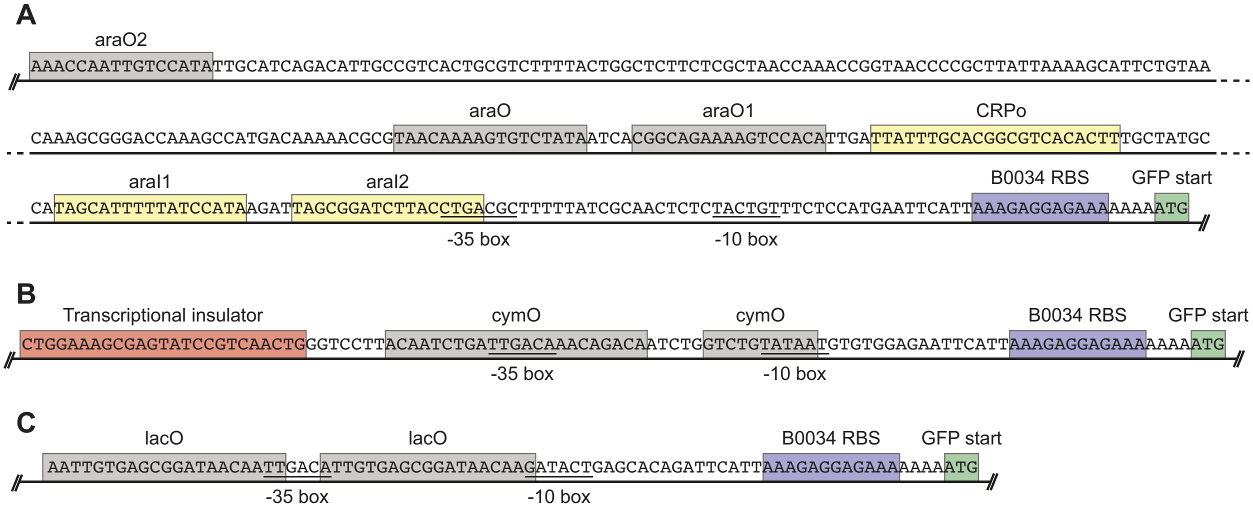

A
araO2
AAACCAATTGTCCATATTGCATCAGACATTGCCGTCACTGCGTCTTTTACTGGCTCTTCTCGCTAACCAAACCGGTAACCCCGCTTATTAAAAGCATTCTGTAA
araO
araO1
CRPo
CAAAGCGGGACCAAAGCCATGACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATGC
araI1
araI2
B0034 RBS
GFP start
CATAGCATTTTTATCCATAAGATTAGCGGATCTTACCTGACGCTTTTTATCGCAACTCTCTACTGTTTCTCCATGAATTCATTAAAGAGGAGAAAAAAAATG
-35 box
-10 box
B
Transcriptional insulator
cymO
cymO
B0034 RBS
GFP start
CTGGAAAGCGAGTATCCGTCAACTGGGTCCTTACAATCTGATTGACAAACAGACAATCTGGTCTGTATAATGTGTGGAGAATTCATTAAAGAGGAGAAAAAAAATG
-35 box
-10 box
C
lacO
lacO
B0034 RBS
GFP start
AATTGTGAGCGGATAACAATTGACATTGTGAGCGGATAACAAGATACTGAGCACAGATTCATTAAAGAGGAGAAAAAAAATG
-35 box
-10 box

Figure 3-4. Nucleotide-level detail for inducible promoters.

Transcription and translation control elements for GFP expression induced by arabinose (**A**), cumate (**B**) and IPTG (**C**) are shown. Grey boxes denote repressor binding sequence motifs. Yellow boxes denote activator binding motifs. Purple boxes denote a common ribosome binding site (RBS). Green boxes denote the start codon for GFP. The red box denotes a synthetic transcriptional insulator designed using R2oDNA Designer[100]. Full sequence details for each associated plasmid can be found in the Benchling repository (links available in Supplementary Table S2). araO1: high-affinity AraC operator; araO2: truncated low-affinity AraC operator; CRPo: cAMP receptor protein operator; araI1: high-affinity AraC-arabinose operator; araI2: low-affinity AraC-arabinose operator; cymO: CymR operator; lacO: LacI operator.

Fold-change induction for each regulator increased with plasmid copy number while no leaky expression was observed for low-copy and medium-copy plasmids. For plasmids with the low-copy replication origin pSC101, we observed mean fold-change induction levels of 107, 68, and 20 for arabinose, cumate, and isopropyl β-D-1-thiogalactopyranoside (IPTG), respectively. For plasmids with the medium-copy replication origin p15A, we observed mean fold-change induction levels of 146, 184, and 30 for arabinose, cumate, and IPTG, respectively. In both copy-number contexts, GFP expression with no inducer was indistinguishable from a control plasmid lacking *gfp*. We found that repressor levels were insufficient to fully repress GFP expression on plasmids with the ColE1 replication origin. We note that AraC-regulated GFP expression saturates near 33 μM (5 μg/ml, 0.0005%) arabinose, a much lower saturation point than common plasmid-based systems (0.1% arabinose).

Single-cell fluorescence distributions observed through flow cytometry revealed unimodal distributions of GFP expression for nearly all induction conditions (**Figure 3-3E**). GFP expression from both cumate- and IPTG-responsive promoters produced monotonic, decreasing coefficient of variation noise profiles for increasing inducer levels (**Figure 3-5**). For arabinose induction, despite introducing modifications consistent with Khlebnikov *et al.*, we observed significant cell–cell variability at two intermediate arabinose levels. Specifically, we observed a maximum coefficient of variation at 3.3 μM (**Figure 3-5**), seen in **Figure 3-3E** as a broad, weakly bimodal fluorescence distribution.

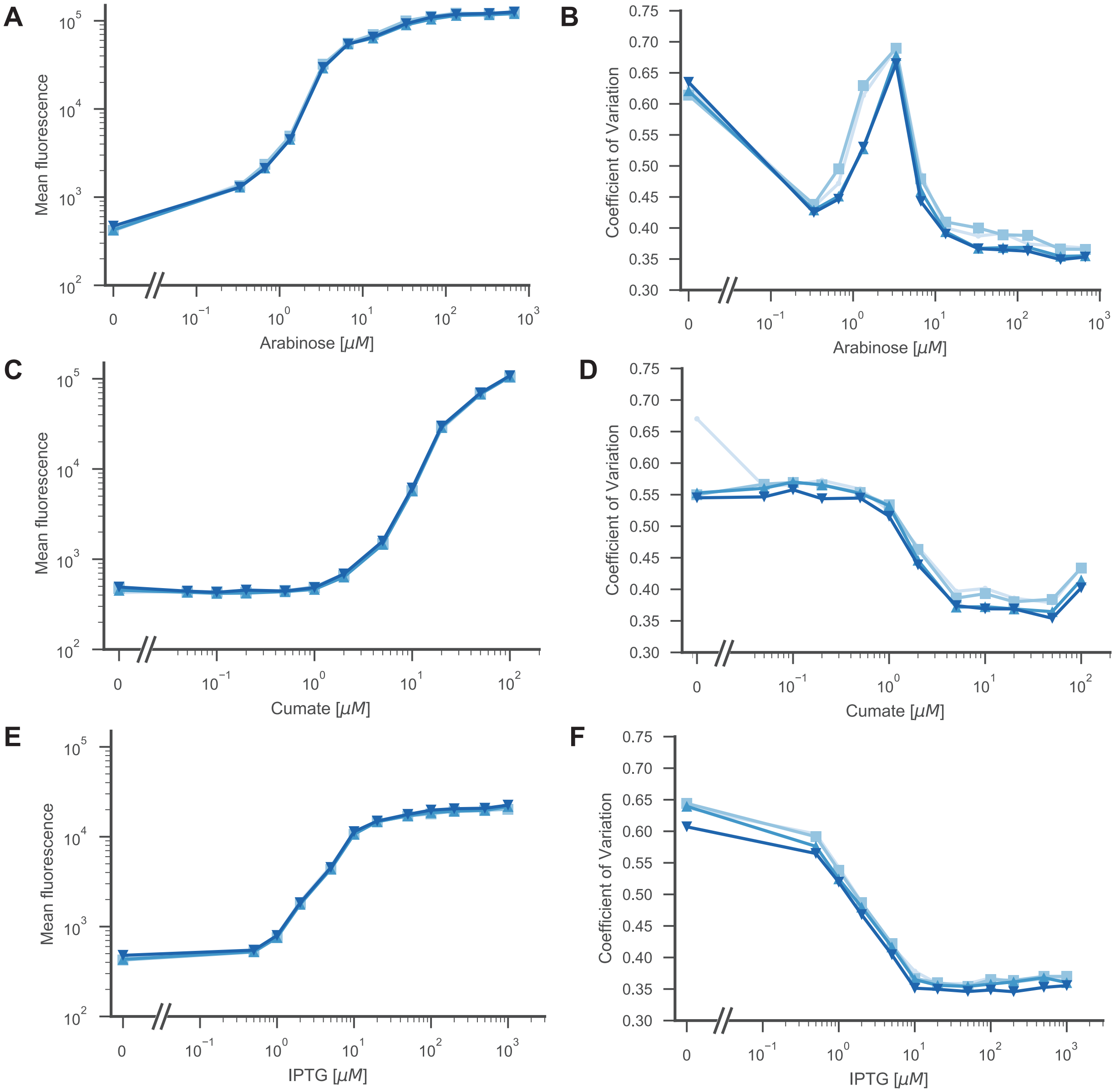


Figure 3-5. Noise profiles for inducible regulators.

Mean fluorescence (**A**, **C**, **E**) and coefficient of variation (**B**, **D**, **F**) profiles for GFP fluorescence measured via flow cytometry. Each profile is a function of inducer concentration for arabinose (**A**, **B**), for cumate (**C**, **D**), and for IPTG (**E**, **F**). In each case, BioDesignER was transformed with the appropriate inducible plasmid as described in the text and shown in **Figure 3-3**. Individual traces represent measurements from biological replicates run on the same day. Mean fluorescence was calculated as the geometric mean of each fluorescence distribution. Coefficient of variation was calculated from the geometric mean and standard deviation of each fluorescence distribution.

3.3.4. Characterization of genome integration Safe Sites

Genome integration Safe Sites

To aid identifying genomic loci that provide reliable gene expression and recombination efficiency for future engineering efforts, we characterized a curated list of integration loci across the *E. coli* K-12 genome. The resulting eight genomic loci, termed Safe Sites, were chosen based on several criteria to minimize disruption to local chromosomal context upon integration of synthetic DNA constructs (**Figure 3-6A** and **Figure 3-7**). Specifically, the integration Safe Sites are intergenic regions located between two convergently transcribed, non-essential genes that do not exhibit any phenotypes or growth defects across the majority of biochemical conditions screened in previous high-throughput studies[16,24], and contain no annotated features (small RNAs, promoters, transcription factor-binding sites) according to RegulonDB[101] (**Table 3-3**).

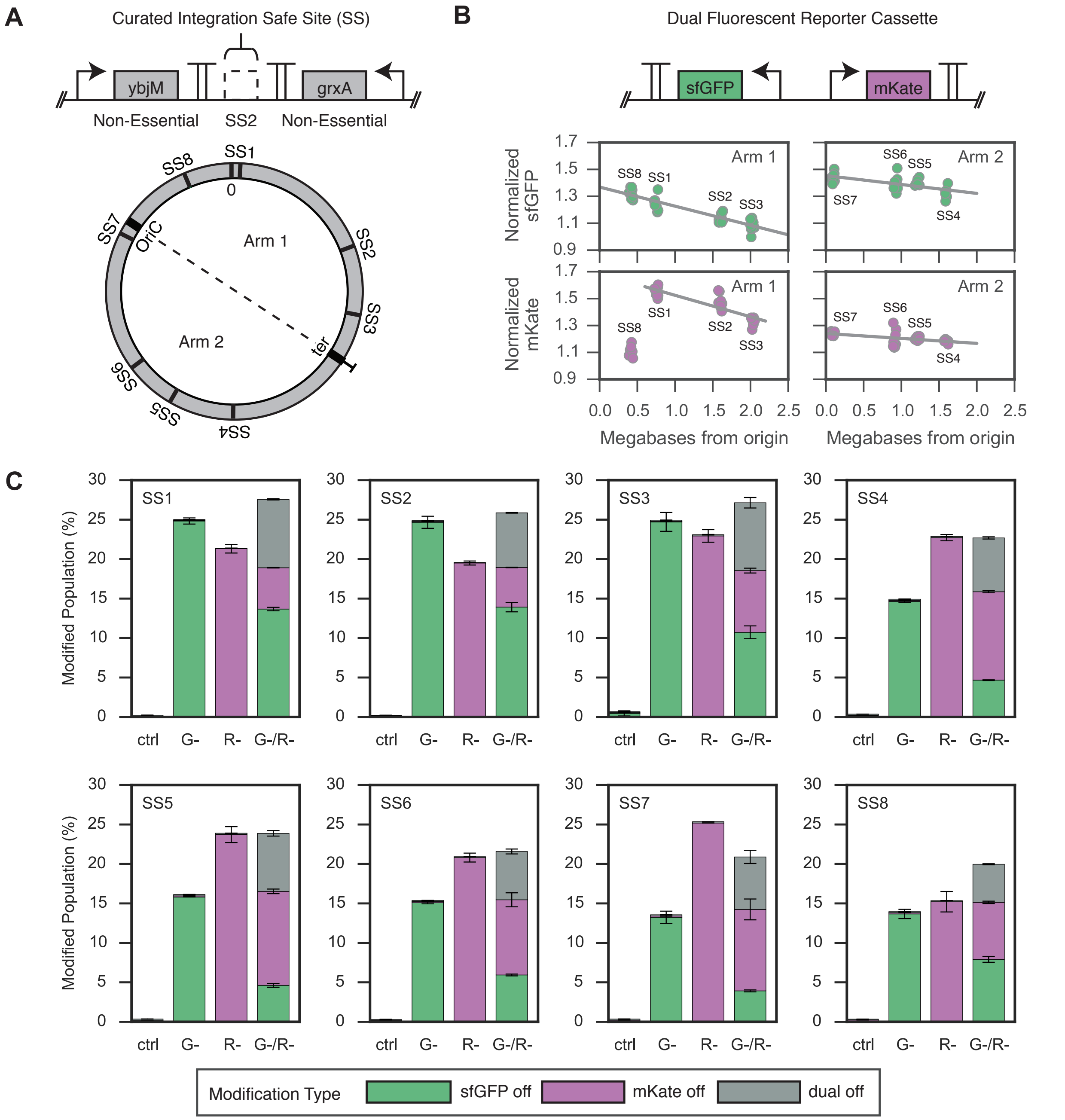
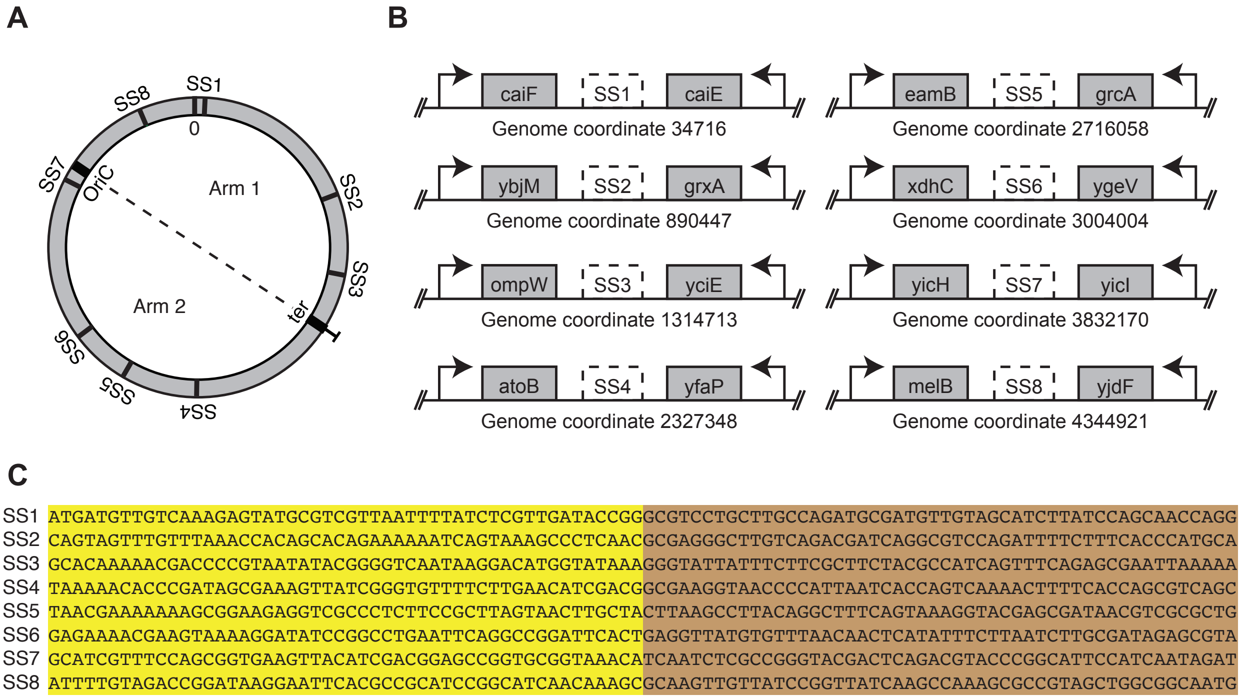


Figure 3-6. Expression and recombination characterization at BioDesignER Safe Sites.

(**A**) Circular map of the BioDesignER chromosome with Safe Sites mapped to corresponding genome position and chromosomal arm (replichore). (**B**) Genetic architecture of dual fluorescent reporter construct (top) and observed expression of reporters when integrated at each Safe Site on the chromosome (bottom). Replicate measurements of normalized expression levels for each reporter arrayed by chromosomal arm on which construct is integrated. (**C**) ssDNA recombination rates at each Safe Site for four independent recombineering reactions. X-axis denotes transformed oligo(s) (G- for sfGFP, R- for mKate) or ctrl (water). Bar height corresponds to the mean of two measurements and error bars represent span of data. Stacked bar plots for each reaction represent population fractions containing one of three possible modifications (sfGFP off, mKate off, or dual off when both reporters inactivated).



```
SS1 ATGATGTTGTCAAAGAGTATGCGTCGTTAATTTTATCTCGTTGATACCGGGCGTCCTGCTTGCCAGATGCGATGTTGTAGCATCTTATCCAGCAACCAGG
SS2 CAGTAGTTTGTTTAAACCACAGCACAGAAAAAATCAGTAAAGCCCTCAACGCGAGGGCTTGTCAGACGATCAGGCGTCCAGATTTTCTTTCACCCATGCA
SS3 GCACAAAAACGACCCCGTAATATACGGGGTCAATAAGGACATGGTATAAAGGGTATTATTTCTTCGCTTCTACGCCATCAGTTTCAGAGCGAATTAAAAA
SS4 TAAAAACACCCGATAGCGAAAGTTATCGGGTGTTTTCTTGAACATCGACGGCGAAGGTAACCCCATTAATCACCAGTCAAAACTTTTCACCAGCGTCAGC
SS5 TAACGAAAAAAAGCGGAAGAGGTCGCCCTCTTCCGCTTAGTAACTTGCTACTTAAGCCTTACAGGCTTTCAGTAAAGGTACGAGCGATAACGTCGCGCTG
SS6 GAGAAAACGAAGTAAAAGGATATCCGGCCTGAATTCAGGCCGGATTCACTGAGGTTATGTGTTTAACAACTCATATTTCTTAATCTTGCGATAGAGCGTA
SS7 GCATCGTTTCCAGCGGTGAAGTTACATCGACGGAGCCGGTGCGGTAAACATCAATCTCGCCGGGTACGACTCAGACGTACCCGGCATTCCATCAATAGAT
SS8 ATTTTGTAGACCGGATAAGGAATTCACGCCGCATCCGGCATCAACAAAGCGCAAGTTGTTATCCGGTTATCAAGCCAAAGCGCCGTAGCTGGCGGCAATG
```

Figure 3-7. Safe Site locations across chromosome with genomic context and sequencing information.

(**A**) Circular map of BioDesignER chromosome with Safe Sites mapped to corresponding genome position and chromosomal arm (replichore). (**B**) Genomic context with flanking genes for each Safe Site. Genome coordinate indicated for the middle position of the 100 bp region defined as a Safe Site. Promoter positions and gene lengths are not scaled to actual positions on the chromosome. Links to RegulonDB for each Safe Site locus can be found in Supplementary Table S6. (**C**) Nucleotide-level information for each Safe Site. Safe Site upstream positions 1-50 are highlighted in yellow. Safe Site downstream positions 51-100 are highlighted in brown.

Table 3-3. RegulonDB feature annotations for Safe Sites

Safe Site	Gene 1	Gene 2	RegulonDB link for local genome context	Annotated Features in intergenic region according to RegulonDB?
1	*caiF*	*caiE*	http://regulondb.ccg.unam.mx/gene?organism=ECK12&term=ECK120002330&format=jsp&type=gene	No
2	*ybjM*	*grxA*	http://regulondb.ccg.unam.mx/gene?organism=ECK12&term=ECK120000410&format=jsp&type=gene	No
3	*ompW*	*yciE*	http://regulondb.ccg.unam.mx/gene?term=ECK120001108&organism=ECK12&format=jsp&type=gene	Yes - annotated rho-independent terminator at 3' end of ompW
4	*atoB*	*yfaP*	http://regulondb.ccg.unam.mx/gene?term=ECK120001615&organism=ECK12&format=jsp&type=gene	No
5	*eamB* *(yfiK)*	*grcA* *(yfiD)*	http://regulondb.ccg.unam.mx/gene?term=ECK120002298&organism=ECK12&format=jsp&type=gene	No
6	*xdhC*	*ygeV*	http://regulondb.ccg.unam.mx/gene?organism=ECK12&term=ECK120004051&format=jsp&type=gene	No

7	*yicH*	*yicI*	http://regulondb.ccg.unam.mx/gene?organism=ECK12&term=ECK120001628&format=jsp&type=gene	No
8	*melB*	*yjdF*	http://regulondb.ccg.unam.mx/gene?organism=ECK12&term=ECK120004322&format=jsp&type=gene	No

To characterize gene expression variation across the chromosome, we measured the expression of dual-fluorescent reporters (*sfGFP* and *mKate2*) integrated into BioDesignER at each Safe Site. We observed a linear decrease in expression for both sfGFP (pearson $r_{\mathrm{sfGFP,arm1}} = -0.91$, pearson $r_{\mathrm{sfGFP,arm2}} = -0.65$, $p_{\mathrm{sfGFP}} < 0.05$, permutation test) and mKate2 (pearson $r_{\mathrm{mKate,arm1}} = -0.85$, pearson$r_{\mathrm{mKate,arm2}} = -0.51$, $p_{\mathrm{mKate}} < 0.05$, permutation test) reporters with respect to distance from the chromosomal origin (**Figure 3-6B**). This result was consistent with expected variations in local chromosomal copy number due to bi-directional replication dynamics during growth[102-104]. Interestingly, we observed a much stronger correlation of expression to distance from replication origin for chromosome Arm 1, though mKate2 expression at Safe Site 8 was a low outlier. We also assessed the effect of integration at each Safe Site on cellular fitness by measuring growth rates for each integration strain. We observed that, in general, genomic integration and expression from each Safe Site did not reduce growth rate, though Safe Site 8 displayed a nominal decrease when grown under aTc induction (**Figure 3-8**). The two unexpected results at Safe Site 8 suggest that it may not be a reliable locus for integration and expression.

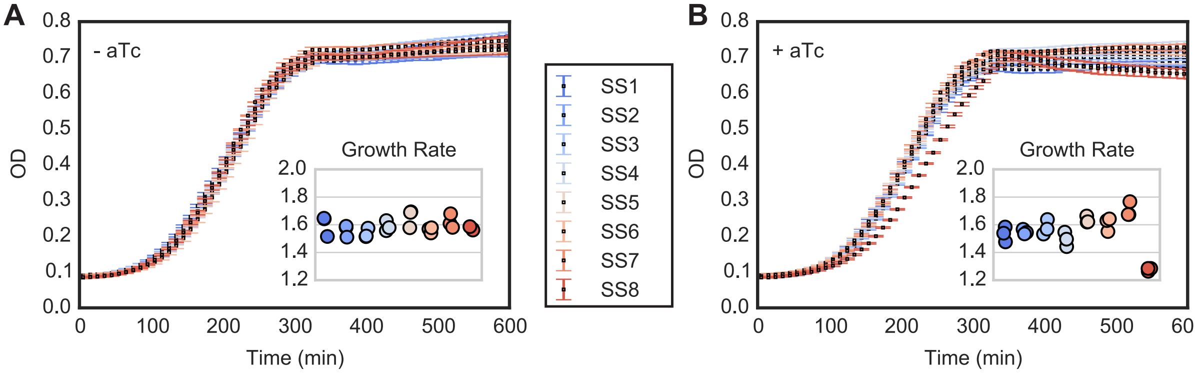


Figure 3-8. Effect of genome integration at Safe Sites.

(**A**) Growth profiles for BioDesignER with the dual-fluorescent reporter construct integrated at different Safe Sites without aTc induction (corresponding growth rates shown in figure inset). (**B**) Growth profiles for BioDesignER with the dual-fluorescent reporter construct integrated at different Safe Sites with aTc induction (corresponding growth rates shown in figure inset).

Recombination rates across Safe Sites

Changes in local chromosomal structure may lead to unexpected fluctuations in recombination efficiency at various locations across the genome. To characterize recombination efficiency as a function of chromosomal locus for BioDesignER, we performed three independent ssDNA oligo-mediated recombination reactions for the panel of eight Safe Site strains. For each strain, we independently transformed (i) an oligo to inactivate *sfGFP*, (ii) an oligo to inactivate *mKate2*, or (iii) an oligo cocktail to inactivate both reporters. We also performed a control reaction in each case using water in place of oligo. For Safe Sites that lie on opposite sides of the replication fork, we designed appropriate oligos to ensure recombination targeting the lagging strand. We found that recombination rates were consistently high across the chromosome with Safe Sites displaying, on average, single cycle, single site conversion rates of 17.0 ± 6.7% and 19.7 ± 5.7% for *sfGFP* and *mKate*2, respectively (**Figure 3-6C**). We also report single cycle, multiplex conversion rates (averaged across Safe Sites) of 7.5 ± 4.4% for the *sfGFP*, 7.9 ± 2.9% for the *mKate2*, and 6.3 ± 2.3% for both reporters when transformed with the dual oligo cocktail.

3.3.5. Analysis of transient hypermethylation effects on mutagenesis

To investigate the effect of BioDesignER modifications on global mutation rate, we developed a mutagenesis detection assay that focuses on a single codon. The mutagenesis cassette utilizes chloramphenicol acetyltransferase (*cat*) gene translationally fused to green fluorescence gene *mNeon* (**Figure 3-9A**). This strategy should allow estimation of mutation rate without mutant fitness biases and second-site suppressor mutations observed in traditional fluctuation analyses such as rifampicin resistance[105]. Following integration of the mutagenesis cassette at Safe Site 1, we introduced a TAA stop codon (ochre) at Lys19 of *cat* via a single nucleotide mutation. In principle, only mutations or suppression of the ochre codon generates chloramphenicol resistant, green fluorescent colonies. Across several fluctuation tests, we sequenced 54 chloramphenicol-resistant clones from distinct mutational events. We observed eight unique genotypes at the ochre codon arising from spontaneous mutations (**Figure 3-10A**). The absence of ochre codons suggests a minimal contribution of ochre suppressor mutants in the fluctuation test.

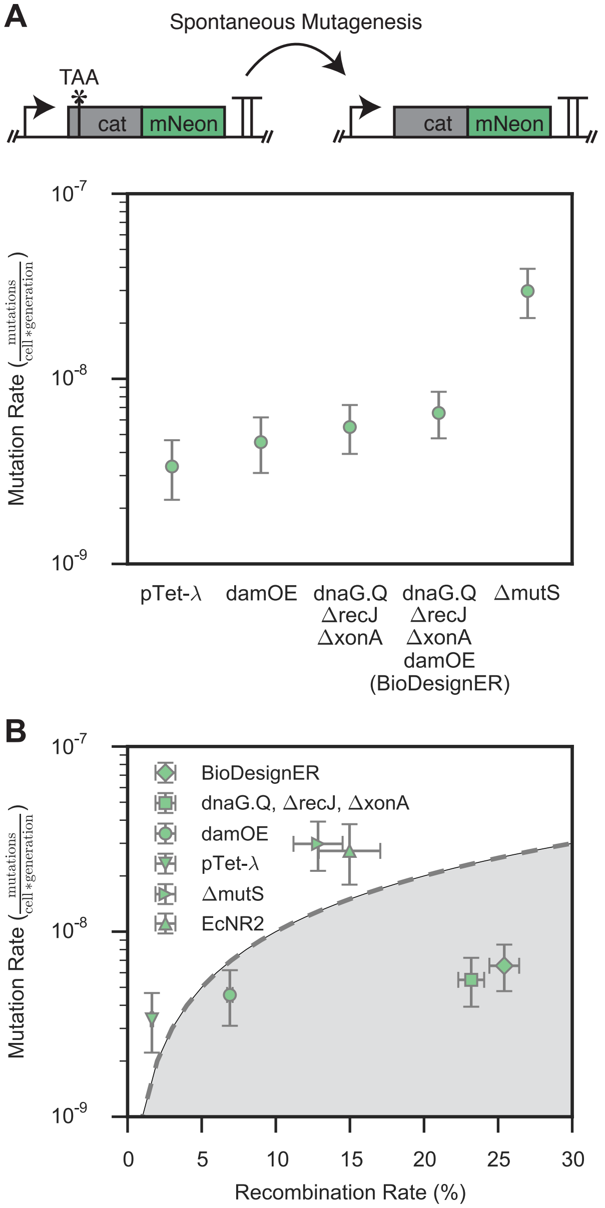


Figure 3-9. Comparative mutational analysis of BioDesignER.

(**A**) Background mutation rates (mutations/cell/generation) as measured via a cat-mNeon fluctuation assay for various stages of BioDesignER strain construction compared to an MMR-deficient (ΔmutS) strain derived from pTet-λ. Error bars represent 95% CI. (**B**) Single-cycle ssDNA recombination efficiency plotted against background mutation rate for each strain to show tradeoffs between recombination and mutation rates. The resulting tradeoff space represents the unit increase in mutation rate observed for a unit increase in recombination rate and is divided by y = β*x, where β = 10−9 is a characteristic scaling factor for the mutation rate. X-error bars represent ± 1 standard deviation and Y-error bars represent 95% CI.

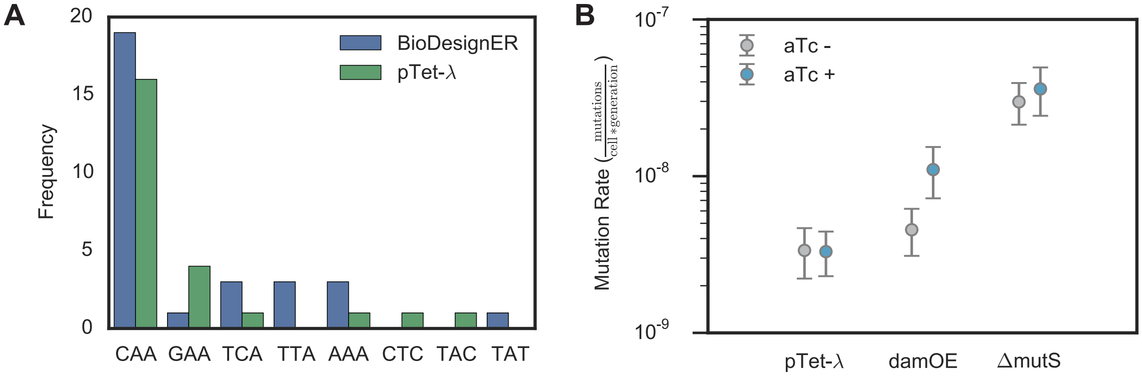


Figure 3-10. Mutational analysis.

(**A**) Stop codon (TAA) reversion genotype distribution of 54 randomly-sequenced colonies from mutational analysis of pTet-λ and BioDesignER. (**B**) Effect of aTc induction on background mutation rates (mutations/cell/generation) as measured via a cat-mNeon fluctuation assay for pTet-λ, damOE, and an MMR-deficient (ΔmutS) strain derived from the pTet-λ strain. Error bars represent 95% confidence intervals.

Using this assay, we benchmarked mutation rates of BioDesignER against (i) strains in the BioDesignER construction lineage, (ii) EcNR2 (reference), and (iii) MMR-deficient (control) strains pTet-λ *ΔmutS* and damOE. All assayed strains utilized the inactivated *cat-mNeon* cassette at Safe Site 1 as shown in **Figure 3-9A**. To allow EcNR2 to be compatible with the *cat-mNeon* fluctuation assay, we replaced the *cmR* selection cassette native to EcNR2 with *kanR*. Under comparable growth conditions, we estimated mutation rates of 3.36×10^{-9} (95% confidence interval (CI): 2.22–4.66×10^{-9}), 4.55×10^{-9} (CI: 3.10–6.19×10^{-9}), and 6.54×10^{-9} (CI: 4.77–8.51×10^{-9}) mutations/cell/generation for pTet-λ, damOE, and BioDesignER, respectively. By comparison, we observed mutation rates of 2.98×10^{-8} (CI: 2.13–3.93×10^{-8}) for the control pTet-λ *ΔmutS*, which was similar to the rate of 2.73×10^{-8} (CI: 1.79–3.81×10^{-8}) observed for EcNR2. For all strains assayed, all chloramphenicol-resistant colonies were also fluorescent.

To investigate the effect of λ-Red induction on global mutation rates and compare the mutagenic effect of *dam* over-expression to deletion of *mutS*, we tested the mutation rates for pTet-λ, damOE, and pTet-λ *ΔmutS* both with and without aTc induction (**Figure 3-10B**). We found no effect on global mutation rates due to aTc induction (i.e. expression of the λ-Red machinery) in pTet-λ and pTet-λ *ΔmutS*. Consistent with prior work[90], we observed an increase in mutation rate for damOE under aTc induction - specifically, 2.4-

fold in this work. Finally, we noted that even with aTc induction damOE was still less mutagenic than pTet-λ *ΔmutS*, suggesting that BioDesignER uniquely strikes a balance between on-target and off-target mutagenesis rates.

To quantify this balance, we compared the recombination and mutagenesis rates for a selection of control strains and BioDesignER (**Figure 3-9B**). The resulting trade-off space can be divided into two regimes where strains falling in the shaded region exhibit a favorable trade-off between recombination rate and mutation rate. BioDesignER falls in the favorable subspace, while MMR-deficient strains such as EcNR2 and pTet-λ Δ*mutS* fall in the unfavorable regime above the tradeoff line. To summarize this result, we introduce the metric recombineering fidelity, which we define as the product of fold-increase in recombination rate and fold-decrease in mutagenesis rate, each relative to EcNR2. Using this metric, we calculate that BioDesignER exhibits 7.3-fold greater recombineering fidelity than EcNR2 (1.75-fold improvement in recombination rate and 4.17-fold decrease in mutagenesis rate) (**Table 3-4**).

Table 3-4. Comparison of recombineering fidelity of relevant strains.

Strain	Recombination efficiency (%)	Mutation rate (mutations/cell/generation)	Recombineering fidelity	Temperature (°C)
pTet-λ	1.6 ± 0.1	3.36×10^{-9}	0.9	37
EcNR2	14.5 ± 2.1	2.73×10^{-8}	1.0	30
BioDesignER	25.4 ± 1.0	6.54×10^{-9}	7.3	37

3.3.6. Multi-cycle recombineering rate enhancements

High single-cycle editing efficiency enables the rapid generation of genotypically diverse populations using multiplexed, cyclical recombineering workflows. To assess how well BioDesignER could generate a population with multiplex edits, we transformed a starting population with an oligo cocktail targeting multiple sites and tracked phenotypic diversity as a function of recombineering cycle for multiple strains. Specifically, we transformed BioDesignER harboring the *sfGFP-mKate2* fluorescence cassette with oligos to inactivate both reporters over four sequential recombineering cycles. In parallel, we compared BioDesignER to pTet-λ (**Figure 3-11A**), EcNR2, and EcNR2.nuc5- (**Figure 3-12**) transformed with the same cocktail. BioDesignER exhibited high multiplex editing efficiency with nearly 60% of the population incorporating both edits (58.8 ± 3.5%) by the fourth recombineering cycle (**Figure 3-11A**), thus outperforming EcNR2 (15.9 ± 3.0%) and showing similar efficiency to EcNR2.nuc5- (54.3 ± 5.6%) (**Figure 3-11B**).

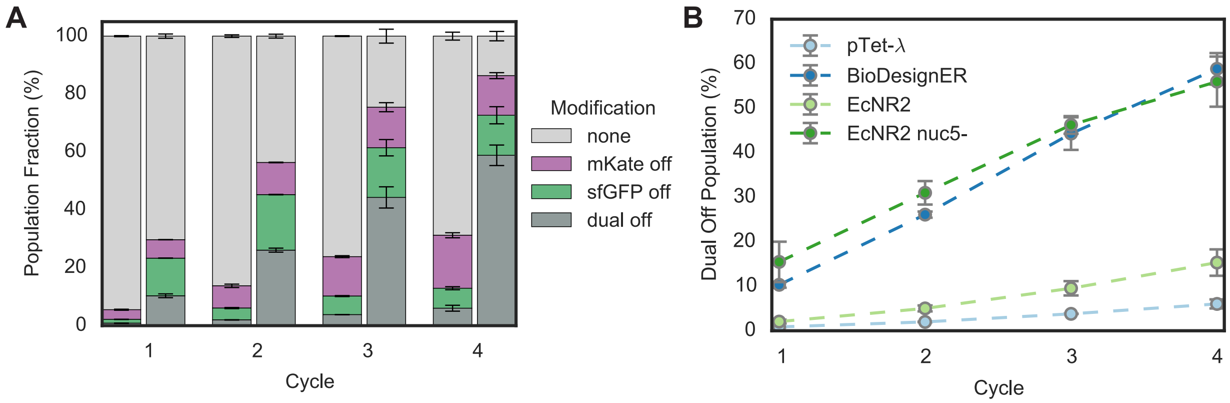


Figure 3-11. Comparative multi-cycle, multiplexed recombineering of BioDesignER.

(**A**) The fraction of each genotype (i.e. modification type) was measured via flow cytometry for pTet-λ (left) and BioDesignER (right) after each cycle of recombineering. Errors bars represent ± 1 standard deviation. (**B**) The fraction of each strain population in which both markers were edited (dual off genotype) is shown across all four recombineering cycles. Errors bars represent ± 1 standard deviation.

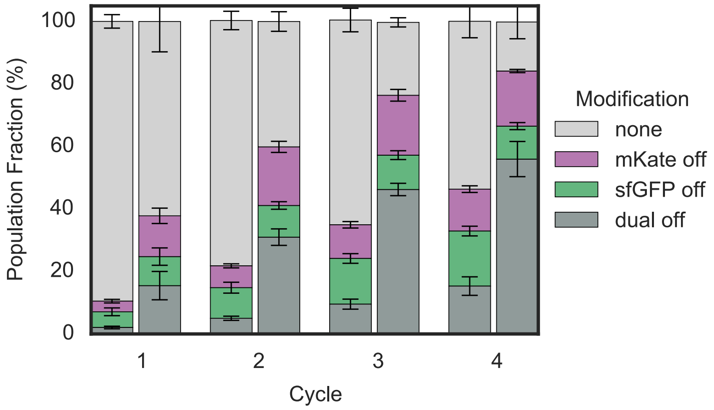


Figure 3-12. Multi-cycle recombineering results for EcNR2 and EcNR2.nuc5-.

The fraction of each genotype (i.e. modification type) measured via flow cytometry for EcNR2 (left) and EcNR2.nuc5- (right) strains after each cycle of recombineering. Errors bars represent ± 1 standard deviation.

Given the higher single-cycle conversion rate of EcNR2.nuc5- compared to BioDesignER (**Figure 3-3C**, right panel), we were surprised by the comparable performance of the two strains over multiple recombineering cycles. We partly attribute this parity to uncharacteristically low and sporadic single-cycle efficiencies that we repeatedly observed for EcNR2.nuc5- replicates (**Figure 3-13**). Regardless, while both BioDesignER and EcNR2.nuc5- exhibited similar multiplex editing efficiencies, EcNR2.nuc5- requires culturing at 30–32°C and is a persistent mutator, which increases recombineering cycle time and basal mutation rate, respectively - thus limiting its overall utility as a reliable strain for multiplex genome editing.

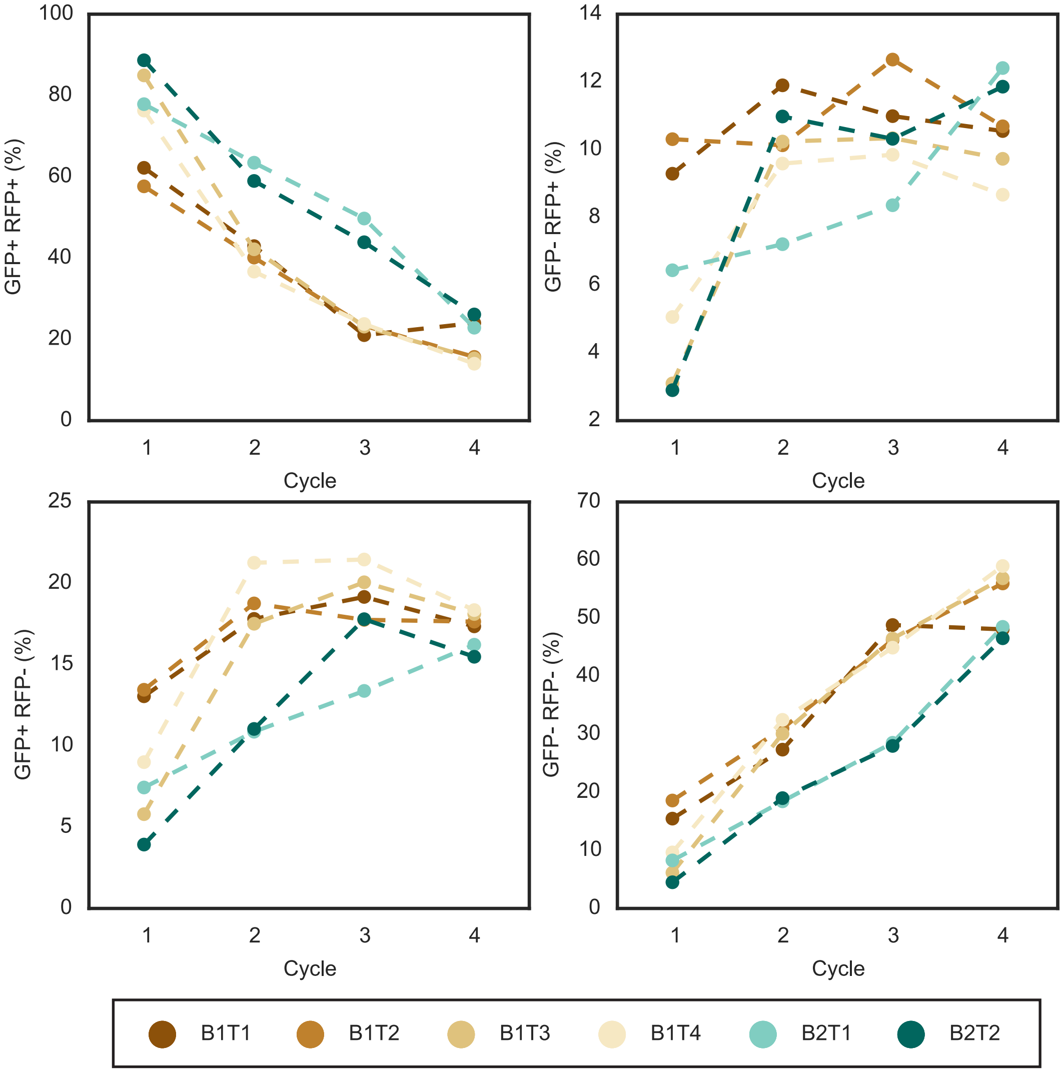


Figure 3-13. Individual traces for multi-cycle recombineering of EcNR2.nuc5-.

The fraction of each resulting strain population (e.g. GFP+ RFP+) for the dual off modification shown across all four recombineering cycles for technical replicates (denoted by T - eg T1, T2) and for two independent EcNR2.nuc5- strain isolates (B1, B2).

3.4. Discussion

High-efficiency genome engineering in bacteria enables breadth[81] and depth[20] explorations of genotypic diversity to enhance engineered behaviors. However, to date, no platform strain exists that incorporates a suite of core functions to provide efficient recombineering and regulate both genome engineering functions and cellular programs. BioDesignER is a high-fidelity recombineering strain constructed to rapidly explore and optimize engineered functions. It incorporates many genomic modifications that increase recombination efficiency and reduce cycle time for recombineering workflows while minimizing off-target mutations. BioDesignER includes four independent inducible regulators to control recombineering and accommodate additional user designs. We have quantified the gene expression and targeted mutagenesis characteristics of eight Safe Site integration loci distributed across the genome and found seven Safe Sites suitable for engineering purposes.

BioDesignER enables rapid selection-based recombineering workflows with no requirements for plasmid transformation or curing. Reliable engineering of sequential genome integrations with established recombineering approaches, such as the use of plasmids pSIM5[74] or pKD20[72], require transformation and curing procedures of plasmid-encoded recombineering functions for each integration stage. These requirements increase the time required for individual genome editing steps by multiple days. Anecdotally, we have found plasmid-based recombineering systems unreliable for conducting multiple editing cycles from a single transformation of the recombineering plasmid. We speculate that the failure to achieve multi-cycle genome editing from plasmid-based recombineering solutions may be related to the accumulation of mutations spurred by the maintenance or leaky expression of λ-Red genes over many generations. In contrast, we have completed all of the scar-free DNA recombineering workflows reported here with no restoration or replacement of the minimized pTet λ-Red cassette.

We have increased recombineering fidelity in BioDesignER by striking a balance between recombination efficiency and mutagenesis rates. A high recombineering fidelity platform such as this may provide new avenues to multiplex genome remodeling using CRISPR-Cas9 techniques. CRISPR-Cas9 genome editing approaches in bacteria are limited by recombination efficiency to rescue double-stranded breaks. Linking CRISPR-Cas9 counterselection of native sequences with high-efficiency, multi-site recombineering may allow concurrent selection of many modifications from a large bacterial population with little off-target activity, thereby enabling researchers to explore unprecedented genetic diversity.

While BioDesignER exhibits robust functionalities with respect to recombineering fidelity, comparing the recombination efficiency of the BioDesignER lineage to EcNR2-derived strains reveals inconsistent results related to culture temperatures. Specifically, we found a nearly 2-fold reduction in recombination efficiency for BioDesignER at 30°C compared to 37°C, resulting in recombineering efficiencies similar to EcNR2 (Figure 2C). This reduction suggests some uncharacterized dependence of recombination efficiency on temperature and could reflect reduced ssDNA access to the replication fork, lower ssDNA half-life at reduced temperatures, or perhaps temperature-dependent expression of the λ-Red machinery from pTet.

While constructing BioDesignER, we developed multiple selection/counter-selection strategies that may be of general use for bacterial genome engineering. These strategies combine selection/counter-selection and fluorescence screening components to accelerate scar-free genome engineering. Specifically, the genetic cassettes utilize selection/counter-selection of *thyA*, building on work from FRUIT[106]. This approach requires two recombineering transformations: a dsDNA integration of the fluorescence-coupled *thyA* cassette at the target genomic locus followed by removal of the cassette using ssDNA or dsDNA. The genetic modification of interest can be incorporated at either integration stage. In comparison, CRISPR-based genome editing workflows, which are gaining popularity, require multiple steps including guide plasmid construction, co-transformation with Cas9, and subsequent curing. Thus, the selection/counter selection methodologies developed here allow a simple and effective approach to genome engineering. Finally, as a resource to the bioengineering community we have generated a variant of BioDesignER (RE1000) with no antibiotic resistance genes (*ΔampR*).

Development of BioDesignER points to genome design strategies for next-generation biotechnology hosts. As synthetic biology matures, the application space is expanding beyond prototypical genetic circuits and metabolic pathways in laboratory environments to robust engineered functions in ecologies with high biotic and abiotic complexity, including soil, wastewater, and the human gut[1]. Efficient and sustained activity of engineered functions in these environments will require programmed behaviors to be optimized in phylogenetically diverse microbes. We anticipate the integrative approach used to develop and characterize BioDesignER can be a template to develop high-efficiency recombineering platforms for new bacterial hosts.

3.5. Materials and Methods

3.5.1. Chemicals, reagents, and media

LB Lennox Medium (10 g/l Tryptone, 5 g/l Yeast Extract, 5 g/l NaCl; Sigma-Aldrich, USA) was used to culture strains for experiments, to prepare electrocompetent cells for recombineering, and as recovery broth following electroporation. Antibiotics concentrations used were 34 μg/ml for chloramphenicol, 100 μg/ml for carbenicillin, and 50 μg/ml for kanamycin. Anhydrotetracycline (aTc) (CAS 13803-65-1; Sigma-Aldrich, USA) was used at 100 ng/ml to induce the λ-Red genes for recombineering. For *thyA*-mediated recombineering steps, M9 minimal media supplemented with 0.4% glucose, 0.2% casamino acids, thymine (100 μg/ml), and trimethoprim (50 μg/ml) was used. M9 minimal media with valine (20 μg/ml) was used to select for the *ilvG*$^+$ genotype. All M9 minimal media was supplemented with biotin at 10 μg/ml to account for the biotin auxotrophy common to all EcNR1-derivative strains.

3.5.2. Oligonucleotides

Oligos were ordered from Integrated DNA Technologies, resuspended in 1× TE buffer at either 500 uM (recombineering oligos) or 100 uM (standard amplification oligos), and stored at −20°C. For recombineering workflows, oligos were designed to target the lagging strand of DNA replication and contain at least 35 bp of homology to the target locus. Oligos for testing recombination efficiency were ordered with 5′ phosphorothioate base modifications. Oligo sequences for individual BioDesignER lineage construction steps are available in **Table 3-5**.

Table 3-5. Useful oligonucleotides for BioDesignER.

ID	Name	Sequence	Function
oRE021	SS1_fwd	ATGATGTTGTCAAAGAGTATGCGTCG	amplify Safe Site 1 from 5 'end
oRE022	SS1_rev	CCTGGTTGCTGGATAAGATGCTACAAC	amplify Sate Site 1 from 3' end
oRE023	SS2_fwd	CAGTAGTTTGTTTAAACCACAGCACAGAAAAAATC	amplify Safe Site 2 from 5 'end
oRE024	SS2_rev	TGCATGGGTGAAAGAAAATCTGGAC	amplify Sate Site 2 from 3' end
oRE025	SS3_fwd	GCACAAAAACGACCCCGTAATATACG	amplify Safe Site 3 from 5 'end
oRE026	SS3_rev	TTTTTAATTCGCTCTGAAACTGATGGC	amplify Sate Site 3 from 3' end

oRE027	SS4_fwd	TAAAAACACCCGATAGCGAAAGTTATCGG	amplify Safe Site 4 from 5 'end
oRE028	SS4_rev	GCTGACGCTGGTGAAAAGTTTTGAC	amplify Sate Site 4 from 3' end
oRE029	SS5_fwd	TAACGAAAAAAAGCGGAAGAGGTCG	amplify Safe Site 5 from 5 'end
oRE030	SS5_rev	CAGCGCGACGTTATCGCTCG	amplify Sate Site 5 from 3' end
oRE031	SS6_fwd	GAGAAAACGAAGTAAAAGGATATCCGGC	amplify Safe Site 6 from 5 'end
oRE032	SS6_rev	TACGCTCTATCGCAAGATTAAGAAATATGAGTTG	amplify Sate Site 6 from 3' end
oRE033	SS7_fwd	GCATCGTTTCCAGCGGTGAAG	amplify Safe Site 7 from 5 'end
oRE034	SS7_rev	ATCTATTGATGGAATGCCGGGTACG	amplify Sate Site 7 from 3' end
oRE035	SS8_fwd	ATTTTGTAGACCGGATAAGGAATTCACGC	amplify Safe Site 8 from 5 'end
oRE036	SS8_rev	CATTGCCGCCAGCTACGG	amplify Sate Site 8 from 3' end
oRE819	sfGFP-off-90_thio	T*G*AAGGTGACGCAACTAATGGTAAACTGACGCTGAAGTTCATCTGaACTACTGGTAAACTGCCGGTACCTTGGCCGACTCTGGTAACGAC	incorporate premature stop codon into sfGFP (replichore 1)
oRE820	mKate-off-90_thio	T*T*CACCCTCAGAGGTGCATTTGAAGTGGTGGTTGTTAACGGTGCCTTaCATGTACAGCTTCATGTGCATGTTTTCCTTAATCAGTTCAGA	incorporate premature stop codon into mKate (replichore 1)
oRE877	sfGFP-off-90_thio_rev	G*T*CGTTACCAGAGTCGGCCAAGGTACCGGCAGTTTACCAGTAGTtCAGATGAACTTCAGCGTCAGTTTACCATTAGTTGCGTCACCTTCA	incorporate premature stop codon into sfGFP (replichore 2)
oRE878	mKate-off-90_thio_rev	T*C*TGAACTGATTAAGGAAAACATGCACATGAAGCTGTACATGtAAGGCACCGTTAACAACCACCACTTCAAATGCACCTCTGAGGGTGAA	incorporate premature stop codon into sfGFP (replichore 2)
oRE406	recJ_off	GGAGGCAATTCAGCGGGCAAGTCTGCCGTTTCATCGACTTCACGTCACGACGAAGTTGTATCTGTTGTTTCACGCGAATTATTTACCGCT	remove RT2P cassette from *recJ* deletion intermediate
oRE571	RT2P_recJ_fwd	agcggtaaataattcgcgtgaaacaacagatacaacttcgtTgAGCAATAGTAAGACAACACGCAAAGTC	amplify RT2P cassette with 5' overhangs for targeting *recJ*
oRE572	RT2P_recJ_rev	caattcagcgggcaagtctgccgtttcatcgacttcacgtGGACCAAAACGAAAAAAGGC	amplify RT2P cassette with 3' overhangs for targeting *recJ*
oBA285	CmR_OFF_R	ATAGGTACATTGAGCAACTGACTGAAATGCCTCAAAATGTTCTTAACGATGCCATTGGGATATATCAACGGTGGTATA TCCAGTGATTTT	incorporate premature stop codon into cmr-mNeon cassette

3.5.3. Strains

A key set of strains used in this work is listed in **Table 3-1**. **Table 3-6** provides an abbreviated summary of strain identification numbers and genotypes for the BioDesignER lineage. Strains used to quantify recombineering and mutagenesis rates are listed in **Table 3-7**.

Table 3-6. Relevant strain genotypes in BioDesignER construction lineage.

Strain	Genotype
EcNR1	MG1655 λ-Red(pN25:*tetR*, *ampR*)::*bioA*/*bioB*
RE065	MG1655 λ-Red(pN25:*tetR*, *ampR*)::*bioA*/*bioB ilvG*$^{+}$
RE089	MG1655 λ-Red(*gfp-kanR* / *mRFP1*::λ-int, pN25:*tetR*, *ampR*)::*bioA*/*bioB ilvG*$^{+}$

RE095	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA*
RE097	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, *thyA*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA*
RE111	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, *dam*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA*
RE117	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, *dam*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *thyA*::SS7
RE119	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, *dam*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7
RE123	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, *dam*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *thyA*::SS1
RE130	MG1655 λ-Red(*gfpC48*-kanR*/*mRFP1*::λ-int, pTet::{λ-*kil*,λ-*gam*}, *dam*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *thyA*::SS1
RE151	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, pTet::{λ-*kil*,λ-*gam*}, *dam*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *thyA*::SS1
RE173	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, pTet::{λ-*kil*,λ-*gam*}, *dam*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1
RE270	MG1655 λ-Red(*gfp-kanR*/*mRFP1*::λ-int, *thyA*::pTet-λ, *dam*::λ-term, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1
RE335	MG1655 λ-Red(*gfp-kanR*/*mRFP1*, pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1
RE369	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1
HR113	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A-*sfGFP-thyA*::*dnaG*
HR114	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A
HR115	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1}-*sfGFP-thyA*::*lacI*
HR117	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1}
HR118	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} *sfGFP-thyA*::*araE*
HR119	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE*
HR122	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* *sfGFP-thyA*::*araC*
HR123	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC*
HR128	MG1655 λ-Red(*sfGFP-kanR*, pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC*
HR132	MG1655 λ-Red(*sfGFP-kanR*, pTet2:*bet-exo-gam-dam*(fs), pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* RT2P::*recJ*
HR139	MG1655 λ-Red(*sfGFP-kanR*, pTet2:*bet-exo-gam-dam*(fs), pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ*
HR143	MG1655 λ-Red(*sfGFP-kanR*, pTet2:*bet-exo-gam-dam*(fs), pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ* RT2P*::xonA*
HR145	MG1655 λ-Red(*sfGFP-kanR*, pTet2:*bet-exo-gam-dam*(fs), pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ* *ΔxonA*
RE603	MG1655 λ-Red(*sfGFP-kanR*, pTet2:*bet-exo-gam-dam*(fs)-RT2P, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ* *ΔxonA*
RE609	MG1655 λ-Red(*sfGFP-kanR*, pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ* *ΔxonA*
RE611	MG1655 λ-Red(*sfGFP-cmR*/*mKate2-kanR*, pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *ΔthyA* *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ* *ΔxonA*
RE613	MG1655 λ-Red(*sfGFP-cmR*/*mKate2-kanR*, pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ* *ΔxonA*
RE625	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *cymR*::SS7 *spoIIID-mKate2*::SS1 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ* *ΔxonA*
RE630	MG1655 λ-Red(pTet2:*bet-exo-gam-dam*, pN25:*tetR*, *ampR*)::*bioA*/*bioB* *ilvG*$^+$ *cymR*::SS7 *dnaG*.Q576A *lacI*^{Q1} Pcp8-*araE* Δ*araBAD* pConst-*araC* Δ*recJ* *ΔxonA*

Table 3-7. Relevant strains for characterizing recombination and mutagenesis rates.

Strain	Host strain	Host strain ID	Integrated cassette	Usage	Host genotype
RE871	EcNR2	RE869	sfGFP-cmR+mKate2::SS1	Recombineering rate estimation	MG1655 λ-Red(*ampR*)::*bioA/bioB* Δ*mutS*::*kanR*
RE875	EcNR2.nuc5-	RE867	sfGFP-cmR+mKate2::SS1	Recombineering rate estimation	MG1655 λ-Red(*ampR*)::*bioA/bioB dnaG.Q576A* Δ*recJ* Δ*xonA* Δ*xseA* Δ*exoX* Δ*red-α* Δ*mutS*::*kanR*
RE811	pTet-λ	RE574	sfGFP-cmR+mKate2::SS1	Recombineering rate estimation	MG1655 pTet2-*gam-bet-exo/tetR/ampR*::*bioA/B ilvG+*
RE849	damOE	RE824	sfGFP-cmR+mKate2::SS1	Recombineering rate estimation	MG1655 pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+*
RE851	dnaG.Q	RE626	sfGFP-cmR+mKate2::SS1	Recombineering rate estimation	pTet2-*gam-bet-exo/tetR/ampR*::*bioA/B ilvG+ dnaG*.Q576A
RE853	exo1	RE628	sfGFP-cmR+mKate2::SS1	Recombineering rate estimation	pTet2-*gam-bet-exo/tetR/ampR*::*bioA/B ilvG+ dnaG*.Q576A Δ*recJ*
RE813	exo2	HR146	sfGFP-cmR+mKate2::SS1	Recombineering rate estimation	pTet2-*gam-bet-exo/tetR/ampR*::*bioA/B ilvG+ dnaG*.Q576A Δ*recJ* Δ*xonA*
RE636	BioDesignER	RE630	sfGFP-cmR+mKate2::SS1	Recombineering rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
RE638	BioDesignER	RE630	sfGFP-cmR+mKate2::SS2	Recombineering rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
RE640	BioDesignER	RE630	sfGFP-cmR+mKate2::SS3	Recombineering rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
RE642	BioDesignER	RE630	sfGFP-cmR+mKate2::SS4	Recombineering rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
RE644	BioDesignER	RE630	sfGFP-cmR+mKate2::SS5	Recombineering rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
RE646	BioDesignER	RE630	sfGFP-cmR+mKate2::SS6	Recombineering rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
RE648	BioDesignER	RE630	sfGFP-cmR+mKate2::SS7	Recombineering rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
RE650	BioDesignER	RE630	sfGFP-cmR+mKate2::SS8	Recombineering rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
BA545	pTet-λ	RE574	cmR-mNeon::SS1	Mutatgenesis rate estimation	MG1655 pTet2-*gam-bet-exo/tetR/ampR*::*bioA/B ilvG+*
BA549	damOE	RE824	cmR-mNeon::SS1	Mutatgenesis rate estimation	MG1655 pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+*
RE1192	exo2	HR146	cmR-mNeon::SS1	Mutatgenesis rate estimation	pTet2-*gam-bet-exo/tetR/ampR*::*bioA/B ilvG+ dnaG*.Q576A Δ*recJ* Δ*xonA*
BA543	BioDesignER	RE630	cmR-mNeon::SS1	Mutatgenesis rate estimation	pTet2-*gam-bet-exo-dam/tetR/ampR*::*bioA/B ilvG+ lacIQ1* Pcp8-*araE* Δ*araBAD* pConst-*araC dnaG*.Q576A Δ*recJ* Δ*xonA*
BA547	pTet-λ Δ*mutS*	RE815	cmR-mNeon::SS1	Mutatgenesis rate estimation	MG1655 pTet2-*gam-bet-exo/tetR/ampR*::*bioA/B ilvG+* Δ*mutS*::*kanR*
RE1171	EcNR2	RE869	cmR-mNeon::SS1	Mutatgenesis rate estimation	MG1655 λ-Red(*ampR*)::*bioA/bioB* Δ*mutS*::*kanR*

3.5.4. Growth rate measurements

Two clones of each strain were cultured overnight in LB Lennox (LB) medium with chloramphenicol. The following morning each strain culture was back-diluted 1:100 into two media types: (i) LB with aTc (LB+aTc) (ii) LB. The resulting inocula were divided into four technical replicates and then grown for up to 18 h in a Biotek Synergy 2 microplate reader. The growth rate at early exponential phase was calculated from the resulting optical density data using custom analysis scripts in Python.

3.5.5. Competent cell preparation and recombineering

Strains were grown overnight in LB Lennox medium (LB) with antibiotics as appropriate at 37°C. The following morning each strain culture was back-diluted 1:100 into 25 ml LB+aTc and grown at 37°C until they reached OD600 0.3–0.4. The resulting mid-log cultures were chilled in a 4°C ice-water bath. Cultures were centrifuged (Beckman-Coulter Allegra 25R) at 8000 × g and subjected to two washes: (i) 25 ml chilled water (ii) 15 ml chilled 10% glycerol. The cell pellets after the final glycerol wash were resuspended in 10% glycerol, yielding ~500 ul of competent cells given the residual cell mass from the wash.

Due to their different induction and growth requirements, EcNR2 and EcNR2.nuc5- strains were grown overnight at 30°C, back-diluted 1:100 into 25 ml LB+chlor media, and cultured at 30°C until they reached OD600 0.3–0.6. The λ-Red machinery was induced by incubating the cultures in a 42°C water bath for 15 min after which the strains were chilled in a 4°C ice-water bath for at least 10 min. The remainder of the preparation for EcNR2 and EcNR2.nuc5- follows the same aforementioned wash steps.

A total of 40 ul of competent cells were used for each recombineering reaction. Oligos were diluted to 50 μM concentration in 10% glycerol and 10 μl of the diluted oligo was added to the competent cell mixture. For water control reactions, 10 μl of water was added. For multiplexed reactions, 10 μl of a cocktail with a total oligo concentration of 50 μM was used. The resulting cell-oligo mix was transferred to a chilled cuvette (1 mm gap, VWR) and electroporated using a BTX™-Harvard Apparatus ECM™ 630 Exponential Decay Wave Electroporator with the following parameters: voltage (1800 V), resistance (200 Ω), and capacitance (25 μF).

3.5.6. Fluorescence-coupled scar-free selection/counter-selection

Working from the *ΔthyA* strain RE095 and derivative strains (**Table 3-6**), a dsDNA *thyA* cassette with or without a fluorescence gene (**Figure 3-2**) was amplified with 35–50 bp flanking homology to a target genomic locus and integrated via standard recombineering as described above, with the exception that cells were made competent by growing in LB supplemented with thymine (100 μg/ml) and trimethoprim (50 μg/ml). Integrants of *thyA* were selected for on LB media. Colonies with fluorescence-coupled *thyA* cassettes were screened visually for fluorescent phenotypes on a blue-light transilluminator. Proper insertion of the cassette was confirmed by locus-specific colony polymerase chain reaction (PCR). Replacement of the *thyA* cassette was performed through recombineering with a ssDNA or dsDNA cassette and selected for on M9 agar plates supplemented with thymine (100 µg/ml), trimethoprim (50 µg/ml), and casamino acids (0.2%). Removal of fluorescence-coupled *thyA* cassettes was screened visually for non-fluorescent colonies via blue-light transillumination and sequences were validated via colony PCR and Sanger sequencing.

3.5.7. Recombineering efficiency and Safe Site expression measurements

Competent cells were transformed with water (control) or oligos to turn off sfGFP, mKate2, or both reporters. Following electroporation, cells were resuspended in 3 ml LB+carb. These cultures were mixed and 30 µl was transferred into an additional 3 ml LB+carb for overnight growth at 37°C (or 30°C for EcNR2 and EcNR2.nuc5-). The following morning, saturated cultures of each transformation were diluted 1:200 into phosphate-buffered saline (PBS) solution and run on a Sony SH800 cell sorter for single-cell flow cytometry analysis. At least 50,000 events were recorded for each reaction, and the fractional abundance of each reporter phenotype (GFP+ RFP+, GFP- RFP+, GFP+ RFP-, GFP- RFP-) in the population was measured. The threshold for each reporter phenotype was determined via a prior calibration in which gates for each fluorescent reporter were measured. For measurement of gene expression across Safe Sites, overnight outgrowths of control reactions from the Safe Site recombineering efficiency transformations were processed on the flow cytometer.

3.5.8. Response curves of inducible regulators

BioDesignER was transformed with plasmids containing a GFP gene regulated by each transcription factor - AraC (pBAD), CymR (pCym), or LacI (pLac) - and controlled by one of two replication origins, p15A or pSC101. Plasmids with the p15A origin contain a *kanR* marker and plasmids with the pSC101 origin use a *cmR* marker. Plasmid sequences are available via Benchling (https://benchling.com/organizations/arkinlab). Individual colonies were inoculated in LB with an appropriate antibiotic to maintain the plasmid and grown overnight. Saturated cultures were diluted 200-fold into a microtiter plate (Corning 3904) and grown at 37°C with shaking in a Biotek Synergy 2 plate reader. Kinetic growth and fluorescence measurements were taken every 5 or 10 min for 12 h. Absorbance was measured at 600 nm. GFP fluorescence was measured using 485/20 and 520/15 nm filter cubes for excitation and emission, respectively. mKate fluorescence was measured using 560/20 and 615/30 nm filter cubes for excitation and emission, respectively. Fluorescence values measured nearest OD 0.5 were used to estimate absorbance-normalized fluorescence in each channel.

3.5.9. Flow cytometry analysis of inducible regulators

Saturated cultures from the kinetic growth assays used to measure regulator inducer responses were diluted 400-fold into PBS and analyzed in a BD LSR Fortessa flow cytometer (488 nm excitation / 525/50 nm emission for GFP; 561 nm excitation / 670/30 nm emission for mKate) using an autosampler. Raw .fcs files were imported for pre-processing and subsequent analysis with custom Python scripts using the FlowCytometryTools software package (https://github.com/eyurtsev/FlowCytometryTools). For each sample, 50,000 events were captured and outliers in forward scatter and side scatter were removed using a filter with cut-offs for events outside the second and third quartile.

3.5.10. Safe Site expression analysis

Data for expression levels at each Safe Site were gathered from the flow cytometry data used to measure recombineering efficiency at each Safe Site. Data files were extracted for four recombineering conditions (no oligo control, GFP-off, mKate-off, and dual-off) and two biological replicates. The geometric mean for each fluorescence channel was calculated from filtered data. Specifically, events outside the second and third quartiles for forward and side scatter channels were removed from analysis for each .fcs data file. The dual-fluorescent subpopulation for each measurement was extracted by gating at a value that excluded non-fluorescent subpopulations but did not truncate the distribution of the dual-fluorescent subpopulation.

3.5.11. Fluctuation assay

Fluctuation tests were performed on an inactivated *cmR-mNeon* translational fusion cassette integrated at Safe Site 1. The cassette was inserted using selection on chloramphenicol and subsequently inactivated for the strains listed in **Table 3-7**. The *cmR-mNeon* cassette was first integrated as dsDNA into the respective strains via recombineering and selected for by plating on LB agar supplemented with 34 μg/ml chloramphenicol. A premature stop codon (AAA to TAA) was generated on the cassette via ssDNA recombineering with an oligo containing the stop codon mutation. The non-fluorescent population was enriched using cell sorting (Sony SH800) and non-fluorescent colonies were isolated on LB agar plates.

Prior to fluctuation tests, individual non-fluorescent colonies were grown at 30°C in LB+carb and stored at −80°C as glycerol stocks normalized to OD600 of 0.5. For the fluctuation tests, cultures were diluted 1000-fold and grown for 16 h in permissive conditions of LB+carb at 30°C (N = 24). For pTet-*λ ΔmutS*, EcNR2 and EcNR2.nuc5-, 20 μl of culture was spotted onto LB agar plates supplemented with chloramphenicol and carbenicillin. For all other strains, 100 μl volume spots were used. Viability counts were estimated for all strains by serial dilutions of six independent cultures on LB agar plates supplemented with carbenicillin. Chloramphenicol-resistant mutants were counted, and mutation rates were inferred by the MSS-MLE method[107,108]

3.5.12. Iterative recombineering cycling

Strains were prepared for transformation using the competent cell protocol described above using 25 ml of culture with a target OD600 of 0.3. Each culture was resuspended in ~500 μl of 10% glycerol after washes. Each transformation consisted of 40 μl competent cells mixed with 10 μl of 50 μM oligo mix. After transformation, cells were recovered in 3 ml LB supplemented with carbenicillin. The recovery culture was grown to saturation before beginning the next round of competent cell prep and recombination. In parallel, the recovery culture was diluted 1:60 into an additional 3 ml of LB supplemented with carbenicillin and grown to saturation prior to measurements using flow cytometry (Sony SH800).

Chapter 4. CRISPR-Cas9 Circular Permutants as Programmable Scaffolds for Genome Modification

4.1. Author Contributions

This chapter represents a manuscript with contributions from Benjamin L. Oakes (B.L.O), Christof Fellman (C.F.), Harneet S. Rishi (H.S.R.), Kian L. Taylor (K.L.T.), Shawn M. Ren (S.M.R.), Dana C. Nadler (D.C.N.), Rayka Yokoo (R.Y.), Adam P. Arkin (A.P.A.), Jennifer A. Doudna (J.A.D.), and David F. Savage (D.F.S.). Given the collaborative nature of this work, it is important to acknowledge the contributions of all authors: B.L.O., C.F., and D.F.S. conceived and designed the study. B.L.O., C.F., K.L.T., S.M.R., D.C.N., and R.Y. conducted experiments. H.S.R. conducted bioinformatics analyses for the protein engineering of Cas9-CPs. A.P.A., J.A.D., and D.F.S. supervised the research. All authors interpreted results. B.L.O., C.F., J.A.D., and D.F.S. wrote the manuscript with input from all authors.

4.2. Introduction

Type II CRISPR-Cas proteins, such as Cas9, are RNA-guided, DNA binding, and cleaving enzymes that function as integral components of adaptive bacterial immune systems[60]. Because of its intuitive and robust function, Cas9 has been adapted as a programmable DNA double-strand break generation tool in vitro and in vivo[60,109-112]. Additionally, enzymatically deactivated Cas9 (dCas9) has been shown to function as a programmable DNA-binding protein that can be harnessed to site specifically deliver additional protein domains, such as chromatin modifiers, methyltransferases, nucleobase deaminases, and fluorescent markers[25,113-118]. As a consequence, Cas9 has revolutionized genome editing and functional interrogation of the genome. Despite this potential, a number of issues still hinder the broad utility of Cas9.

Foremost, Cas9's always-on nature can lead to off-target genome editing due to prodigious activity[119]. It is, relatedly, inherently difficult to generate cell or tissue specificity in an in vivo therapeutic delivery context unless the protein can be directly delivered to a defined compartment, such as the inner ear, eye, or brain[120-122]. A lack of activity control also limits the potential of using Cas9 in post-translational genetic circuits, such as a sensor that can respond to a defined cellular input or act as a molecular recorder of cellular events[123]. Although numerous strategies now control Cas9 activity via exogenous ligands and other inputs, the ability to control Cas9 activity via an endogenous signal would be highly desirable[119,124-126].

Additionally, Cas9 did not evolve to function as a modular DNA-binding scaffold. Thus, its fusion to protein domains possessing additional activity has required elaborate optimization, such as long linkers or daisy-chained fusions[115,117,118,127,128]. Such tools may be greatly enhanced by the ability to fuse protein domains at a precise location in the Cas9 topology. Hence, developing an optimized Cas9 architecture for controlled nuclease activity and facilitating efficient construction of fusion proteins would expand and improve the future applications of this incredibly important enzyme.

One unique route for creating novel and highly functional CRISPR architectures is by protein circular permutation (CP). CP is the topological rearrangement of a protein's primary sequence, connecting its N- and C-terminus with a peptide linker, while concurrently splitting its sequence at a different position to create new, adjacent N and C termini[129] (**Figure 4-1A**). Termini repositioning can change a protein's behavior and naturally occurring or engineered circularly permuted proteins possess altered stability, substrate specificity, enzymatic rate, and novel quaternary structure[130-133]. The inherent requirement of linking the N termini and the C termini has also been exploited to create "caged" zymogen pro-enzymes in which a protease cleavage site is used as the linker sequence[134].

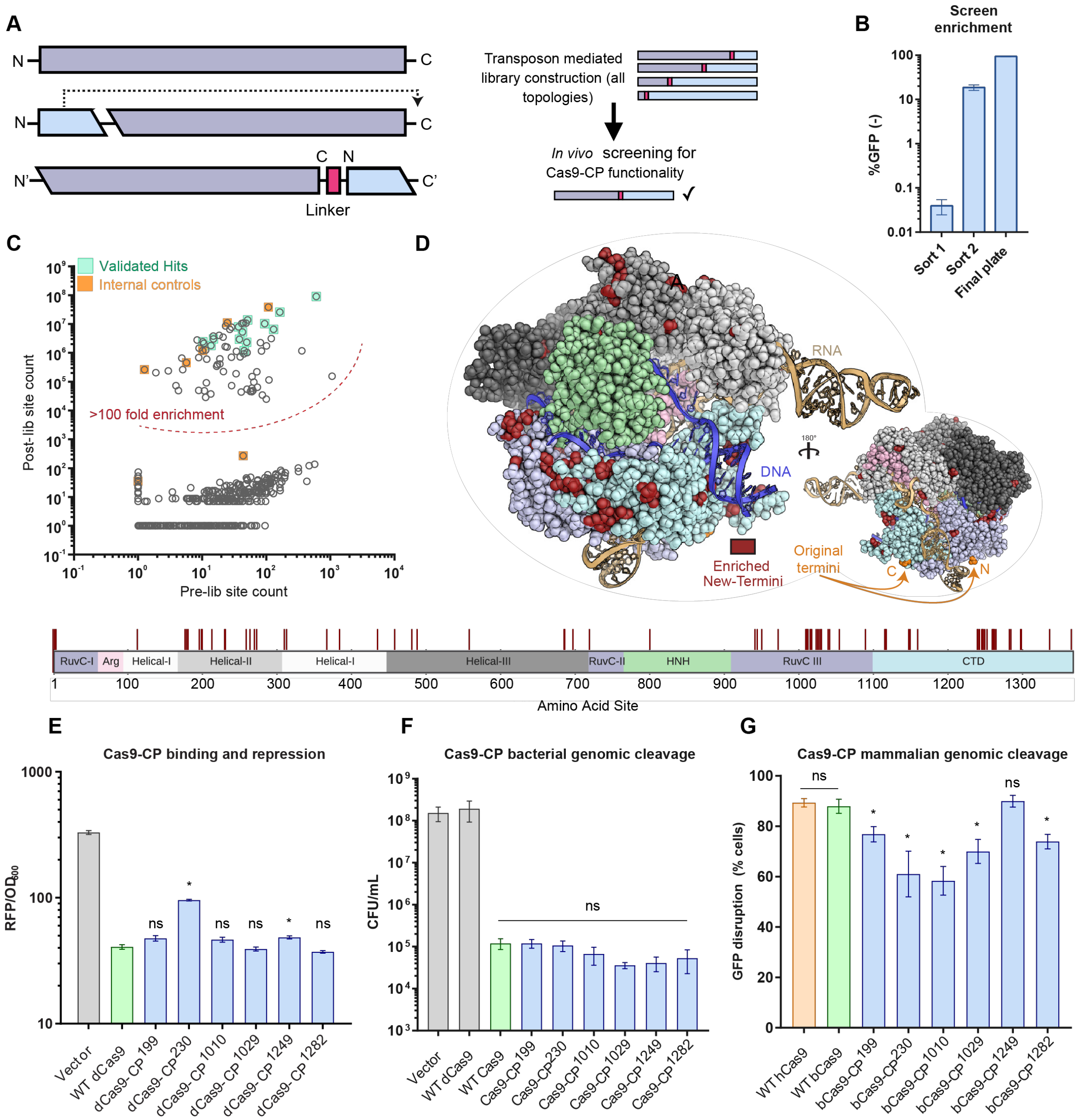


Figure 4-1. An unbiased Cas9 library screen identifies active circularly permuted Cas9 proteins.

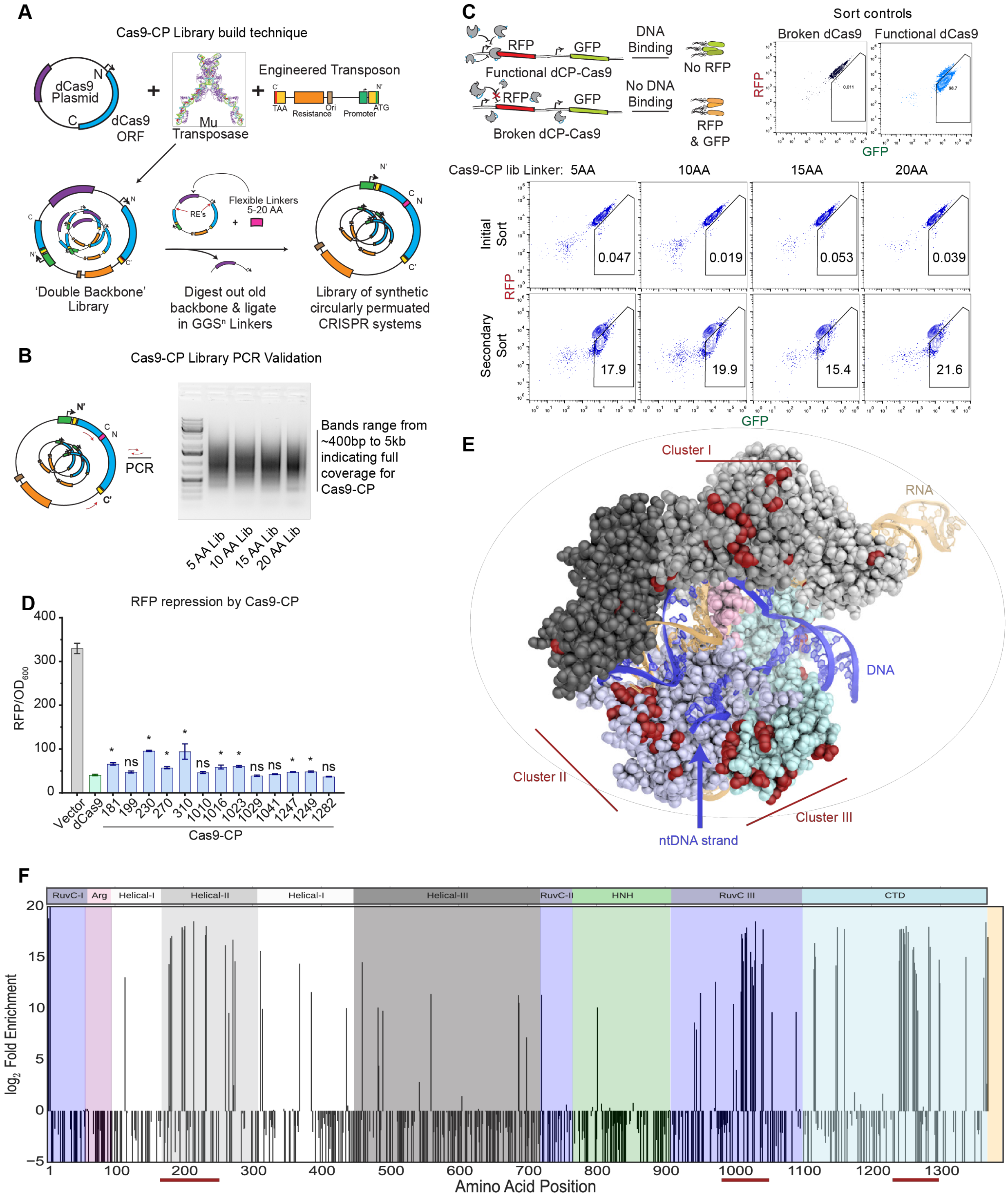
(**A**) Overview of circular permutation and library generation for Cas9. (**B**) Enrichment values of the unbiased screen as determined by flow cytometry and colony-forming units (CFU). Error bars represent standard deviation in all panels. (**C**) Deep-sequencing read averages for the pre- and post-Cas9-CP library members, demonstrating a strong clustering of highly enriched library members with internal (within 4 aa of the N and C termini) and empirically validated controls. The dotted line highlights an approximate boundary that represents >100-fold enrichment in the screen. (**D**) Model of new Cas9-CP termini (in red) based on PDB: 5F9R with domains colored according to the sequence bar (below). New termini are mapped onto the aa sequence bar. (**E**) Endpoint values for dCas9-CP 12-hr *E. coli* CRISPRi DNA binding and RFP repression system compared with WT dCas9 and a protein expression vector control in triplicate (error bars represent SD; $*p < 0.05$; ns, not significant, t test). (**F**) CFU/mL readings in an E. coli genomic cleavage assay readout by cell death compared with a protein expression vector control, WT dCas9, and WT Cas9 (n = 3, error bars represent SD; $*p < 0.05$; ns, not significant, t test). (**G**) Cleavage efficiency of a genomic reporter in mammalian cells in triplicate (described in Figures S2B and S2C), observed via indel formation, and GFP reporter disruption. hCas9 is human codon-optimized Cas9; bCas9 indicates bacterial codon-based Cas9 constructs (error bars represent SD; $*p < 0.05$; ns, not significant, t test).

Here, we demonstrate how circular permutation can be used to re-engineer the molecular sequence of Cas9 to both better control its activity and create a more optimal DNA binding scaffold for fusion proteins. By coupling systematic library creation with high-throughput fitness assays and deep sequencing, we define the set of possible Cas9 circular permutants (collectively, Cas9-CPs). Our analysis shows that Cas9 is highly malleable to circular permutation, and several regions of the protein—notably the Helical-II, RuvC, and C-terminal domain (CTD)—possess hotspots that can be opened at numerous positions to generate a diversity of Cas9-CPs. We further show that engineering of the linker sequence with site-specific protease sequences, derived from a variety of pathogenic plant and human viruses, yields “caged” pro-enzyme Cas9 variants that can be activated by proteolytic cleavage. This modular approach is generalizable and the proteins, which we term ProCas9s, are capable of sensing and stably recording or responding to the presence of numerous, distinct families of viral proteases—including those from Flaviviruses—in both E. coli and various mammalian cell types. In total, this work establishes an architecture for generating Cas9s that can be activated by cell-, tissue-, or pathway-specific endogenous signals and provides a resource of Cas9-CPs to simplify and optimize the process of constructing Cas9-fusion proteins for precision genome modification.

4.3. Results

4.3.1. Circular permutation of Cas9

To investigate the topological malleability of Streptococcus pyogenes Cas9 (hereafter Cas9), we generated a random transposon insertion library in vitro by adapting an engineered transposon from prior work[135] to contain a plasmid backbone, inducible promoter, and stop codon (**Figure 4-2A**; **Table 4-1**). As the original N and C termini of Cas9 are 40 to 60 Å apart[136], the requirements for Cas9 circular permutation are not known. We therefore permuted dCas9 using a series of linkers (GGS repeats, varying from 5 to 20 amino acids [aa]) between the original N and C termini, providing increasing steric freedom (**Figure 4-2A**). Transposition of the engineered cassette and pooled molecular cloning yielded high insertional diversity for all libraries, as indicated by the length distributions of polymerase chain reaction (PCR) amplicons (**Figure 4-2B**). Deep sequencing of the 20 aa linker library further demonstrated that ~1 of every 2 aa in Cas9 were observed transposition sites in the original pool, for a total of 661 circular permutant (CP) variants in the library (**Table 4-2**).


A
Cas9-CP Library build technique
dCas9 Plasmid
N
C
dCas9 ORF
+
Mu Transposase
+
Engineered Transposon
C'
TAA
Resistance
Ori
N'
Promoter
ATG
'Double Backbone' Library
RE's
Flexible Linkers 5-20 AA
+
Digest out old backbone & ligate in GGS^n Linkers
Library of synthetic circularly permuted CRISPR systems
B
Cas9-CP Library PCR Validation
N'
C
N
C'
PCR
Bands range from ~400bp to 5kb indicating full coverage for Cas9-CP
5 AA Lib
10 AA Lib
15 AA Lib
20 AA Lib
C
RFP
GFP
DNA Binding
No RFP
Functional dCP-Cas9
RFP
GFP
No DNA Binding
RFP & GFP
Broken dCP-Cas9
Sort controls
Broken dCas9
Functional dCas9
RFP
GFP
0.011
98.7
Cas9-CP lib Linker: 5AA
10AA
15AA
20AA
Initial Sort
Secondary Sort
RFP
GFP
0.047
0.019
0.053
0.039
17.9
19.9
15.4
21.6
D
RFP repression by Cas9-CP
RFP/OD600
400
300
200
100
0
Vector
dCas9
181
199
230
270
310
1010
1016
1023
1029
1041
1247
1249
1282
*
ns
*
*
*
ns
*
*
ns
ns
*
*
ns
Cas9-CP
E
Cluster I
RNA
DNA
Cluster II
Cluster III
ntDNA strand
F
RuvC-I
Arg
Helical-I
Helical-II
Helical-I
Helical-III
RuvC-II
HNH
RuvC III
CTD
log2 Fold Enrichment
20
15
10
5
0
−5
1
100
200
300
400
500
600
700
800
900
1000
1100
1200
1300
Amino Acid Position

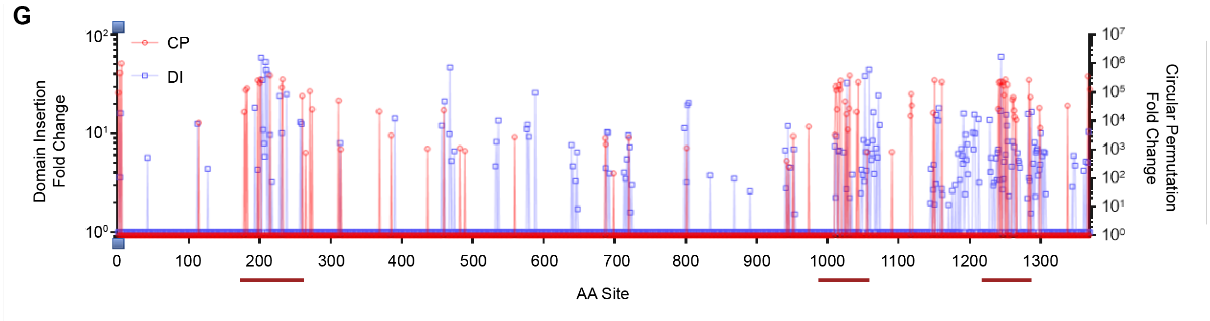


Figure 4-2. Cas9 circular permutation.

(**A**) Detailed schematic of the transposition method used to build the Cas9-CP libraries, REs = Restriction Enzyme sites. (**B**) Schematic and uncropped gel of the PCR system used to validate the creation of CP libraries. (**C**) Schematic and flow cytometry from the screen and enrichment of active Cas9-CPs in all four Cas9-CP libraries. (**D**) Endpoint values for 13 new Cas9-CPs in a 12 hr *E. coli* CRISPRi DNA binding and RFP repression system compared with WT dCas9 and a protein expression vector control (n = 3). (**E**) Alternate view of the model of new Cas9-CP termini (in red) based on PDB: 5F9R. The HNH domain has been removed to clearly demonstrate the new termini flanking either side of the non-targeting (nt) DNA strand. Inset highlights distances between various new Cas9-CP termini and R-loop. (**F**) Deep sequencing analysis and log2-fold change for new termini in the 20 aa library as mapped onto the primary sequence of Cas9. Red bars indicate clusters of CPs in specific domains. (**G**) Overlay of enrichment values for Domain Insertion (DI) [126] and CP, demonstrating clustering of events.

Table 4-1. Prokaryotic and eukaryotic vectors.

Vector type	Vector name	Key features	Length
dCas9 staging vector	pBLO5.01	pUC19_BsaI-dCas9-BsaI	6807
Transposase vector	pBLO 4.01	Enginnered Mu transposon	3451
Bacterial expression Cas9-CP vector	pBLO 9.2_bCas9-CP^{199}	TetR_tetR/tetA promoters_dCas9-CP_CM_P15A	6863
Bacterial expression Cas9-CP vector	pBLO 9.3_bCas9-CP^{230}	TetR_tetR/tetA promoters_dCas9-CP_CM_P15A	6863
Bacterial expression Cas9-CP vector	pBLO 9.6_bCas9-CP^{1010}	TetR_tetR/tetA promoters_dCas9-CP_CM_P15A	6863
Bacterial expression Cas9-CP vector	pBLO 9.9_bCas9-CP^{1029}	TetR_tetR/tetA promoters_dCas9-CP_CM_P15A	6863
Bacterial expression Cas9-CP vector	pBLO 9.15_bCas9-CP^{1249}	TetR_tetR/tetA promoters_dCas9-CP_CM_P15A	6863
Bacterial expression Cas9-CP vector	pBLO 9.16_bCas9-CP^{1282}	TetR_tetR/tetA promoters_dCas9-CP_CM_P15A	6863
ProCas9 *E. coli* vector	pBLO 36.02_ProCas9^{TEV}	TetR_tetR/tetA promoters_bCas9-CP$^{199-TEVlinker}$_2xflag_CM_P15A	6881
ProCas9 *E. coli* vector	pBLO 41.2.3_ProCas9^{3C}	TetR_tetR/tetA promoters_bCas9-CP$^{199-3Clinker}$_2xflag_CM_P15A	6881
ProCas9 *E. coli* vector	pBLO 41.2.3_ProCas9^{Poty}	TetR_tetR/tetA promoters_bCas9-CP$^{199-Potylinker}$_2xflag_CM_P22	6881

ProCas9 *E. coli* vector	pBLO 41.2.3_ProCas9^{Flavi}	TetR_tetR/tetA promoters_bCas9-CP$^{199\text{-}Flavilinker}$-2xflag_CM_P23	6881
ProCas9 mammalian cell vector	pBLO43.2_human ProCas9^{Poty}	U6-sgRNAdest_CMV-intron_hProCas9^{Poty}_T2A Mcherry_AmpR_ColE1	9263
ProCas9 mammalian cell vector	pBLO43.3_human ProCas9^{Flavi}	U6-sgRNAdest_CMV-intron_hProCas9^{Flavi}_T2A Mcherry_AmpR_ColE1	9263
ProCas9 mammalian cell vector	pBLO43.3.-6_human ProCas9$^{Flavi\text{-}6}$	U6-sgRNAdest_CMV-intron_hProCas9$^{Flavi\text{-}6}$_T2A Mcherry_AmpR_ColE1	9245
Mammalian protease expression vector	pBLO44.1_dProtease	CMV-intron_Pro_P2A-mTagBFP2_AmpR_ColE1	5368
Mammalian protease expression vector	pBLO44.3_TuMVpro	CMV-intron_Pro_P2A-mTagBFP2_AmpR_ColE1	5389
Mammalian protease expression vector	pBLO44.4_PPVpro	CMV-intron_Pro_P2A-mTagBFP2_AmpR_ColE1	5389
Mammalian protease expression vector	pBLO44.5_PVYpro	CMV-intron_Pro_P2A-mTagBFP2_AmpR_ColE1	5392
Mammalian protease expression vector	pBLO44.6_ZIKVpro	CMV-intron_Pro_P2A-mTagBFP2_AmpR_ColE1	5368
Mammalian protease expression vector	pBLO44.7_WNVpro	CMV-intron_Pro_P2A-mTagBFP2_AmpR_ColE1	5380
Lentiviral vector	pCF204	U6-sgRNA EFS-Cas9-wt-P2A-Puro	13029
Lentiviral vector	pCF704	U6-sgRNA EFS-ProCas9-Flavi-P2A-Puro	13027
Lentiviral vector	pCF711	U6-sgRNA EFS-ProCas9-FlaviS6-P2A-Puro	13009
Lentiviral vector	pCF712	U6-sgRNA EF1a-ProCas9-Flavi-P2A-Puro	13973
Lentiviral vector	pCF713	U6-sgRNA EF1a-ProCas9-FlaviS6-P2A-Puro	13955
Lentiviral vector	pCF732	U6-sgRNA EF1a-ProCas9-Flavi-P2A-Puro (without NLS)	13919
Lentiviral vector	pCF226	EFS-Cas9-wt-P2A-Puro	12649
Lentiviral vector	pCF730	EF1a-ProCas9-Flavi-P2A-Puro	13593
Lentiviral vector	pCF221	U6-sgRNA EF1a-mCherry	9841
Lentiviral vector	pCF708	EF1a-dTEV-protease-mTagBFP2	10284
Lentiviral vector	pCF709	EF1a-ZIKV-protease-mTagBFP2	10284
Lentiviral vector	pCF710	EF1a-WNV-protease-mTagBFP2	10296
Lentiviral vector	pCF736	EF1a-dTEV-protease-GFP	10296
Lentiviral vector	pCF738	EF1a-WNV-protease-GFP	10308

Table 4-2. Cas9 circular permutation screen.

Cas9-CP Amino Acid Position	New start site (AA)	Name	Pre-sort Library Merged Raw Counts	Post-sort Library Bio Rep 1 Merged Raw Counts	Post-sort Library Bio Rep 2 Merged Raw Counts	FoldChange	log2FoldChange
1	0	Cas9-CP-0	1092	7	58448	96905.608	16.5642925
3	2	Cas9-CP-2	5544	465492	1065496	466294.663	18.8308824
4	3	Cas9-CP-3	0	2	3	inf	inf
5	4	Cas9-CP-4	61	9697	26101	998722.218	19.9297239
7	6	Cas9-CP-6	1	0	0	0	-inf
8	7	Cas9-CP-7	5678	1	0	0.24828084	-2.0099551
9	8	Cas9-CP-8	1402	0	0	0	-inf
11	10	Cas9-CP-10	0	0	0		
12	11	Cas9-CP-11	4611	1	0	0.30573382	-1.709652
17	16	Cas9-CP-16	6781	1	0	0.20789539	-2.2660703
20	19	Cas9-CP-19	1	0	0	0	-inf
23	22	Cas9-CP-22	1616	0	0	0	-inf
24	23	Cas9-CP-23	4016	0	0	0	-inf
26	25	Cas9-CP-25	80	0	0	0	-inf
29	28	Cas9-CP-28	0	0	1	inf	inf

35	34	Cas9-CP-34	21978	2	2	0.29302792	-1.77089
36	35	Cas9-CP-35	9985	0	0	0	-inf
40	39	Cas9-CP-39	1	0	0	0	-inf
41	40	Cas9-CP-40	26719	4	1	0.27880159	-1.8426893
42	41	Cas9-CP-41	4522	0	0	0	-inf
43	42	Cas9-CP-42	3357	0	0	0	-inf
44	43	Cas9-CP-43	1	0	0	0	-inf
49	48	Cas9-CP-48	19837	1	1	0.16232716	-2.6230237
50	49	Cas9-CP-49	2	0	0	0	-inf
54	53	Cas9-CP-53	3445	0	2	1.05099866	0.07176083
55	54	Cas9-CP-54	5102	0	1	0.3548305	-1.4947981
56	55	Cas9-CP-55	137	0	0	0	-inf
57	56	Cas9-CP-56	2811	1	1	1.14552964	0.19601479
59	58	Cas9-CP-58	1598	0	1	1.13288185	0.17999741
61	60	Cas9-CP-60	1	0	0	0	-inf
68	67	Cas9-CP-67	6596	1	0	0.21372629	-2.2261637
71	70	Cas9-CP-70	138	0	0	0	-inf
73	72	Cas9-CP-72	5189	0	0	0	-inf
74	73	Cas9-CP-73	2213	0	0	0	-inf
75	74	Cas9-CP-74	26490	3	2	0.2963347	-1.7547005
76	75	Cas9-CP-75	1	0	0	0	-inf
78	77	Cas9-CP-77	5224	1	0	0.26985808	-1.8897272
80	79	Cas9-CP-79	5757	0	0	0	-inf
81	80	Cas9-CP-80	628	0	0	0	-inf
82	81	Cas9-CP-81	15141	0	0	0	-inf
83	82	Cas9-CP-82	3501	0	0	0	-inf
84	83	Cas9-CP-83	83	0	0	0	-inf
89	88	Cas9-CP-88	5	0	0	0	-inf
93	92	Cas9-CP-92	2649	0	0	0	-inf
94	93	Cas9-CP-93	7612	0	0	0	-inf
95	94	Cas9-CP-94	1578	0	0	0	-inf
99	98	Cas9-CP-98	6893	0	0	0	-inf
100	99	Cas9-CP-99	27018	5	0	0.26088878	-1.9384932
102	101	Cas9-CP-101	1540	0	0	0	-inf
103	102	Cas9-CP-102	5895	0	0	0	-inf
107	106	Cas9-CP-106	1809	0	0	0	-inf
108	107	Cas9-CP-107	1	0	0	0	-inf
109	108	Cas9-CP-108	77	0	0	0	-inf
114	113	Cas9-CP-113	2477	8205	5312	8552.06261	13.0620567
116	115	Cas9-CP-115	1420	0	0	0	-inf
117	116	Cas9-CP-116	6167	0	3	0.88066087	-0.1833415
123	122	Cas9-CP-122	0	0	0		
124	123	Cas9-CP-123	7380	1	0	0.19102149	-2.3881931
125	124	Cas9-CP-124	212	0	0	0	-inf
128	127	Cas9-CP-127	6479	0	1	0.27941738	-1.8395063
129	128	Cas9-CP-128	8830	1	0	0.1596533	-2.6469857
130	129	Cas9-CP-129	23324	1	1	0.13805882	-2.8566451
133	132	Cas9-CP-132	507	0	0	0	-inf
138	137	Cas9-CP-137	4521	0	0	0	-inf
139	138	Cas9-CP-138	1	0	0	0	-inf
144	143	Cas9-CP-143	998	0	0	0	-inf
146	145	Cas9-CP-145	677	0	0	0	-inf
147	146	Cas9-CP-146	16163	0	0	0	-inf
148	147	Cas9-CP-147	825	0	0	0	-inf
150	149	Cas9-CP-149	2259	0	0	0	-inf
152	151	Cas9-CP-151	2857	1	0	0.49343319	-1.0190733
160	159	Cas9-CP-159	1853	0	0	0	-inf
163	162	Cas9-CP-162	2736	1	0	0.51525535	-0.9566405
165	164	Cas9-CP-164	7901	1	0	0.17842534	-2.4866076
166	165	Cas9-CP-165	17115	5	1	0.51761836	-0.9500393
171	170	Cas9-CP-170	22107	0	0	0	-inf
172	171	Cas9-CP-171	12000	1	0	0.11747822	-3.0895348
176	175	Cas9-CP-175	851	0	0	0	-inf
177	176	Cas9-CP-176	32048	5	2	0.33291886	-1.5867575
178	177	Cas9-CP-177	11488	46023	96058	20785.04	14.3432579
179	178	Cas9-CP-178	2301	1	1	1.399428	0.48483726
180	179	Cas9-CP-179	456	4848	26826	121488.45	16.8904596
181	180	Cas9-CP-180	9	0	0	0	-inf
182	181	Cas9-CP-181	2979	106439	148650	140704.593	17.1023099
187	186	Cas9-CP-186	22602	1	0	0.0623723	-4.0029508
189	188	Cas9-CP-188	0	0	0		

192	191	Cas9-CP-191	19279	1	2	0.2609279	-1.9382769
194	193	Cas9-CP-193	1	0	0	0	-inf
197	196	Cas9-CP-196	10081	711694	857484	253511.017	17.9516889
198	197	Cas9-CP-197	0	1	2	inf	inf
200	199	Cas9-CP-199	3277	166614	262172	216510.226	17.7240756
201	200	Cas9-CP-200	6497	232878	811993	276786.938	18.0784163
202	201	Cas9-CP-201	0	1	1	inf	inf
203	202	Cas9-CP-202	6	0	0	0	-inf
205	204	Cas9-CP-204	11623	1	1	0.27704412	-1.8518123
207	206	Cas9-CP-206	535	0	0	0	-inf
209	208	Cas9-CP-208	12240	3	0	0.34552417	-1.5331415
213	212	Cas9-CP-212	2	0	0	0	-inf
214	213	Cas9-CP-213	5188	437695	755066	382414.351	18.5447771
215	214	Cas9-CP-214	1	0	0	0	-inf
216	215	Cas9-CP-215	1	0	0	0	-inf
221	220	Cas9-CP-220	1	0	0	0	-inf
226	225	Cas9-CP-225	1	0	0	0	-inf
228	227	Cas9-CP-227	1472	0	0	0	-inf
230	229	Cas9-CP-229	1971	0	0	0	-inf
231	230	Cas9-CP-230	140396	5470733	7404127	150405.491	17.1984977
232	231	Cas9-CP-231	3241	229005	316679	276499.691	18.0769183
233	232	Cas9-CP-232	0	2	4	inf	inf
234	233	Cas9-CP-233	0	2	1	inf	inf
235	234	Cas9-CP-234	0	1	0	inf	inf
239	238	Cas9-CP-238	1952	0	0	0	-inf
241	240	Cas9-CP-240	689	0	0	0	-inf
242	241	Cas9-CP-241	5958	0	0	0	-inf
245	244	Cas9-CP-244	0	0	1	inf	inf
247	246	Cas9-CP-246	86	0	0	0	-inf
250	249	Cas9-CP-249	353	0	0	0	-inf
260	259	Cas9-CP-259	5112	108159	127404	74945.4497	16.1935533
261	260	Cas9-CP-260	0	0	0		
262	261	Cas9-CP-261	38	0	0	0	-inf
265	264	Cas9-CP-264	13734	7442	1	764.022513	9.57747134
266	265	Cas9-CP-265	2	0	0	0	-inf
268	267	Cas9-CP-267	1	0	0	0	-inf
269	268	Cas9-CP-268	1	0	0	0	-inf
271	270	Cas9-CP-270	21791	737504	746456	109725.616	16.7435408
272	271	Cas9-CP-271	567	1	1	5.67916017	2.5056776
274	273	Cas9-CP-273	674	12108	2	25330.4688	14.6285862
280	279	Cas9-CP-279	0	1	0	inf	inf
281	280	Cas9-CP-280	0	0	1	inf	inf
283	282	Cas9-CP-282	5423	0	0	0	-inf
285	284	Cas9-CP-284	1625	1	0	0.86753146	-0.205012
287	286	Cas9-CP-286	6046	0	1	0.29942858	-1.7397162
288	287	Cas9-CP-287	5732	0	1	0.31583133	-1.6627738
289	288	Cas9-CP-288	0	0	1	inf	inf
298	297	Cas9-CP-297	2032	0	0	0	-inf
299	298	Cas9-CP-298	514	0	0	0	-inf
302	301	Cas9-CP-301	0	0	0		
304	303	Cas9-CP-303	2140	0	0	0	-inf
306	305	Cas9-CP-305	229	0	0	0	-inf
307	306	Cas9-CP-306	1	0	0	0	-inf
308	307	Cas9-CP-307	146	0	0	0	-inf
311	310	Cas9-CP-310	30138	437290	508170	50979.7505	15.6376367
314	313	Cas9-CP-313	13697	9842	1	1013.10199	9.9845637
316	315	Cas9-CP-315	2026	0	0	0	-inf
317	316	Cas9-CP-316	5906	0	1	0.30652645	-1.7059165
319	318	Cas9-CP-318	9149	0	1	0.19787356	-2.3373492
320	319	Cas9-CP-319	0	0	1	inf	inf
324	323	Cas9-CP-323	877	0	0	0	-inf
326	325	Cas9-CP-325	24354	7	2	0.55386634	-0.8523902
327	326	Cas9-CP-326	17679	3	1	0.34162346	-1.5495211
328	327	Cas9-CP-327	4465	0	1	0.40545245	-1.3023954
329	328	Cas9-CP-328	5189	0	0	0	-inf
330	329	Cas9-CP-329	1685	0	1	1.07438884	0.10351622
331	330	Cas9-CP-330	10303	1	1	0.31253847	-1.6778943
332	331	Cas9-CP-331	11116	0	1	0.16285941	-2.618301
345	344	Cas9-CP-344	3151	0	2	1.14906074	0.20045506
346	345	Cas9-CP-345	1662	0	0	0	-inf
347	346	Cas9-CP-346	1	0	0	0	-inf

353	352	Cas9-CP-352	4547	0	0	0	-inf
354	353	Cas9-CP-353	1	0	0	0	-inf
358	357	Cas9-CP-357	2466	2	0	1.14334033	0.1932549
360	359	Cas9-CP-359	5044	0	0	0	-inf
361	360	Cas9-CP-360	16006	3	1	0.37733107	-1.4060972
362	361	Cas9-CP-361	20	0	0	0	-inf
365	364	Cas9-CP-364	25712	3	3	0.37570984	-1.4123092
366	365	Cas9-CP-365	5682	1	0	0.24810606	-2.0109711
368	367	Cas9-CP-367	3702	5841	39577	21578.151	14.3972836
375	374	Cas9-CP-374	0	0	1	inf	inf
385	384	Cas9-CP-384	16411	26193	8022	3134.9627	11.6142326
387	386	Cas9-CP-386	24412	3	1	0.24740132	-2.0150749
391	390	Cas9-CP-390	1	0	0	0	-inf
395	394	Cas9-CP-394	2590	1	0	0.54430063	-0.8775244
396	395	Cas9-CP-395	10959	0	1	0.16519255	-2.5977794
398	397	Cas9-CP-397	2927	0	0	0	-inf
400	399	Cas9-CP-399	16233	1	2	0.30988905	-1.6901763
401	400	Cas9-CP-400	9838	0	0	0	-inf
403	402	Cas9-CP-402	5159	1	1	0.62416821	-0.6799932
404	403	Cas9-CP-403	6717	2	0	0.41975246	-1.2523893
406	405	Cas9-CP-405	2154	0	0	0	-inf
407	406	Cas9-CP-406	3870	1	1	0.832063	-0.2652353
409	408	Cas9-CP-408	3745	1	0	0.37643221	-1.409538
410	409	Cas9-CP-409	1	0	0	0	-inf
411	410	Cas9-CP-410	3852	1	1	0.83595115	-0.2585095
412	411	Cas9-CP-411	99920	3	2	0.07856191	-3.6700261
413	412	Cas9-CP-412	10488	0	2	0.3452222	-1.5344029
415	414	Cas9-CP-414	2393	0	0	0	-inf
416	415	Cas9-CP-415	5942	0	0	0	-inf
417	416	Cas9-CP-416	2	0	0	0	-inf
418	417	Cas9-CP-417	910	0	0	0	-inf
419	418	Cas9-CP-418	2335	0	1	0.77530843	-0.3671577
420	419	Cas9-CP-419	138	0	0	0	-inf
423	422	Cas9-CP-422	2740	1	0	0.51450315	-0.9587482
425	424	Cas9-CP-424	1	0	0	0	-inf
428	427	Cas9-CP-427	626	0	0	0	-inf
430	429	Cas9-CP-429	850	0	0	0	-inf
431	430	Cas9-CP-430	7630	0	1	0.23726674	-2.0754182
435	434	Cas9-CP-434	8966	0	0	0	-inf
436	435	Cas9-CP-435	25448	12345	5236	1056.3577	10.0448827
437	436	Cas9-CP-436	2530	0	2	1.43110292	0.51712743
438	437	Cas9-CP-437	1260	0	1	1.4367819	0.52284108
439	438	Cas9-CP-438	6713	0	1	0.26967752	-1.8906928
440	439	Cas9-CP-439	1726	0	0	0	-inf
441	440	Cas9-CP-440	8721	0	0	0	-inf
444	443	Cas9-CP-443	25	0	0	0	-inf
445	444	Cas9-CP-444	0	0	0		
446	445	Cas9-CP-445	2206	0	0	0	-inf
447	446	Cas9-CP-446	2151	0	0	0	-inf
450	449	Cas9-CP-449	279	0	0	0	-inf
451	450	Cas9-CP-450	133	0	0	0	-inf
453	452	Cas9-CP-452	6542	0	0	0	-inf
454	453	Cas9-CP-453	2507	0	0	0	-inf
456	455	Cas9-CP-455	1	0	0	0	-inf
458	457	Cas9-CP-457	3225	0	1	0.56134735	-0.8330343
459	458	Cas9-CP-458	3807	35068	22780	23818.3288	14.5397846
461	460	Cas9-CP-460	300	0	0	0	-inf
463	462	Cas9-CP-462	1371	0	0	0	-inf
464	463	Cas9-CP-463	2198	0	0	0	-inf
467	466	Cas9-CP-466	1655	0	0	0	-inf
470	469	Cas9-CP-469	29006	4	1	0.25681927	-1.9611746
475	474	Cas9-CP-474	8386	1	0	0.1681062	-2.5725551
480	479	Cas9-CP-479	6624	0	0	0	-inf
482	481	Cas9-CP-481	9243	1	5649	1106.57251	10.1118823
483	482	Cas9-CP-482	66	1	0	21.3596761	4.41681787
484	483	Cas9-CP-483	8984	0	0	0	-inf
485	484	Cas9-CP-484	898	0	0	0	-inf
486	485	Cas9-CP-485	5077	1	0	0.27767158	-1.8485486
489	488	Cas9-CP-488	16777	1	8212	886.21115	9.79150667
494	493	Cas9-CP-493	1	0	0	0	-inf
495	494	Cas9-CP-494	2556	0	0	0	-inf

499	498	Cas9-CP-498	472	0	0	0	-inf
500	499	Cas9-CP-499	1	0	0	0	-inf
503	502	Cas9-CP-502	2196	0	0	0	-inf
510	509	Cas9-CP-509	9192	2	1	0.50367955	-0.9894219
511	510	Cas9-CP-510	772	0	0	0	-inf
513	512	Cas9-CP-512	4135	0	0	0	-inf
514	513	Cas9-CP-513	5	0	0	0	-inf
515	514	Cas9-CP-514	1	0	0	0	-inf
516	515	Cas9-CP-515	2	0	0	0	-inf
517	516	Cas9-CP-516	1912	0	0	0	-inf
519	518	Cas9-CP-518	33	0	0	0	-inf
521	520	Cas9-CP-520	8655	2	0	0.32576282	-1.6181061
522	521	Cas9-CP-521	16844	2	3	0.48981909	-1.0296791
525	524	Cas9-CP-524	1	0	0	0	-inf
526	525	Cas9-CP-525	5538	0	1	0.32689512	-1.6131002
531	530	Cas9-CP-530	5320	0	0	0	-inf
532	531	Cas9-CP-531	8542	1	1	0.37697071	-1.4074757
533	532	Cas9-CP-532	1	0	0	0	-inf
535	534	Cas9-CP-534	3620	0	1	0.50009536	-0.9997249
542	541	Cas9-CP-541	255	0	1	7.09939291	2.82769566
543	542	Cas9-CP-542	10	0	0	0	-inf
544	543	Cas9-CP-543	3	0	0	0	-inf
547	546	Cas9-CP-546	1029	0	0	0	-inf
548	547	Cas9-CP-547	1	0	0	0	-inf
550	549	Cas9-CP-549	1	0	0	0	-inf
552	551	Cas9-CP-551	1	0	0	0	-inf
554	553	Cas9-CP-553	4184	0	0	0	-inf
559	558	Cas9-CP-558	2405	4705	2	2759.43489	11.4301571
562	561	Cas9-CP-561	5980	0	0	0	-inf
563	562	Cas9-CP-562	5766	1	0	0.24449161	-2.0321431
565	564	Cas9-CP-564	27	0	0	0	-inf
566	565	Cas9-CP-565	4562	0	0	0	-inf
568	567	Cas9-CP-567	4133	1	0	0.3410933	-1.5517617
569	568	Cas9-CP-568	1	0	0	0	-inf
570	569	Cas9-CP-569	103	0	0	0	-inf
573	572	Cas9-CP-572	5	0	0	0	-inf
576	575	Cas9-CP-575	13626	2	0	0.20691892	-2.2728625
577	576	Cas9-CP-576	37791	2	2	0.17041538	-2.5528725
578	577	Cas9-CP-577	1702	0	0	0	-inf
579	578	Cas9-CP-578	0	0	0		
582	581	Cas9-CP-581	11221	1	5	0.93231125	-0.1011164
584	583	Cas9-CP-583	3032	0	0	0	-inf
586	585	Cas9-CP-585	6811	0	1	0.26579727	-1.9116018
588	587	Cas9-CP-587	1	0	0	0	-inf
589	588	Cas9-CP-588	4522	1	1	0.71209284	-0.4898628
590	589	Cas9-CP-589	1001	0	0	0	-inf
591	590	Cas9-CP-590	272	0	0	0	-inf
593	592	Cas9-CP-592	245	0	0	0	-inf
594	593	Cas9-CP-593	17620	1	1	0.18275164	-2.4520438
595	594	Cas9-CP-594	16726	2	1	0.27680392	-1.8530637
598	597	Cas9-CP-597	353	0	0	0	-inf
599	598	Cas9-CP-598	1	0	0	0	-inf
601	600	Cas9-CP-600	4	0	0	0	-inf
602	601	Cas9-CP-601	3849	1	0	0.36626101	-1.449056
603	602	Cas9-CP-602	1419	0	0	0	-inf
604	603	Cas9-CP-603	667	0	1	2.71416071	1.44050615
608	607	Cas9-CP-607	0	1	0	inf	inf
609	608	Cas9-CP-608	1317	0	0	0	-inf
610	609	Cas9-CP-609	6683	0	0	0	-inf
611	610	Cas9-CP-610	1991	0	0	0	-inf
613	612	Cas9-CP-612	19639	1	3	0.34832599	-1.52149
614	613	Cas9-CP-613	1726	1	0	0.81676629	-0.2920048
617	616	Cas9-CP-616	20429	3	0	0.20702021	-2.2721565
618	617	Cas9-CP-617	1528	0	0	0	-inf
623	622	Cas9-CP-622	1	0	0	0	-inf
624	623	Cas9-CP-623	2	0	0	0	-inf
625	624	Cas9-CP-624	3419	0	0	0	-inf
627	626	Cas9-CP-626	28024	3	3	0.34471351	-1.5365302
628	627	Cas9-CP-627	1233	1	0	1.14334033	0.1932549
629	628	Cas9-CP-628	37982	2	1	0.12189517	-3.0362871
630	629	Cas9-CP-629	302	0	0	0	-inf

633	632	Cas9-CP-632	10130	0	0	0	-inf
634	633	Cas9-CP-633	3331	1	0	0.42321784	-1.2405277
638	637	Cas9-CP-637	5036	1	2	0.99889377	-0.0015968
640	639	Cas9-CP-639	898	0	0	0	-inf
641	640	Cas9-CP-640	2889	0	0	0	-inf
642	641	Cas9-CP-641	23004	1	4	0.37607022	-1.410926
644	643	Cas9-CP-643	1	0	0	0	-inf
645	644	Cas9-CP-644	49	0	0	0	-inf
646	645	Cas9-CP-645	4760	1	0	0.29616358	-1.7555339
648	647	Cas9-CP-647	1	0	0	0	-inf
650	649	Cas9-CP-649	2	0	0	0	-inf
651	650	Cas9-CP-650	1	0	0	0	-inf
652	651	Cas9-CP-651	1466	0	0	0	-inf
654	653	Cas9-CP-653	2245	0	1	0.80638984	-0.3104506
655	654	Cas9-CP-654	45594	5	3	0.27371428	-1.8692574
656	655	Cas9-CP-655	125	0	0	0	-inf
657	656	Cas9-CP-656	1	0	0	0	-inf
658	657	Cas9-CP-657	22094	5	2	0.48290864	-1.0501778
660	659	Cas9-CP-659	6000	1	0	0.23495644	-2.0895348
664	663	Cas9-CP-663	47	0	0	0	-inf
666	665	Cas9-CP-665	0	0	1	inf	inf
668	667	Cas9-CP-667	22048	0	4	0.32843708	-1.6063111
669	668	Cas9-CP-668	35476	5	3	0.35177948	-1.5072568
671	670	Cas9-CP-670	18912	2	1	0.24480872	-2.0302732
672	671	Cas9-CP-671	1	0	0	0	-inf
673	672	Cas9-CP-672	3975	1	0	0.35465123	-1.4955272
674	673	Cas9-CP-673	1	0	0	0	-inf
676	675	Cas9-CP-675	73862	6	5	0.23706585	-2.0766402
677	676	Cas9-CP-676	3414	0	0	0	-inf
681	680	Cas9-CP-680	1	0	0	0	-inf
682	681	Cas9-CP-681	2	0	0	0	-inf
684	683	Cas9-CP-683	0	1	0	inf	inf
686	685	Cas9-CP-685	13761	8233	12929	2544.3159	11.3130621
687	686	Cas9-CP-686	14703	1	12455	1533.65022	10.5827538
688	687	Cas9-CP-687	1112	0	0	0	-inf
690	689	Cas9-CP-689	4126	1	0	0.34167199	-1.5493161
691	690	Cas9-CP-690	1	0	0	0	-inf
695	694	Cas9-CP-694	13097	4	3	0.84523097	-0.2425825
696	695	Cas9-CP-695	1	0	0	0	-inf
697	696	Cas9-CP-696	444	0	0	0	-inf
698	697	Cas9-CP-697	247988	25	19966	145.896558	7.18880204
699	698	Cas9-CP-698	29903	2	0	0.09428744	-3.4067906
700	699	Cas9-CP-699	1980	0	0	0	-inf
701	700	Cas9-CP-700	8752	1	0	0.16107617	-2.634185
705	704	Cas9-CP-704	34	0	0	0	-inf
706	705	Cas9-CP-705	549	0	0	0	-inf
708	707	Cas9-CP-707	3883	0	1	0.46622333	-1.1009069
709	708	Cas9-CP-708	9978	0	0	0	-inf
712	711	Cas9-CP-711	1025	0	0	0	-inf
715	714	Cas9-CP-714	12806	1	0	0.11008423	-3.1833203
717	716	Cas9-CP-716	2	0	0	0	-inf
718	717	Cas9-CP-717	16364	1	1	0.19677853	-2.3453553
720	719	Cas9-CP-719	6487	11988	3	2606.03941	11.3476432
721	720	Cas9-CP-720	4	0	0	0	-inf
722	721	Cas9-CP-721	11428	4	2	0.81025944	-0.3035442
723	722	Cas9-CP-722	9464	0	0	0	-inf
729	728	Cas9-CP-728	2108	0	0	0	-inf
730	729	Cas9-CP-729	14239	3	1	0.42415627	-1.2373322
731	730	Cas9-CP-730	2194	0	0	0	-inf
732	731	Cas9-CP-731	2	0	0	0	-inf
736	735	Cas9-CP-735	1956	0	0	0	-inf
739	738	Cas9-CP-738	16	0	0	0	-inf
741	740	Cas9-CP-740	22953	2	0	0.12283698	-3.0251831
744	743	Cas9-CP-743	1	0	0	0	-inf
745	744	Cas9-CP-744	11038	2	2	0.58345422	-0.7773086
746	745	Cas9-CP-745	8186	3	1	0.73779148	-0.438715
748	747	Cas9-CP-747	2	0	0	0	-inf
749	748	Cas9-CP-748	19967	2	0	0.14120685	-2.824118
753	752	Cas9-CP-752	3	0	0	0	-inf
756	755	Cas9-CP-755	1185	0	1	1.52771746	0.61137775
757	756	Cas9-CP-756	26664	2	2	0.24153044	-2.0497231

760	759	Cas9-CP-759	24	0	0	0	-inf
761	760	Cas9-CP-760	0	0	0		
762	761	Cas9-CP-761	3321	0	1	0.5451205	-0.8753529
765	764	Cas9-CP-764	12967	0	0	0	-inf
768	767	Cas9-CP-767	15975	0	1	0.11332364	-3.1414792
774	773	Cas9-CP-773	108436	11	7	0.25987256	-1.9441238
775	774	Cas9-CP-774	1	0	0	0	-inf
779	778	Cas9-CP-778	9611	1	0	0.1466797	-2.7692588
780	779	Cas9-CP-779	10966	2	0	0.25711082	-1.9595378
782	781	Cas9-CP-781	17483	2	0	0.16126965	-2.6324532
783	782	Cas9-CP-782	7633	0	0	0	-inf
785	784	Cas9-CP-784	1284	0	0	0	-inf
786	785	Cas9-CP-785	4815	0	0	0	-inf
787	786	Cas9-CP-786	2189	0	0	0	-inf
792	791	Cas9-CP-791	2533	0	1	0.71470398	-0.4845823
794	793	Cas9-CP-793	1396	0	1	1.29680888	0.37496587
798	797	Cas9-CP-797	6275	0	0	0	-inf
799	798	Cas9-CP-798	494	0	0	0	-inf
800	799	Cas9-CP-799	12080	0	1	0.14986301	-2.7382837
801	800	Cas9-CP-800	11506	2	7081	1114.36414	10.122005
802	801	Cas9-CP-801	0	0	1	inf	inf
807	806	Cas9-CP-806	1106	0	1	1.63684014	0.71091343
808	807	Cas9-CP-807	6202	0	1	0.291897	-1.7764687
809	808	Cas9-CP-808	2524	0	0	0	-inf
812	811	Cas9-CP-811	2871	0	0	0	-inf
813	812	Cas9-CP-812	98	0	0	0	-inf
814	813	Cas9-CP-813	19897	2	1	0.23268947	-2.1035222
816	815	Cas9-CP-815	1277	0	0	0	-inf
817	816	Cas9-CP-816	26561	2	3	0.31062508	-1.6867538
819	818	Cas9-CP-818	1464	0	1	1.23657459	0.30634926
820	819	Cas9-CP-819	2745	0	0	0	-inf
821	820	Cas9-CP-820	1	0	0	0	-inf
822	821	Cas9-CP-821	1057	0	0	0	-inf
824	823	Cas9-CP-823	8709	1	1	0.36974208	-1.4354088
825	824	Cas9-CP-824	3680	0	0	0	-inf
826	825	Cas9-CP-825	3246	1	0	0.43430025	-1.2032353
827	826	Cas9-CP-826	1638	0	0	0	-inf
829	828	Cas9-CP-828	1	0	0	0	-inf
830	829	Cas9-CP-829	6044	2	0	0.46649193	-1.100076
832	831	Cas9-CP-831	19169	1	0	0.07354263	-3.7652755
833	832	Cas9-CP-832	2303	0	0	0	-inf
834	833	Cas9-CP-833	8483	3	1	0.71196052	-0.4901309
835	834	Cas9-CP-834	7671	1	1	0.41977367	-1.2523164
836	835	Cas9-CP-835	5357	0	0	0	-inf
837	836	Cas9-CP-836	12467	1	1	0.25828859	-1.9529442
838	837	Cas9-CP-837	7299	0	0	0	-inf
840	839	Cas9-CP-839	1	0	0	0	-inf
841	840	Cas9-CP-840	1	0	0	0	-inf
842	841	Cas9-CP-841	1	0	0	0	-inf
843	842	Cas9-CP-842	4495	0	1	0.40274643	-1.3120563
848	847	Cas9-CP-847	4314	0	1	0.41964423	-1.2527614
849	848	Cas9-CP-848	2732	1	0	0.51600975	-0.9545298
853	852	Cas9-CP-852	4183	0	4	1.7311453	0.79172682
859	858	Cas9-CP-858	10733	1	1	0.30001713	-1.7368832
861	860	Cas9-CP-860	7839	0	1	0.23094083	-2.1144048
864	863	Cas9-CP-863	130	0	0	0	-inf
865	864	Cas9-CP-864	8866	3	1	0.68120472	-0.5538397
868	867	Cas9-CP-867	1633	0	0	0	-inf
870	869	Cas9-CP-869	4508	0	1	0.401585	-1.3162227
871	870	Cas9-CP-870	3172	1	1	1.01515883	0.02170547
873	872	Cas9-CP-872	4005	0	0	0	-inf
874	873	Cas9-CP-873	3	0	0	0	-inf
877	876	Cas9-CP-876	0	0	0		
880	879	Cas9-CP-879	0	0	0		
883	882	Cas9-CP-882	3083	0	0	0	-inf
885	884	Cas9-CP-884	1	0	0	0	-inf
887	886	Cas9-CP-886	9779	1	0	0.14415979	-2.7942592
889	888	Cas9-CP-888	2	0	0	0	-inf
890	889	Cas9-CP-889	84	0	0	0	-inf
893	892	Cas9-CP-892	164	0	0	0	-inf
894	893	Cas9-CP-893	116	0	0	0	-inf

895	894	Cas9-CP-894	0	1	0	inf	inf
896	895	Cas9-CP-895	1198	0	0	0	-inf
898	897	Cas9-CP-897	9075	1	1	0.35483017	-1.4947994
899	898	Cas9-CP-898	6785	1	0	0.20777283	-2.2669211
901	900	Cas9-CP-900	3	0	0	0	-inf
904	903	Cas9-CP-903	5110	1	1	0.63015339	-0.666225
905	904	Cas9-CP-904	402	0	0	0	-inf
906	905	Cas9-CP-905	5347	4	0	1.05460155	0.07669803
907	906	Cas9-CP-906	2	0	0	0	-inf
910	909	Cas9-CP-909	4435	0	1	0.40819508	-1.2926693
912	911	Cas9-CP-911	413	0	0	0	-inf
913	912	Cas9-CP-912	1156	0	0	0	-inf
915	914	Cas9-CP-914	15678	1	1	0.20538869	-2.2835714
919	918	Cas9-CP-918	14	0	0	0	-inf
920	919	Cas9-CP-919	21943	1	1	0.14674766	-2.7685906
923	922	Cas9-CP-922	6643	2	1	0.69694753	-0.520878
925	924	Cas9-CP-924	4051	0	0	0	-inf
929	928	Cas9-CP-928	469	0	0	0	-inf
931	930	Cas9-CP-930	7757	1	3	0.881884	-0.1813392
932	931	Cas9-CP-931	21	0	0	0	-inf
933	932	Cas9-CP-932	11	0	0	0	-inf
938	937	Cas9-CP-937	7085	0	0	0	-inf
939	938	Cas9-CP-938	7546	0	0	0	-inf
941	940	Cas9-CP-940	1125	0	0	0	-inf
942	941	Cas9-CP-941	20024	3352	1810	399.628879	8.64251703
944	943	Cas9-CP-943	4044	0	0	0	-inf
945	944	Cas9-CP-944	23084	4075	1	248.938453	7.95964529
947	946	Cas9-CP-946	3655	0	0	0	-inf
948	947	Cas9-CP-947	3701	0	0	0	-inf
949	948	Cas9-CP-948	1	0	0	0	-inf
951	950	Cas9-CP-950	2955	6158	1	2938.40298	11.5208165
953	952	Cas9-CP-952	1	0	0	0	-inf
954	953	Cas9-CP-953	1554	0	0	0	-inf
955	954	Cas9-CP-954	77	0	0	0	-inf
961	960	Cas9-CP-960	0	1	0	inf	inf
965	964	Cas9-CP-964	133	0	0	0	-inf
966	965	Cas9-CP-965	30	0	0	0	-inf
967	966	Cas9-CP-966	10790	1	1	0.29843224	-1.7445247
971	970	Cas9-CP-970	1481	0	1	1.22238028	0.28969317
973	972	Cas9-CP-972	1825	0	6373	6321.82461	12.6261253
975	974	Cas9-CP-974	5600	0	1	0.32327593	-1.629162
976	975	Cas9-CP-975	4829	0	0	0	-inf
977	976	Cas9-CP-976	15211	1	1	0.21169442	-2.2399449
979	978	Cas9-CP-978	16725	1	2	0.30077303	-1.7332529
980	979	Cas9-CP-979	1	0	0	0	-inf
981	980	Cas9-CP-980	3	0	0	0	-inf
982	981	Cas9-CP-981	46051	6	4	0.34092229	-1.5524852
983	982	Cas9-CP-982	129	0	0	0	-inf
984	983	Cas9-CP-983	7421	3	0	0.56989838	-0.8112234
985	984	Cas9-CP-984	32244	8	1	0.40591286	-1.3007581
986	985	Cas9-CP-985	5998	1	0	0.23503478	-2.0890538
987	986	Cas9-CP-986	4640	0	1	0.3901606	-1.35786
989	988	Cas9-CP-988	655	0	0	0	-inf
990	989	Cas9-CP-989	4044	0	1	0.44766202	-1.1595182
993	992	Cas9-CP-992	5784	1	0	0.24373074	-2.0366399
994	993	Cas9-CP-993	3705	0	0	0	-inf
996	995	Cas9-CP-995	0	1	0	inf	inf
997	996	Cas9-CP-996	1	0	0	0	-inf
999	998	Cas9-CP-998	1	1	0	1409.73862	10.461212
1003	1002	Cas9-CP-1002	88	1	0	16.0197571	4.00178037
1005	1004	Cas9-CP-1004	1	0	0	0	-inf
1006	1005	Cas9-CP-1005	1	0	0	0	-inf
1010	1009	Cas9-CP-1009	83673	92654	78360	3256.44559	11.6690824
1011	1010	Cas9-CP-1010	37778	1559295	2144493	160952.697	17.2962772
1012	1011	Cas9-CP-1011	2417	96278	84638	119549.364	16.8672469

1013	1012	Cas9-CP-1012	6456	41280	58257	25349.952	14.6296954
1014	1013	Cas9-CP-1013	22	0	0	0	-inf
1016	1015	Cas9-CP-1015	889	34161	57711	171692.815	17.3894701
1017	1016	Cas9-CP-1016	9680	287087	454011	126718.416	16.9512667
1018	1017	Cas9-CP-1017	6089	323821	586413	249320.566	17.9276424
1019	1018	Cas9-CP-1018	40	0	0	0	-inf
1020	1019	Cas9-CP-1019	683	0	0	0	-inf
1024	1023	Cas9-CP-1023	11355	130012	199248	47907.5823	15.5479664
1026	1025	Cas9-CP-1025	23207	173721	89017	17496.9924	14.0948193
1027	1026	Cas9-CP-1026	3	0	0	0	-inf
1028	1027	Cas9-CP-1027	2288	7965	0	4907.59097	12.2607993
1029	1028	Cas9-CP-1028	1	0	0	0	-inf
1030	1029	Cas9-CP-1029	10326	70475	98666	26919.5088	14.7163645
1031	1030	Cas9-CP-1030	1715	154706	240148	380668.106	18.5381742
1032	1031	Cas9-CP-1031	0	0	0		
1038	1037	Cas9-CP-1037	1	0	0	0	-inf
1039	1038	Cas9-CP-1038	16502	1	0	0.08542835	-3.5491413
1040	1039	Cas9-CP-1039	10227	85920	51412	20944.3835	14.3542758
1041	1040	Cas9-CP-1040	8	0	0	0	-inf
1042	1041	Cas9-CP-1041	8398	380476	730854	221418.164	17.7564141
1043	1042	Cas9-CP-1042	0	0	3	inf	inf
1046	1045	Cas9-CP-1045	24064	2	4	0.41808752	-1.2581231
1049	1048	Cas9-CP-1048	406	0	0	0	-inf
1053	1052	Cas9-CP-1052	1287	0	0	0	-inf
1055	1054	Cas9-CP-1054	14245	4	6306	801.802439	9.647103
1056	1055	Cas9-CP-1055	2520	0	0	0	-inf
1058	1057	Cas9-CP-1057	1569	0	0	0	-inf
1061	1060	Cas9-CP-1060	38160	0	2	0.09488182	-3.3977245
1064	1063	Cas9-CP-1063	2	0	0	0	-inf
1067	1066	Cas9-CP-1066	132793	16	5	0.23802116	-2.0708383
1068	1067	Cas9-CP-1067	691	0	0	0	-inf
1070	1069	Cas9-CP-1069	44978	1	5	0.2325907	-2.1041347
1073	1072	Cas9-CP-1072	1	0	0	0	-inf
1075	1074	Cas9-CP-1074	1277	0	0	0	-inf
1084	1083	Cas9-CP-1083	5254	0	0	0	-inf

1090	1089	Cas9-CP-1089	13028	0	5909	821.102989	9.68141938
1093	1092	Cas9-CP-1092	3985	1	0	0.35376126	-1.499152
1095	1094	Cas9-CP-1094	1622	0	1	1.11611911	0.15849099
1096	1095	Cas9-CP-1095	10327	1	1	0.31181213	-1.6812511
1101	1100	Cas9-CP-1100	6803	0	0	0	-inf
1103	1102	Cas9-CP-1102	17353	0	2	0.20864925	-2.2608484
1104	1103	Cas9-CP-1103	11606	1	4	0.7454006	-0.4239121
1107	1106	Cas9-CP-1106	2	0	0	0	-inf
1108	1107	Cas9-CP-1107	2276	0	0	0	-inf
1112	1111	Cas9-CP-1111	1	0	0	0	-inf
1116	1115	Cas9-CP-1115	764	8073	2	14901.1003	13.8631312
1117	1116	Cas9-CP-1116	1828	29329	65276	87263.8495	16.4130965
1118	1117	Cas9-CP-1117	4181	27138	58335	34408.9868	15.0704978
1119	1118	Cas9-CP-1118	32851	5	4	0.43499662	-1.2009239
1121	1120	Cas9-CP-1120	4947	1	0	0.28496839	-1.8111262
1122	1121	Cas9-CP-1121	8604	2	0	0.32769378	-1.6095798
1127	1126	Cas9-CP-1126	4798	1	0	0.29381797	-1.7670055
1132	1131	Cas9-CP-1131	40494	4	1	0.18396058	-2.4425315
1133	1132	Cas9-CP-1132	27443	1	1	0.11733716	-3.0912681
1137	1136	Cas9-CP-1136	7751	1	0	0.18187829	-2.4589548
1140	1139	Cas9-CP-1139	1	0	0	0	-inf
1141	1140	Cas9-CP-1140	2766	0	0	0	-inf
1144	1143	Cas9-CP-1143	5717	1	0	0.24658713	-2.0198306
1145	1144	Cas9-CP-1144	12041	1	1	0.26742661	-1.9027851
1147	1146	Cas9-CP-1146	1549	0	0	0	-inf
1148	1147	Cas9-CP-1147	8881	51684	53206	19049.8995	14.2174958
1149	1148	Cas9-CP-1148	2	0	0	0	-inf
1150	1149	Cas9-CP-1149	9459	679472	814472	257146.991	17.9722337
1151	1150	Cas9-CP-1150	0	3	2	inf	inf
1152	1151	Cas9-CP-1151	1	0	0	0	-inf
1156	1155	Cas9-CP-1155	2	0	0	0	-inf
1160	1159	Cas9-CP-1159	11016	621231	892572	226183.439	17.7871338
1161	1160	Cas9-CP-1160	1	0	0	0	-inf
1164	1163	Cas9-CP-1163	1	0	0	0	-inf
1165	1164	Cas9-CP-1164	19659	3	1	0.30721609	-1.7026743

1169	1168	Cas9-CP-1168	2140	0	1	0.8459557	-0.241346
1170	1169	Cas9-CP-1169	0	0	0		
1171	1170	Cas9-CP-1170	4723	0	1	0.38330408	-1.3834387
1175	1174	Cas9-CP-1174	1310	2	1	3.5342156	1.82139005
1178	1177	Cas9-CP-1177	6065	0	2	0.5969811	-0.7442428
1180	1179	Cas9-CP-1179	984	0	0	0	-inf
1183	1182	Cas9-CP-1182	238	0	0	0	-inf
1184	1183	Cas9-CP-1183	1152	0	1	1.5714802	0.6521241
1185	1184	Cas9-CP-1184	547	0	0	0	-inf
1186	1185	Cas9-CP-1185	5724	0	2	0.63254549	-0.6607589
1188	1187	Cas9-CP-1187	3341	0	0	0	-inf
1189	1188	Cas9-CP-1188	2541	0	0	0	-inf
1192	1191	Cas9-CP-1191	1	0	0	0	-inf
1197	1196	Cas9-CP-1196	536	0	0	0	-inf
1200	1199	Cas9-CP-1199	4011	1	0	0.35146812	-1.5085343
1202	1201	Cas9-CP-1201	595	0	0	0	-inf
1204	1203	Cas9-CP-1203	66	0	0	0	-inf
1207	1206	Cas9-CP-1206	548	0	0	0	-inf
1210	1209	Cas9-CP-1209	6556	1	0	0.2150303	-2.2173882
1211	1210	Cas9-CP-1210	7308	1	0	0.19290348	-2.3740489
1212	1211	Cas9-CP-1211	1236	0	0	0	-inf
1216	1215	Cas9-CP-1215	567	0	1	3.19284866	1.67484417
1218	1217	Cas9-CP-1217	73575	10	5	0.31463285	-1.6682588
1221	1220	Cas9-CP-1220	778	0	0	0	-inf
1228	1227	Cas9-CP-1227	6320	0	2	0.57289405	-0.8036597
1231	1230	Cas9-CP-1230	1031	0	0	0	-inf
1233	1232	Cas9-CP-1232	4971	0	0	0	-inf
1239	1238	Cas9-CP-1238	1	0	0	0	-inf
1240	1239	Cas9-CP-1239	27782	288013	179859	26334.6745	14.684676
1241	1240	Cas9-CP-1240	2252	118092	181711	219999.329	17.7471396
1243	1242	Cas9-CP-1242	4056	234258	335168	231018.819	17.8176509
1246	1245	Cas9-CP-1245	210	241	26334	228635.13	17.8026876
1247	1246	Cas9-CP-1246	4698	208101	289477	173993.468	17.4086736
1248	1247	Cas9-CP-1247	8682	203728	223463	79676.1574	16.2818605
1249	1248	Cas9-CP-1248	3405	0	0	0	-inf

1250	1249	Cas9-CP-1249	11704	789748	1166761	275596.243	18.0721967
1251	1250	Cas9-CP-1250	0	0	1	inf	inf
1253	1252	Cas9-CP-1252	1856	120972	93040	182636.54	17.4786159
1260	1259	Cas9-CP-1259	2519	63555	28113	55772.2004	15.7672586
1261	1260	Cas9-CP-1260	924	30159	10230	66056.427	16.0114113
1262	1261	Cas9-CP-1261	4165	15307	41351	23154.4906	14.4990044
1263	1262	Cas9-CP-1262	19485	54029	107508	13897.5293	13.7625408
1264	1263	Cas9-CP-1263	2	0	0	0	-inf
1265	1264	Cas9-CP-1264	4781	27106	7884	10977.8575	13.4223089
1266	1265	Cas9-CP-1265	189	0	0	0	-inf
1267	1266	Cas9-CP-1266	2	0	1	905.172596	9.8220491
1268	1267	Cas9-CP-1267	13652	0	0	0	-inf
1270	1269	Cas9-CP-1269	1	0	0	0	-inf
1271	1270	Cas9-CP-1270	38856	6	6	0.49723345	-1.0080047
1272	1271	Cas9-CP-1271	1592	0	0	0	-inf
1274	1273	Cas9-CP-1273	40597	8	3	0.41158077	-1.2807525
1275	1274	Cas9-CP-1274	2714	0	2	1.33407899	0.41584409
1278	1277	Cas9-CP-1277	4072	2	0	0.692406	-0.5303099
1279	1278	Cas9-CP-1278	2343	0	0	0	-inf
1280	1279	Cas9-CP-1279	22580	3	1	0.26747392	-1.9025299
1282	1281	Cas9-CP-1281	1	0	0	0	-inf
1283	1282	Cas9-CP-1282	2155	108547	229874	264117.953	18.0108228
1285	1284	Cas9-CP-1284	5455	75691	144994	67679.8748	16.0464393
1288	1287	Cas9-CP-1287	0	0	0		
1292	1291	Cas9-CP-1291	26640	0	3	0.2038677	-2.2942949
1293	1292	Cas9-CP-1292	1534	0	0	0	-inf
1294	1293	Cas9-CP-1293	1861	0	0	0	-inf
1295	1294	Cas9-CP-1294	1	0	0	0	-inf
1296	1295	Cas9-CP-1295	16	0	0	0	-inf
1298	1297	Cas9-CP-1297	18691	126402	193420	28267.6555	14.7868646
1299	1298	Cas9-CP-1298	7634	21460	7566	5757.14732	12.4911384
1304	1303	Cas9-CP-1303	1878	0	0	0	-inf
1307	1306	Cas9-CP-1306	1401	0	0	0	-inf
1319	1318	Cas9-CP-1318	13192	1	1	0.24409368	-2.0344932
1322	1321	Cas9-CP-1321	1	0	0	0	-inf

1332	1331	Cas9-CP-1331	1878	0	0	0	-inf
1335	1334	Cas9-CP-1334	1	0	0	0	-inf
1337	1336	Cas9-CP-1336	7733	90687	74938	34075.8586	15.0564624
1350	1349	Cas9-CP-1349	5134	0	0	0	-inf
1354	1353	Cas9-CP-1353	10052	2	0	0.28048918	-1.833983
1355	1354	Cas9-CP-1354	15338	1	1	0.20994157	-2.2519402
1356	1355	Cas9-CP-1355	6429	0	1	0.28159048	-1.8283295
1359	1358	Cas9-CP-1358	15	0	0	0	-inf
1360	1359	Cas9-CP-1359	0	2	0	inf	inf
1362	1361	Cas9-CP-1361	0	0	0		
1363	1362	Cas9-CP-1362	2	8	3	8354.47229	13.028333
1364	1363	Cas9-CP-1363	0	6	2	inf	inf
1366	1365	Cas9-CP-1365	25037	1921223	3467358	358890.329	18.4531835
1367	1366	Cas9-CP-1366	9962	20	19	6.28300855	2.65145554
1369	1368	Cas9-CP-1368	2135	23150	136062	130657.909	16.9954349

CP libraries, constructed around dCas9, were screened for function in an *E. coli*-based repression (i.e., CRISPRi) assay targeting the expression of either RFP or GFP[25,126,137]. In brief, dCas9-CP libraries were targeted to repress RFP expression while GFP was used as a control for cell viability. Functional dCas9-CP library members were isolated through a sequential double-sorting procedure that enriched functional clones 100-10,000-fold (**Figure 4-1B-C**, **Figure 4-2A-C**, **Table 4-2**). A subset of isolated clones was plated for each of the libraries (i.e., 5, 10, 15 and 20 aa linkers) and sequenced. For the 5 and 10 aa linker library, only a minimal number of CPs around the original termini was observed. However, the 15 and 20 aa linker libraries yielded a number of CP variants and isolated clones were found to be highly functional in bacterial CRISPRi assays (**Figure 4-1E**, **Figure 4-2D**, **Table 4-3**).

Table 4-3. Key Cas9 circular permutants.

Name	Domain at CP Site	Original Sequence at CP Site	New Start Site (aa)
Cas9-CP181	Helical-II	…PDNSD ǀ VDKLF…	181
Cas9-CP199	Helical-II	…QLFEE ǀ NPINA…	199
Cas9-CP230	Helical-II	…LIAQL ǀ PGEKK…	230
Cas9-CP270	Helical-II	…QLSKD ǀ TYDDD…	270
Cas9-CP310	Helical-II	…ILRVN ǀ TEITK…	310
Cas9-CP1010	RuvC-III	…ESEFV ǀ YGDYK…	1010
Cas9-CP1016	RuvC-III	…GDYKV ǀ YDVRK…	1016
Cas9-CP1023	RuvC-III	…VRKMI ǀ AKSEQ…	1023
Cas9-CP1029	RuvC-III	…KSEQE ǀ IGKAT…	1029
Cas9-CP1041	RuvC-III	…YFFYS ǀ NIMNF…	1041
Cas9-CP1247	CTD	…YEKLK ǀ GSPED…	1247
Cas9-CP1249	CTD	…KLKGS ǀ PEDNE…	1249
Cas9-CP1282	CTD	…SKRVI ǀ LADAN…	1282

Because the majority of functional clones were found in the 20 aa linker library, we proceeded to deep sequence this library, to generate an enrichment profile of permutation across Cas9. We identified 77 sites that were highly enriched (>100-fold) following the double sorting procedure (**Figure 4-1C**, **Figure 4-2F**). Notably, all confirmed hits (**Figure 4-1E**) and internal controls fell within this group. Mapping the observed sites onto the protein sequence (**Figure 4-1D**) revealed three hotspots of CPs (all numbering based on Streptococcus pyogenes Cas9 protein sequence): in the Helical-II (aa 178–314), in the RuvC-III (aa 940–1150) and in the CTD (aa 1240–1299) domains (**Figure 4-1D**, **Figure 4-2E-G**). These hotspots qualitatively correspond with those we have previously identified for Cas9 domain insertion[126], indicating that the underlying structural and biochemical constraints may be similar (**Figure 4-2G**). Intriguingly, among the newly discovered termini, a number are in direct contact (< 5 Å) with the non-target strand, yielding Cas9-CPs containing ideal fusion points for protein domains to modify the isolated single-strand that heretofore required long linkers to gain such access (i.e., base editors) [115,117,118,138] (**Figure 4-2E**).

The isolated Cas9-CPs were next tested for their cleavage activity relative to wild-type (WT) Cas9. Briefly, two variants from each of the three hotspots (specifically, CP sites 199, 230, 1010, 1029, 1249, and 1282) were constructed with a 20 aa linker between the original N and C termini and recoded with functional nuclease active sites (**Table 4-3**). Testing of these constructs for genomic cleavage and killing activity in *E. coli* demonstrated that all possessed similar activity as WT Cas9 (**Figure 4-1F**). To assess how well these findings extrapolate to mammalian systems, we established a rapid human genome editing reporter assay with a quantitative fluorescence-based readout of target disruption activity and editing efficiency (**Figure 4-3A-C**, **Materials and Methods**). When compared relative to WT Cas9 in this assay, our Cas9-CPs showed surprisingly high genome editing efficiency (**Figure 4-1G**). While we observed more variation than in the *E. coli*-based experiments, four tested CP variants (CP199, CP1029, CP1249, CP1282) showed 80% or more of WT activity. Overall, these results demonstrate that Cas9 can be circularly permuted to create novel proteins that may maintain wild-type like levels of DNA binding and cleavage activity.

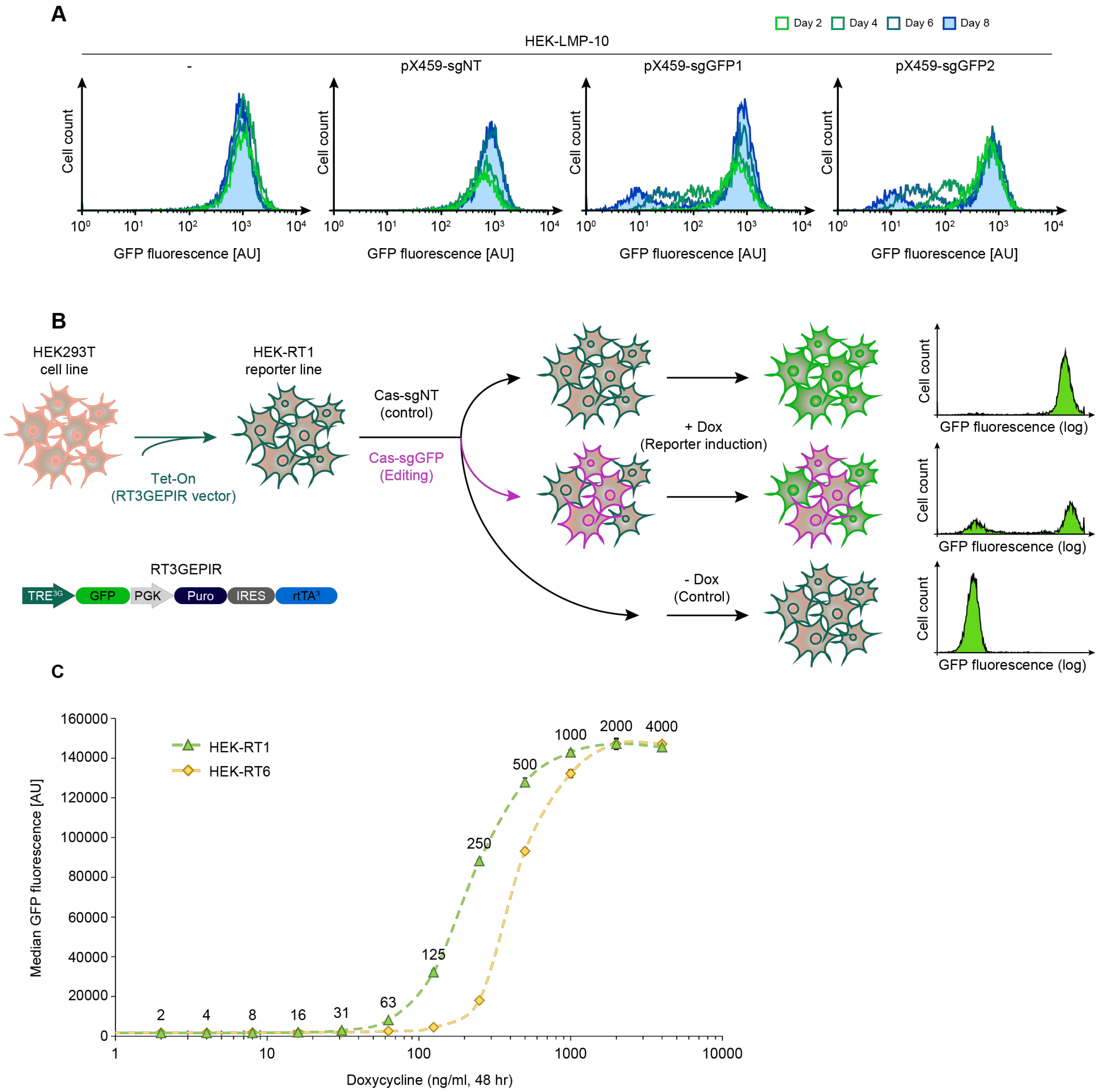


Figure 4-3. Mammalian genome editing reporter cell lines.

(**A**) Flow cytometry time course of GFP fluorescence decay after editing. Monoclonal HEK-LMP-10 reporter cells stably expressing GFP were transfected with a vector (pX459) expressing wild-type Cas9 and the indicated sgRNAs targeting the reporter, or a negative control (sgNT). Note, full fluorescence decay after transfection of editing reagents took up to eight days. (**B**) Schematic showing the concept of a rapid mammalian genome editing reporter assay. Monoclonal reporter cell lines were established by stably integrating and all-in-one Tet-On cassette enabling doxycycline-inducible GFP expression, followed by selection and characterization of single clones. To assess editing efficiency of novel variants, reporter cells are transduced with Cas constructs of interest and guide RNAs targeting GFP, or a non-targeting control. At 24+ hours post-transduction, the GFP fluorescence reporter is induced by doxycycline treatment for 24-48 hr and genome editing quantified by flow cytometry. (**C**) Activation curves (doxycycline titration) of two monoclonal genome editing reporter cell lines. HEK-RT1 and HEK-RT6 reporter cell lines were treated with the indicated doxycycline concentrations for 48 hours. The median GFP fluorescence intensity was quantified by flow cytometry and normalized to parental HEK293T cells. Both cell lines show full reporter induction at 1000-2000 ng/ml doxycycline and similar EC50 values (HEK-RT1: 214.5 ± 2.3 ng/ml; HEK-RT6: 433.0 ± 9.5 ng/ml).

4.3.2. Cas9-CP activity can be regulated by proteolytic cleavage

Characterization of the libraries described above revealed that circular permutation is highly sensitive to the linker length connecting the original N and C terminus. PCR analysis of pooled libraries (**Figure 4-4A**) indicated that a linker length of 5 aa or 10 aa was not sufficient to generate Cas9-CP diversity. Conversely, libraries of 15 or 20 aa linkers (**Figure 4-4A**) qualitatively possessed extensive permutable diversity. We therefore tested the importance of linker length on confirmed sites identified above (**Figure 4-1E**). The same six Cas9-CPs (i.e., Cas9-CP199 through Cas9-CP1282) were cloned with linkers (GGS repeats) from 5 to 30 aa and tested for repression of GFP in an *E. coli*-based CRISPRi assay (**Figure 4-5A**). In agreement with the pooled libraries, we found that all Cas9-CPs with linkers of 5 and 10 aa in length were markedly disrupted in activity, while those with longer linkers were active. Notably, activity did not increase with linker length beyond 15 aa (**Figure 4-5A**).

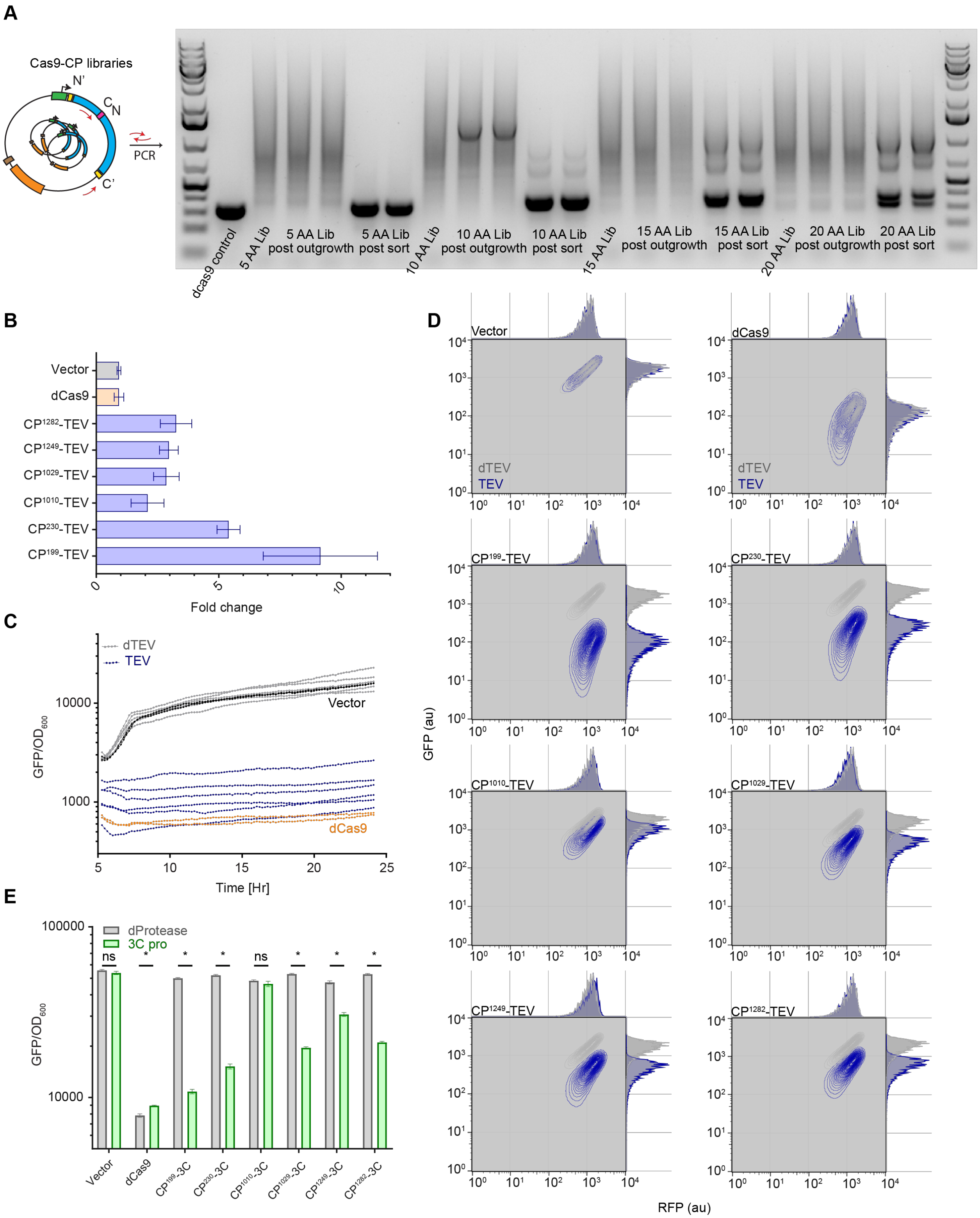


Figure 4-4. CP linker length and activation.

(**A**) Schematic of the PCR system and uncropped gel of the PCRs for each library, in biological replicate, pre and post sorting. (**B**) Fold changes of the TEV based activation of

CP-TEV linker clones from Figure 2C. (**C**) Time course values from the CRISPRi assay in Figure 2C, demonstrating constancy of activity for clones with TEV (blue) versus dTEV (gray). (**D**) Single cell analysis of Cas9-CP-TEV linkers. (**E**) Endpoint analysis of an *E. coli* CRISPRi based GFP expression assay with all six Cas9-CPs containing a 8 aa 3C linker (LEVLFQ/GP) in the presence of a functional 3C protease (3C pro, green) or a deactivated TEV protease with a catalytic triad mutant C151A (dProtease, gray).

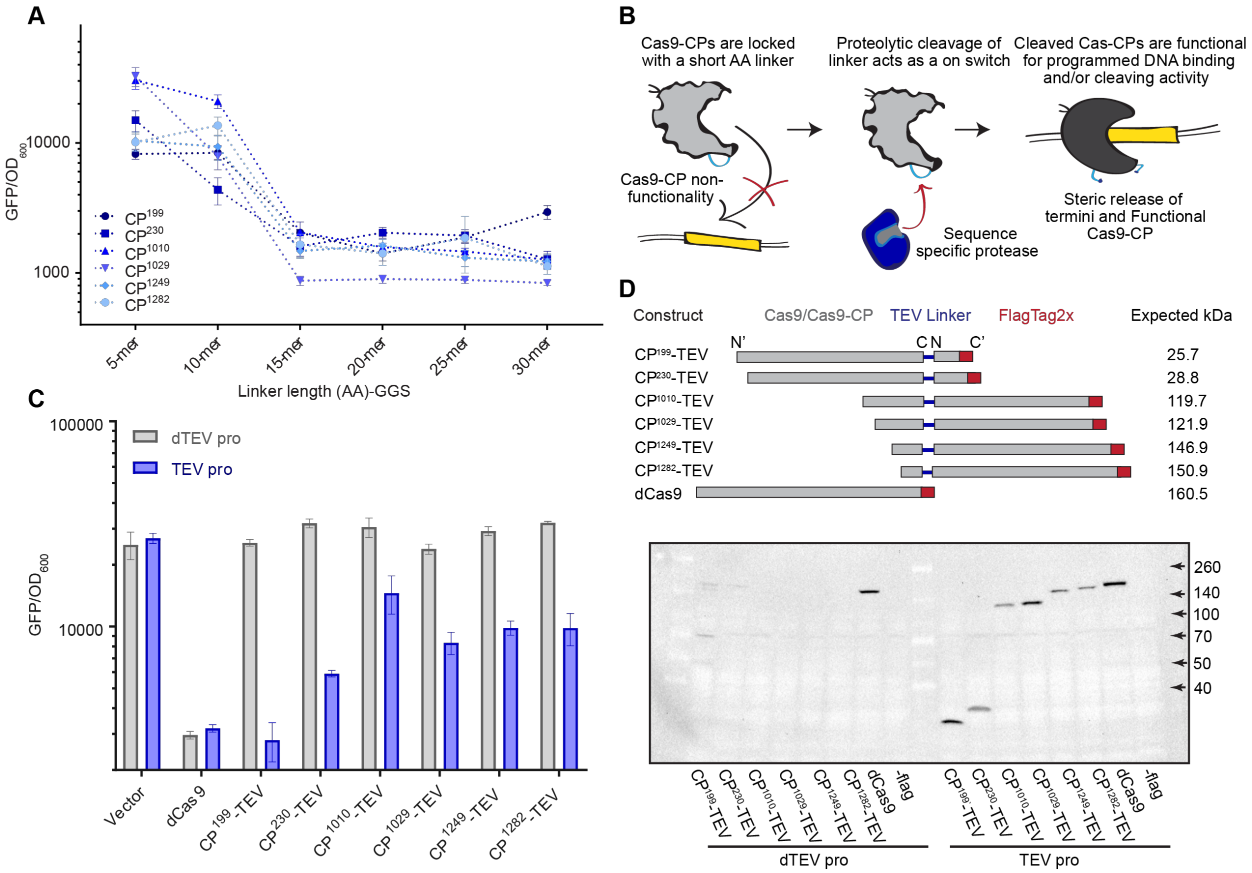


Figure 4-5. Linker length can be utilized to control Cas9-CP activity.

(**A**) Endpoint analysis of an *E. coli* CRISPRi-based GFP repression assay run in triplicate using Cas9-CPs identified as functional with 20 aa linkers, evaluated with GGSn linkers of length 5, 10, 15, 20, 25, and 30 aa. Error bars represent standard deviation in all panels. (**B**) Schematic describing the rationale behind using a Cas9-CP with a short aa linker as a "caged" molecule. (**C**) Endpoint analysis of an *E. coli* CRISPRi-based GFP expression time course with all six Cas9-CPs containing a 7 aa TEV linker (ENLYFQ/S) in the presence of a functional TEV protease (TEV, blue) compared with deactivated TEV protease with the catalytic triad mutant C151A (dTEV, gray) (n = 3, error bars represent SD; $*p < 0.05$; ns, not significant, t test).

(D) Schematic and western blot against the Flag epitope on the C terminus of the CP-TEVs after the endpoint measurement (Figure 2C). Expected kilodaltons indicate the predicted band size if cleavages occur at the TEV site in the CP linker region.

The sensitivity of CPs to linker length led us to hypothesize that Cas9-CPs could be made into "caged" variants that could switch from an inactive form to an active one upon post-translational modification (**Figure 4-5B**). It has previously been observed that circularly permuted proteins can be sensitive to the length of the linker between their old N and C termini[129]. This requirement has been exploited to create zymogen pro-enzymes by replacing the linker with a site-specific protease sequence, such that proteolytic cleavage converts a short linker into an effectively infinite linker with concomitant turn-on in protein activity. Although potentially useful for applications in biosensing (e.g., pathogen or cancer detection) existing sensors were constructed around either RNase A[134,139] or barnase[140] and possess limited in vivo potential because of their inherent nonspecific, toxic activity.

To test the possibility of turning Cas9-CPs into activatable switches using a well-studied protease, we engineered the six representative CP variants with the 7 aa cleavage site (ENLYFQ/S) of the tobacco etch virus (TEV) nuclear inclusion antigen (NIa) protease as the linker sequence[141]. We found that this 7 aa linker was able to fully disrupt Cas9-CP activity in our E. coli CRISPRi GFP repression assay (**Figure 4-5C**, **Figure 4-4B**). Upon addition of a fully active TEV protease, activity was restored to a varying degree in all six Cas9-CPTEV constructs. Notably, Cas9-CP199 switched from completely off to fully on (**Figure 4-5C**) and performed consistently over a 20-hr time course (**Figure 4-4C**). This switch behaved well across the population in single cell assays and did not activate when a TEV catalytic triad mutant, C151A, was expressed (dTEV) (**Figure 4-4D**). Finally, to verify if TEV is cleaving Cas9-CPs at the CP linker, we recovered cells from the endpoint of the CRISPRi assay (**Figure 4-5C**) for western blot analysis against a 2x Flag-tag cloned onto the C terminus of the protein. As expected, when an active TEV protease was present, products were observed corresponding to the size of the C-terminal circularly permuted fragment (**Figure 4-5D**).

4.3.3. ProCas9s exemplify a general strategy for regulating caged Cas9's with site-specific proteases

Next, we sought to determine whether this uncaging mechanism could be generalized to other families of proteases. For example, the human rhinovirus 3C is responsible for about 30% of cases of the common cold and contains a well-studied protease (3Cpro), unrelated to that from TEV[142]. Thus, we replaced the TEV linker sequence with the 8 aa recognition site for 3Cpro (LEVLFQ/GP) in the six Cas9-CPs and tested for bacterial CRISPRi activity with and without active protease. Protease-dependent activation of Cas9-CPs was observed, with varying amounts of turn-on in activity, thus demonstrating that the mechanism can be extended to other proteases. Cas9-CP199 again had the greatest response (**Figure 4-4E**) and was used for all experiments described below.

Next, we sought to apply our protease sensing Cas9-CPs (hereafter ProCas9s) to agriculturally and medically relevant viruses. We examined the Potyvirus proteases from turnip mosaic virus (TuMV), plum pox virus (PPV), potato virus Y (PVY), and cassava brown streak virus (CBSV), all of which are plant viruses responsible for significant crop losses each year[141,143]. The NIa protease genes from these viruses were cloned for co-expression in conjunction with our ProCas9s. Cognate protease cleavage sites (**Materials and Methods**) were used as the CP linker in Cas9-CP199, yielding the respective ProCas9s that were systematically tested against all co-expressed Nla proteases (**Figure 4-6A**, controls in **Figure 4-7AB**). CRISPRi experiments revealed a general trend of proteases activating their respective ProCas9 and also that the PPV linker (QVVVHQ/SK) enabled a ProCas9 response to three different Nla proteases with distinct specificity from TEV (**Figure 4-6AB**). We term this variant ProCas9Poty for a Cas9 that can recognize and respond to a number of agriculturally important Potyvirus proteases.

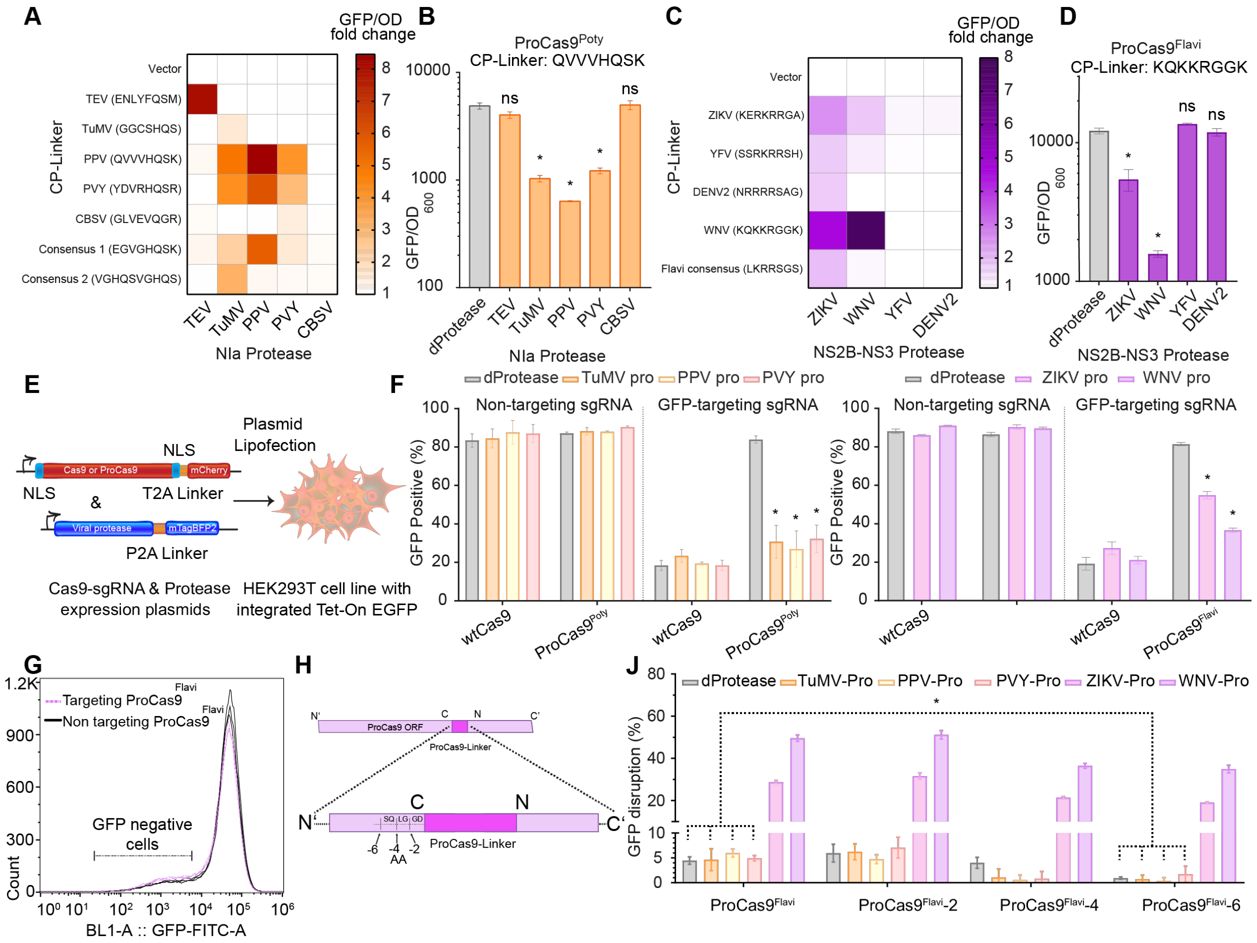


Figure 4-6. Generation of ProCas9s for sensing and responding to *Potyvirus* and *Flavivirus* proteases.

(**A**) Heatmap depicting the fold activation of a suite of ProCas9 CP linkers for Potyviral NIa proteases. Data are normalized to a non-active protein expression control (dTEV) in an *E. coli*-based CRISPRi GFP repression assay. Darker coloration indicates greater activity (n = 2). (**B**) Endpoint analysis of the *E. coli* CRISPRi assay utilizing the linker derived from Plum Pox virus (PPV) comparing the response to distinct NIa proteases and a dead protease (n = 3, error bars represent SD; $*p < 0.05$; ns, not significant, t test compared to dProtease). (**C**) Heatmap depicting the fold activation of a suite of ProCas9 CP linkers for Flavivirus NS2B-NS3 proteases, normalized to a non-active protein expression control (dTEV) in an *E. coli*-based CRISPRi GFP repression assay. Darker coloration indicates greater activity (n = 2). (**D**) Endpoint analysis of the *E. coli* CRISPRi assay utilizing the linker derived from West Nile virus (WNV) showing the response to distinct NS2B-NS3 proteases and a dead protease (n = 3, error bars represent SD; $*p < 0.05$; ns, not significant; t test compared to dProtease). (**E**) Schematic of the constructs used for the transient transfection and testing in HEK293T cells. (**F**) Mammalian GFP disruption assay (**Figure 4-3A-C**). HEK293T-based reporter cells were transfected with the indicated sgRNAs, WT Cas9, or a ProCas9 variant and the respective proteases. Reduction in GFP-positive cells indicates genome cleavage by a Cas9 construct (n = 3; error bars represent SD; $*p < 0.05$, t test compared to dProtease). (**G**) Flow cytometry plots from (F) with overlay of GFP-targeting (pink) versus non-targeting (black) ProCas9^{Flavi} systems, demonstrating a small degree of background activity. (**H**) Truncation of the ProCas9 aa linker sequence to prevent leakiness. (**I**) Leakiness and orthogonality of the original and shortened ProCas9^{Flavi} constructs. Displayed as a percentage of GFP disrupted via normalization to the non-targeting guide for each construct-protease pairing. In addition to the deactivated protease (dProtease) control, the active *Potyvirus* NIa proteases were used to assess orthogonality (n = 3; error bars represent SD; $*p < 0.05$; ns, not significant, t test).

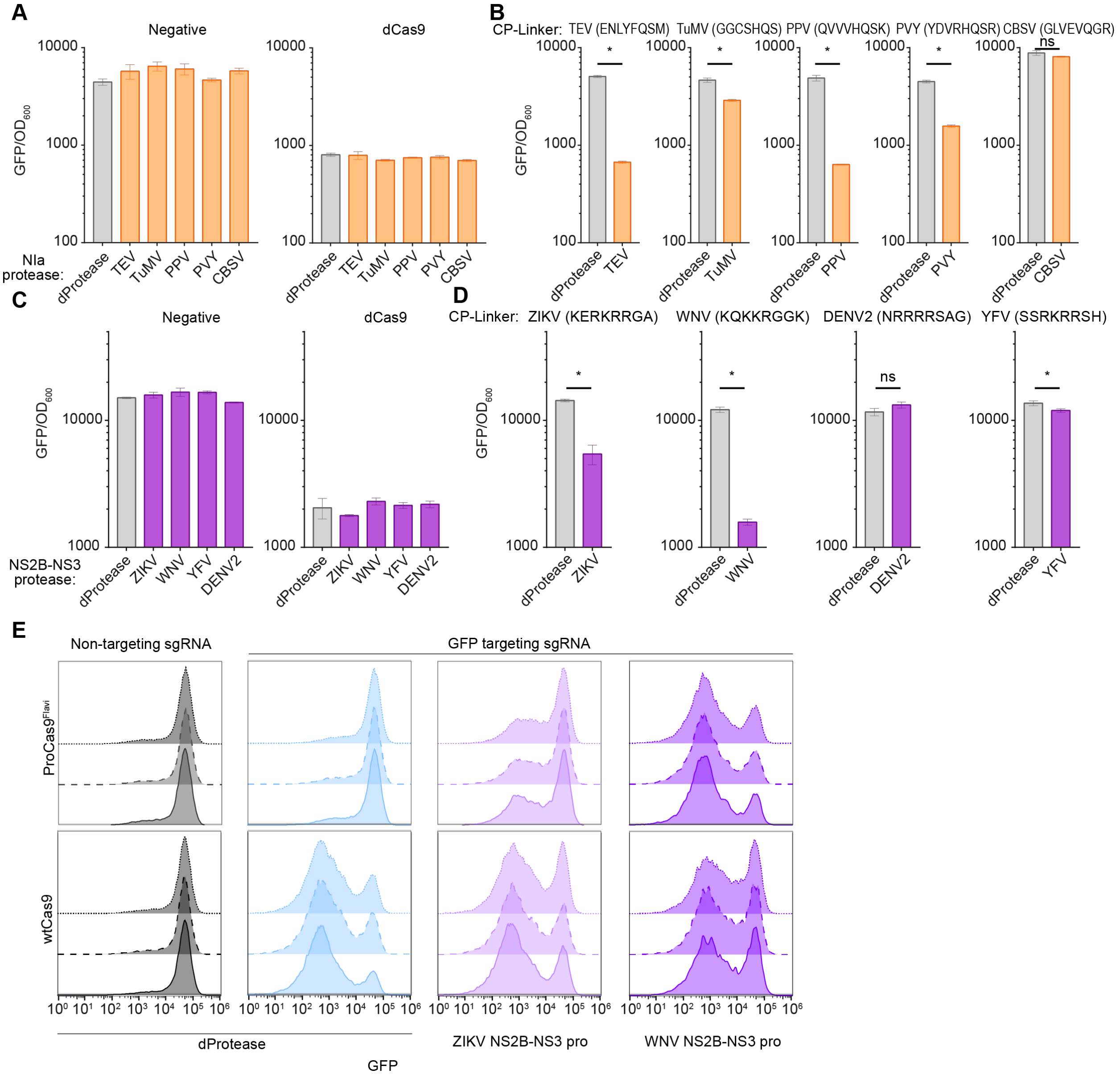


Figure 4-7. ProCas9 specificity assessment.

(**A**) Endpoint analysis of an *E. coli* CRISPRi based GFP expression assay with negative and positive controls in the presence of all NIa proteases to determine if any protease changes the GFP expression levels. (**B**) Endpoint analysis of an *E. coli* CRISPRi based GFP expression assay for each Cas9-CP-Potyviral linker against its respective protease. Significance was assessed by comparing each sample to its respective dProtease control (unpaired, two-tailed t test, n = 3, ∗ = $p < 0.05$, ns = not significant). (**C**) Endpoint analysis of an *E. coli* CRISPRi based GFP expression assay with negative and positive controls in the presence of all Flavirus NS2B-NS3 proteases to determine if any protease changes the GFP expression levels. (**D**) Endpoint analysis of an *E. coli* CRISPRi based GFP expression assay for each Cas9-CP-Flaviviral linker against its respective protease. Significance was assessed by comparing each sample to its respective dProtease control (unpaired, two-tailed t test, n = 3, ∗ = $p < 0.05$, ns = not significant). (**E**) Raw Flow cytometry plots from **Figure 4-6F** demonstrating the always on nature of WT Cas9 and the activation of ProCas9Flavi in the presence of Flavivirus proteases.

We repeated this process with a set of proteases from the medically important Flavivirus genus. Briefly, the capsid protein C cleavage sequences from Zika virus (ZIKV), West Nile virus (WNV, Kunjin strain), Dengue virus 2 (DENV2), and yellow fever virus (YFV) [144,145] were used as the CP linker sequence to generate a set of flavivirus-specific ProCas9s. In the viral life cycle, these cleavage sequences are cut by the NS2B-NS3 protease from the respective virus to mature the polyprotein[145]. Screening of these Flavivirus ProCas9 variants against their cognate proteases revealed a variant—hereafter called ProCas9Flavi—that possesses a WNV linker sequence (KQKKR/GGK) and was activated by NS2B-NS3 proteases from both Zika and WNV (**Figure 4-6CD**, **Figure 4-7CD**). We did not observe any activation with the CBSV, DENV2, or YFV proteases; this may be due to non-optimal CP linkers, poor expression of the cognate proteases, or a steric hindrance blocking the protease from reaching the CP linker site.

Next, we validated and optimized the function of ProCas9s in eukaryotic cells using a transient transfection system in our HEK293T-based GFP disruption assay (**Figure 4-6E**, **Figure 4-3BC**). In this model, expression of either ProCas9Poty or ProCas9Flavi resulted in GFP disruption only in the presence of the active proteases (**Figure 4-6F**, **Figure 4-7E**). Nevertheless, we also observed a small amount of leaky activation (~5%) in the absence of protease activity (**Figure 4-6FG**). Hence, we tested whether progressively shortening the distance between the original N and C termini by 2, 4, or 6 aa would reduce unwanted background activity (**Figure 4-6H**). While removing 2 aa from ProCas9Flavi had no apparent effect, removing six aa (ProCas9Flavi-S6) significantly reduced activity levels for non-active or non-corresponding active proteases while still enabling a response, albeit weaker, to both ZIKV and WNV (corresponding) proteases (**Figure 4-6FJ**). Thus, linker "tightening" optimization provides an additional safety mechanism, allowing a ProCas9 to exist in cells with little risk of untriggered genome cleavage activity.

4.3.4. ProCas9 can be stably integrated into mammalian genomes without leaky activity

A prerequisite for using activatable genome editors in sensing or molecular recording applications is that they possess low background activity under stable expression conditions. To confirm that ProCas9s function accordingly, we built lentiviral vectors expressing ProCas9 from either a weak EF1a core promoter (EFS) or strong full-length EF1a promoter, along with single guide RNAs (sgRNAs) driven from a U6 promoter, and tested ProCas9Flavi and ProCas9Flavi-S6 activity in HEK-RT1 reporter cells (**Figure 4-8A**). When measured 6 to 10 days post-transduction, none of the four tested ProCas9 constructs showed any background activity (**Figure 4-8B**), indicating that the systems are not leaky.

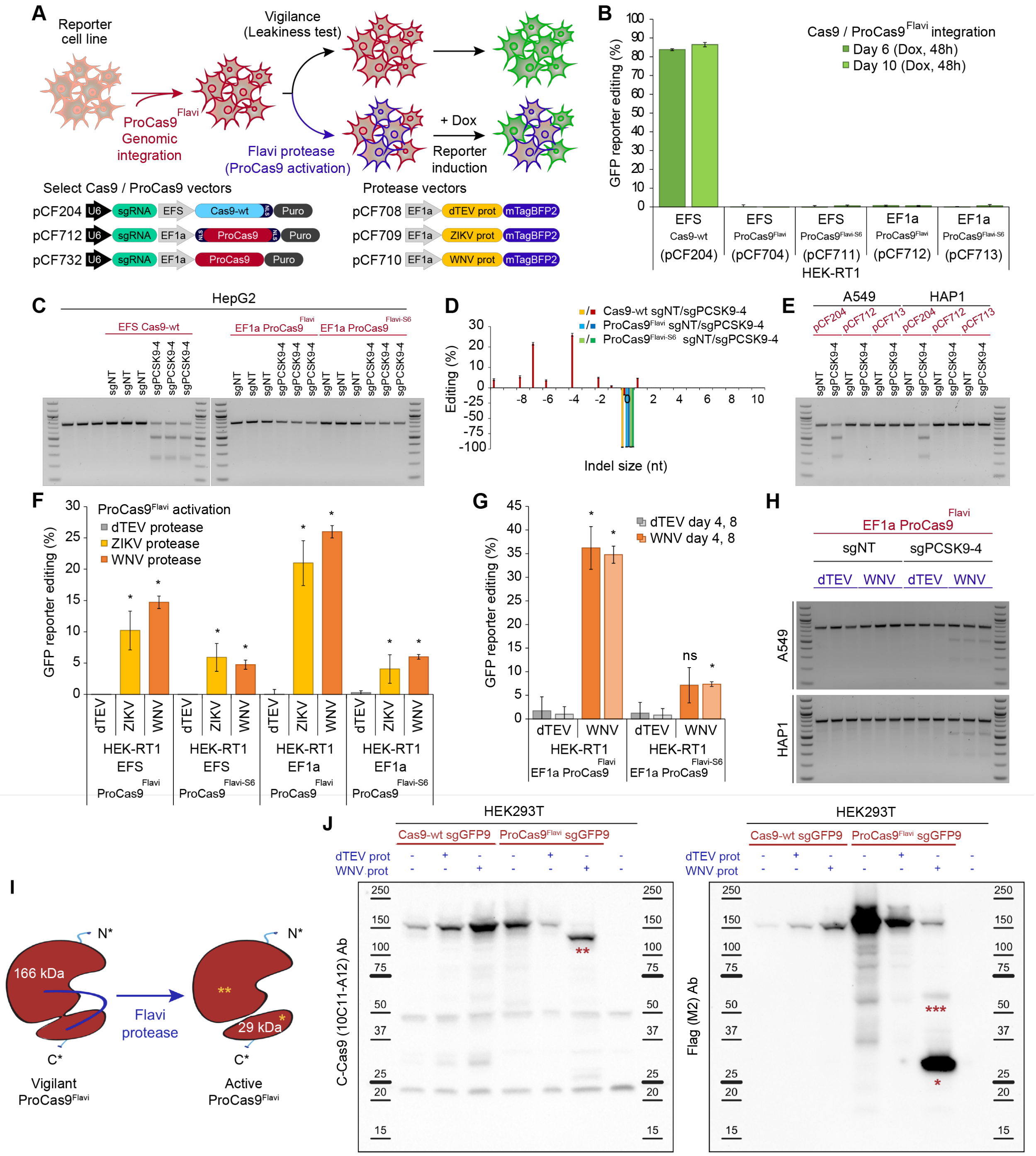


Figure 4-8. ProCas9 stably integrated into mammalian genomes can sense and respond to *Flavivirus* proteases.

(**A**) Genomic integration and testing of *Flavivirus* protease-sensitive ProCas9s. HEK-RT1 genome editing reporter cells are stably transduced with various ProCas9 lentiviral vectors, followed by puromycin selection of ProCas9 cell lines. These cell lines are then either tested for leaky ProCas9 activity in the absence of a stimulus or stably transduced with a vector expressing the indicated proteases, followed by assessment of genome editing using the GFP reporter. (**B**) Leakiness assessment of ProCas9 variants expressed from either the EFS or EF1a promoter. HEK-RT1 reporter cells were stably transduced with the indicated ProCas9 variants or Cas9 WT. Genome editing activity was quantified at the indicated days post-transduction. Error bars represent the standard deviation of triplicates. (**C**) Leakiness assessment at the endogenous *PCSK9* locus. HepG2 cells stably transduced with the indicated sgRNAs and ProCas9 variants or Cas9 WT. Cells were selected on puromycin and harvested at day 8 post-transduction for T7E1 analysis. (**D**) Mutational patterns and editing efficiency at the PCSK9 locus of samples shown in (C). Indels were quantified using TIDE. For clarity, the fraction of non-edited cells is represented as negative percentages. (**E**) ProCas9 leakiness quantification, like in (C), in A549 and HAP1 cells. Cells were selected on puromycin and harvested at day 7 post-transduction for T7E1 analysis. (**F**) Quantification of *Flavivirus* ProCas9 activation in response to various control (dTEV, pCF708) or *Flavivirus* (ZIKV, pCF709; WNV, pCF710) proteases. ProCas9 reporter cell lines were stably transduced with the indicated protease vectors. At day 3 post-transduction, cells were treated with doxycycline to induce GFP reporter expression. Error bars represent the standard deviation of triplicates. Significance was assessed by comparing each sample to its respective dTEV control (unpaired, two-tailed t test, n = 3, $*p < 0.05$; ns, not significant). (**G**) Genome editing activity in Flavivirus ProCas9 reporter cell lines, like in (F), at day 4 or 8 post-transduction. (**H**) Protease-sensitive editing at the endogenous *PCSK9* locus. T7E1 assay of A549 and HAP1 *Flavivirus* ProCas9 cell lines (sgNT, sgPCSK9-4) stably transduced with the indicated mTagBFP2-tagged viral proteases. At day 4 post-transduction, mTagBFP2-positive cells were sorted and harvested for T7E1 analysis. (**I**) ProCas9^{Flavi} activation by *Flavivirus* (Flavi) proteases. ∗, small subunit of the activated ProCas9^{Flavi} (29 kDa). ∗∗, large subunit of the activated ProCas9^{Flavi} (137 kDa). (**J**) Immunoblotting for Cas9 in HEK293T co-transfected with plasmids expressing Cas9 WT or ProCas9^{Flavi} and dTEV or WNV proteases. The C-Cas9 (clone 10C11-A12) antibody recognizes the large subunit of the activated ProCas9^{Flavi} (∗∗ 137 kDa). The Flag-tag (clone M2) antibody recognizes the small subunit of the activated ProCas9^{Flavi} (∗ 29 kDa). ∗∗∗, likely small-subunit-ProCas9^{Flavi}-T2A-mCherry (55 kDa). Protein ladders indicate reference molecular weight markers.

To further confirm these findings at an endogenous locus, we targeted the non-essential PCSK9 locus in the hepatocellular carcinoma cell line HepG2. Eight days after stable transduction with ProCas9Flavi, ProCas9Flavi-S6 or WT Cas9 PCSK9 editing efficiency was assessed by T7 endonuclease 1 (T7E1) assay (**Figure 4-8C**). While WT Cas9 showed high levels of editing, no leakiness was observed with any of the ProCas9 constructs. TIDE analysis[146] was used to quantify editing outcome (**Figure 4-8D**), revealing 71.1% editing with WT Cas9 (11.6% non-edited, 17.3% undetected in the −10- to +10-nt indel range) and confirming the absence of background editing with the ProCas9 constructs. Finally, editing at the PCSK9 locus was also tested in the lung carcinoma cell line A549 and the haploid chronic myeloid leukemia derived line HAP1, two cell lines often used for Flavivirus assays (**Figure 4-8E**). Again, the ProCas9 constructs displayed no background activity.

4.3.5. Genomic ProCas9 can be activated by *Flavivirus* proteases to induce target editing

An activatable switch for molecular sensing must display repeatable induction upon stimulation. In an initial test, HEK-RT1 reporter lines (**Figure 4-8B**) containing stably integrated Flavivirus ProCas9s were transiently transfected with vectors expressing dTEV, ZIKV, and WNV proteases, each tagged with mTagBFP2 to enable tracking of activity (**Figure 4-8A**, **Figure 4-9A**). Two days post-transfection, the GFP reporter was induced by doxycycline treatment for 24 hr and quantified for editing efficiency by flow cytometry in mTagBFP2-positive cells (**Figure 4-9B**). While dTEV protease expression did not lead to genome editing in any reporter cell line, both ZIKV and WNV protease activity led to genome editing, especially with the ProCas9Flavi system. Not surprisingly, the ProCas9Flavi system driven by the stronger EF1a promoter showed the highest genome editing efficiency (**Figure 4-8F, Figure 4-9A**). Together, this suggests that ProCas9 constructs can sense and record Flavivirus protease activity associated with transient expression.

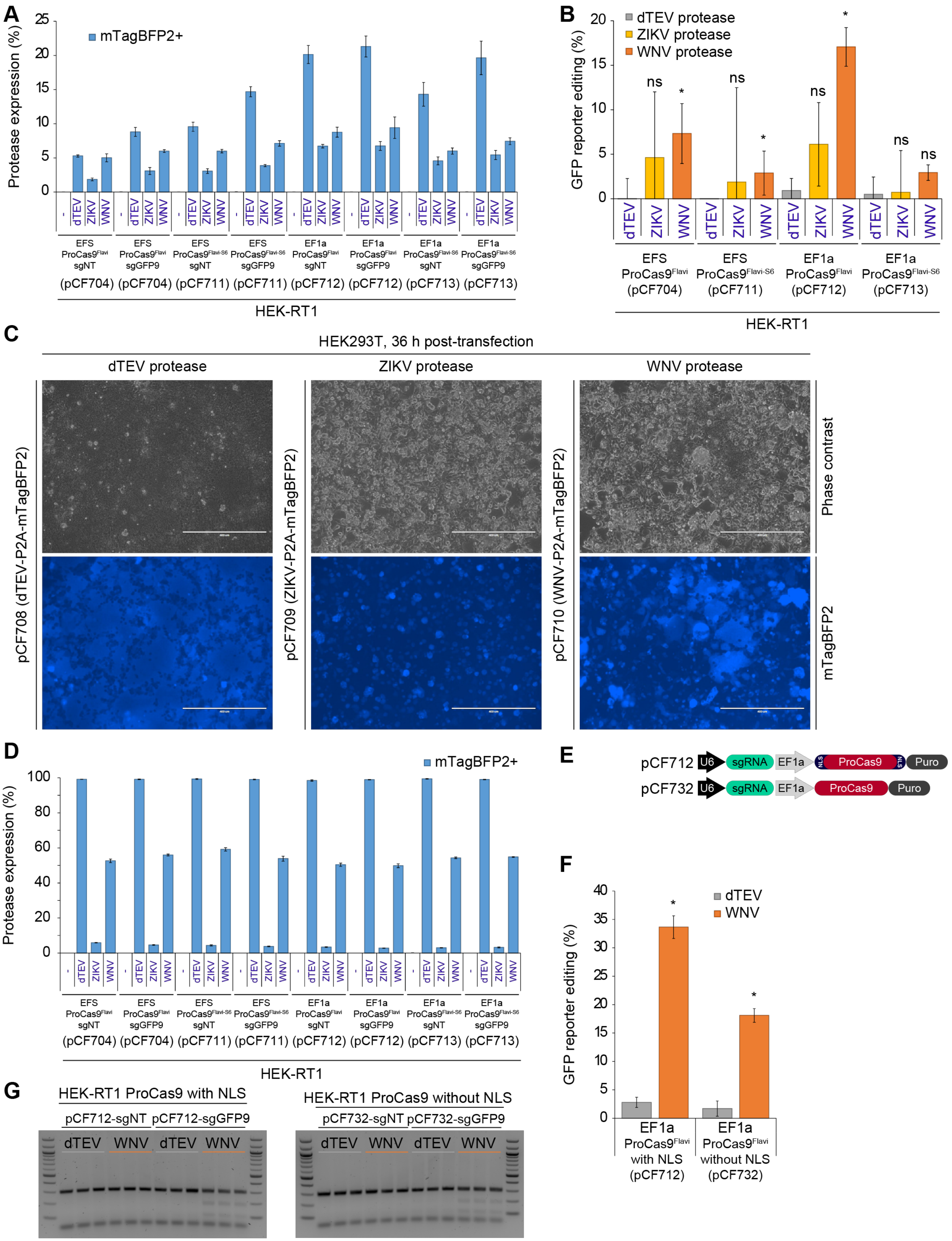


Figure 4-9. ProCas9 activation by Flavivirus proteases.

(**A**) Fluorescence analysis of the indicated HEK-RT1 based cell lines stably expressing a ProCas9 variant and an sgRNA targeting the reporter (sgGFP9) or a non-targeting control (sgNT). All cell lines were either non-transfected or transfected with vectors expressing the dTEV (pCF708), ZIKV (pCF709) or WNV (pCF710) protease. The percentage mTagBFP2+ cells was measured three days post-transfection along with the median fluorescence intensity (MFI) of the mTagBFP2+ cells. AU, arbitrary units. Error bars indicate the standard deviation of triplicates. (**B**) Activation of *Flavivirus* ProCas9 by transfection of various proteases. ProCas9 cell lines were transiently transfected to express the indicated mTagBFP2-tagged viral proteases. At day 2 post-transfection, cells were treated with doxycycline for 24 hr to induce GFP reporter expression. GFP fluorescence was quantified in mTagBFP2-positive cells, for samples expressing either a non-targeting guide (sgNT) or sgGFP9 targeting the reporter. Editing efficiency is reported as the normalized difference between the two in each case. Error bars indicate the standard deviation of triplicates. Significance was assessed by comparing each sample to its respective dTEV control (unpaired, two-tailed t test, $n = 3$, $* = p < 0.05$, ns = not significant). (**C**) Fluorescence imaging of mTagBFP2 in HEK293T cells 36 hr after transfection of the indicated lentiviral plasmids expressing viral proteases. Lentiviral helper plasmids were co-transfected in each case. Scale bar: 400 μm. (**D**) Fluorescence analysis of the indicated HEK-RT1-ProCas9 reporter cell lines expressing an sgRNA targeting the reporter (sgGFP9) or a non-targeting control (sgNT). All cell lines were either non-transduced or stably transduced with vectors expressing the dTEV (pCF708), ZIKV (pCF709) or WNV (pCF710) protease. The percentage mTagBFP2+ cells was quantified four days post-transduction along with the median fluorescence intensity (MFI) of the mTagBFP2+ cells. AU, arbitrary units. Error bars indicate the standard deviation of triplicates. (**E**) Schematic vector maps. (**F**) Activity comparison of *Flavivirus* ProCas9 constructs with and without nuclear localization sequences (NLSs). Genome editing efficiency was assessed in the indicated HEK-RT1-ProCas9 reporter cell lines at day 4 post-transduction of the indicated proteases, followed by 24 hr of GFP reporter induction. Error bars indicate the standard deviation of triplicates. Significance was assessed by comparing each sample to its respective dTEV control (unpaired, two-tailed t test, $n = 3$, $* = p < 0.05$, ns = not significant). (**G**) T7E1 assay of samples shown in (F). Note that while the flow cytometry-based editing quantification was based on cells expressing the respective proteases (mTagBFP2+), the T7E1 assay is based on the total population of cells.

To mimic a viral infection more closely, we next evaluated whether a stably integrated viral vector expressing Flavivirus proteases could also activate ProCas9Flavi enzymes. To generate viral particles, HEK293T packaging cell lines were transfected with dTEV, ZIKV, or WNV protease-encoding lentiviral vectors (**Figure 4-9C**). Expressing the NS2B-NS3 or NS3 protease is known to be toxic[147], and we observed a similar effect with ZIKV and WNV proteases, which led to reduced viral titers and target cell transduction efficiency (**Figure 4-9D**). Nevertheless, we were able to stably transduce the HEK-RT1-ProCas9 reporter cell lines with protease constructs and followed the effects of dTEV, ZIKV, and WNV protease expression (**Figure 4-8F**). While the dTEV protease did not lead to any editing, both the ZIKV and WNV proteases induced genome editing in all four tested ProCas9 lines, with the strongest effect (over 25% editing) again observed with the EF1a-ProCas9Flavi system induced by the WNV protease.

To assess the dynamic range of ProCas9Flavi induction, we repeated the above experiments out to 8 days (**Figure 4-8G**). Here, stable expression of the WNV protease led to ~35% genome editing when sensed by the EF1a-ProCas9Flavi system. In further tests, we tested an EF1a-ProCas9Flavi construct that did not contain any nuclear localization sequence (NLS) (**Figure 4-9E**) and observed that WNV protease-mediated induction was reduced compared to NLS containing constructs (**Figure 4-9F**). Finally, we qualitatively confirmed these results, based on mTagBFP2-positive cells expressing the protease, using a T7E1 assay (**Figure 4-9G**).

As with background activity testing, the activation of ProCas9s by proteases was further validated by targeting the endogenous PCSK9 locus (**Figure 4-8H**). Qualitative T7E1-based analysis showed that while no genome editing was observed with a non-targeting guide, the EF1a-ProCas9Flavi system equipped with a guide targeting PCSK9 (sgPCSK9-4) showed clear genome editing in the presence of WNV protease, but not a negative control (dTEV). Together with the absence of leakiness, this clearly demonstrates that ProCas9 can be stably integrated into mammalian genomes to sense, record and respond to endogenous or exogenous protease activity.

4.3.6. Mechanism of ProCas9 activation in mammalian cells

Conceptually, the underlying idea of ProCas9s is that they are present in cells in an inactive, or "vigilant," state due to the linker sterically inhibiting activity (**Figure 4-8I**). The presence of a cognate protease recognizing the peptide linker relieves inhibition through target cleavage, and leads to an "active" ProCas9 composed of two distinct subunits. To explore this hypothesis, we co-transfected HEK239T cells with vectors expressing either Cas9 WT or ProCas9Flavi and the dTEV or WNV protease (**Figure 4-10A**). Immunoblotting with antibodies for the full-length Cas9 WT and vigilant ProCas9Flavi—as well as both the small (~29 kDa) and large (~137 kDa) subunit of active ProCas9Flavi—showed that Cas9 WT and ProCas9Flavi are expressed to comparable extents in the absence of a cognate protease (**Figure 4-8J**, **Figure 4-10B**). In the presence of the WNV protease, however, the vast majority of vigilant ProCas9Flavi was activated and observed as two distinct subunits, confirming the hypothesized mechanism.

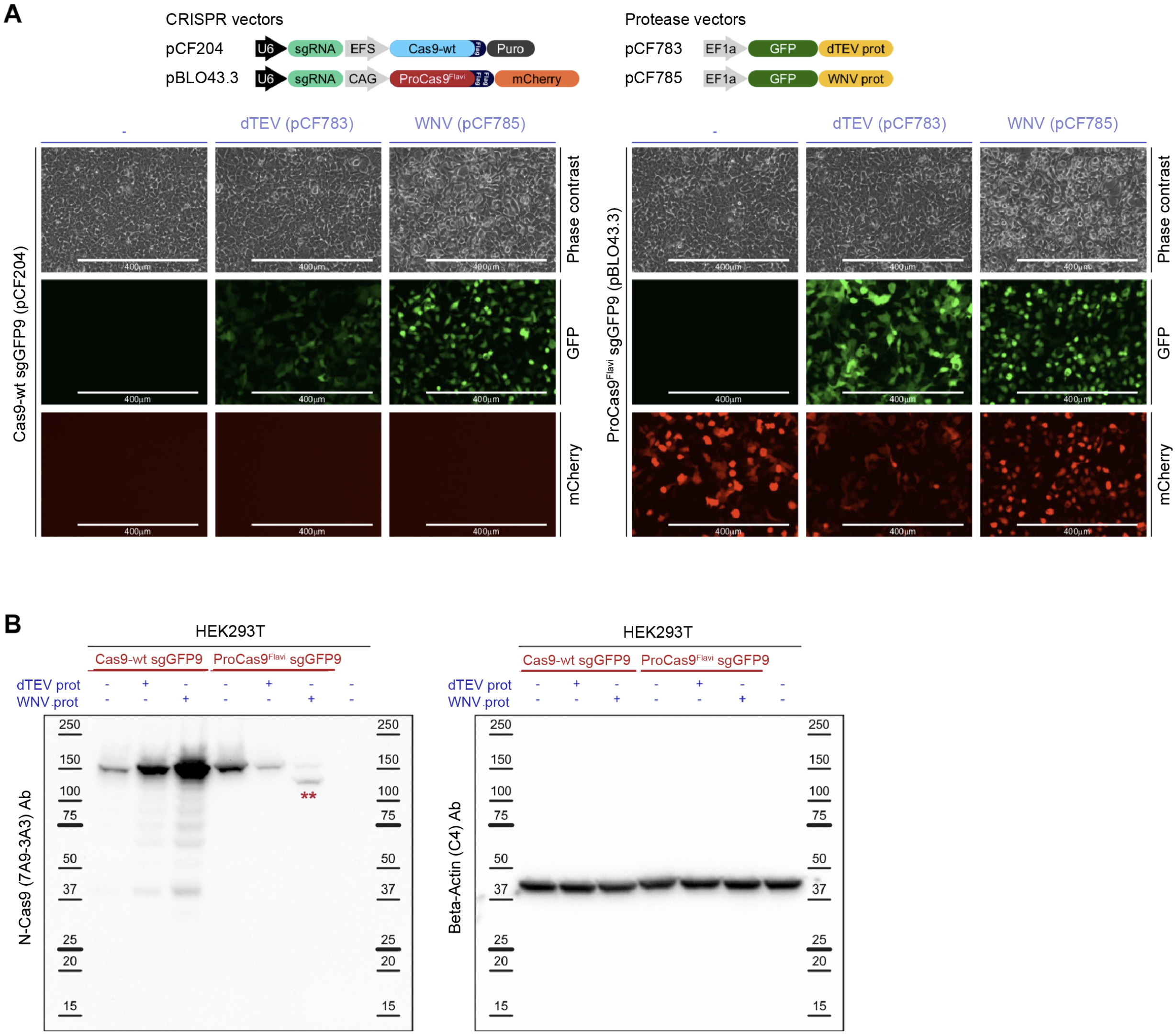


Figure 4-10. Mechanism of ProCas9 activation.

(**A**) Phase contrast and fluorescence imaging in HEK293T cells 36 hr after co-transfection of the indicated plasmids expressing Cas9-wt (pCF204-sgGFP9) or ProCas9^{Flavi} (pBLO43.3-sgGFP9) and plasmids expressing the dTEV (pCF783) or WNV (pCF785) proteases. Scale bars: 400 μ m. (**B**) Immunoblotting for Cas9 in HEK293T co-transfected with the indicated plasmids expressing Cas9-wt or ProCas9^{Flavi} (including sgGFP9) and plasmids expressing the dTEV or WNV proteases. The N-Cas9 (clone 7A9-3A3) antibody recognize the large subunit of the activated ProCas9^{Flavi} (∗∗, 137 kDa). Beta-actin (ACTB, 42 kDa) was used as loading control. Protein ladders indicate reference molecular weight markers.

4.3.7. Rapid CRISPR-Cas-controlled cell depletion

A molecular sensor, such as ProCas9, could actuate many types of outputs. One unique effect would be to induce cell death upon sensing viral infection, as a form of altruistic defense. Since activated ProCas9 is capable of inducing DNA double-strand breaks, we sought to identify sgRNAs that could induce rapid cell death. As Flaviviruses replicate rapidly upon target cell infection, such sgRNAs would have to kill their host cells in less time. Targeting essential genes such as the single-stranded DNA binding protein RPA1, which is involved in DNA replication, could be one option. Alternatively, targeting highly repetitive sequences within a cell's genome to induce massive DNA damage and cellular toxicity could be another avenue. Indeed, sgRNAs targeting even only moderately amplified loci have been shown to lead to cell depletion under certain conditions[148], independent of whether the sgRNA targets a gene or intergenic region. While these effects have been observed over long assay periods, targeting highly repetitive sequences might provide sufficient DNA damage to trigger rapid cell death.

To compare the two strategies, both HEK293T and HAP1 cells were stably transduced to express WT Cas9 and an sgRNA coupled to an mCherry fluorescence marker (**Figure 4-11A**). The effect of guide RNA expression on cell viability was assessed using a competitive proliferation assay in which cells expressing a specific sgRNA were mixed with parental cells expressing only Cas9 WT, and the mCherry-positive population was followed over time. Negative control guides targeting an olfactory receptor gene (sgOR2B6-1, sgOR2B6-2) showed no depletion. As expected, guide RNAs targeting the essential RPA1 gene depleted over the eight-day assay period. To potentially accelerate depletion, we also designed and tested several sgRNAs targeting repetitive sequences in the human genome (~125,000–300,000 target loci each; **Materials and Methods**), which could cause CRISPR-Cas induced death by editing or "CIDE." Indeed, CIDE guide RNAs (sgCIDE-1, sgCIDE-2, sgCIDE-4, sgCIDE-5) led to rapid elimination of the mCherry-positive population (**Figure 4-11A**) and show promise as a simple genetic output module for an altruistic defense system based on CRISPR-Cas-mediated cell death.

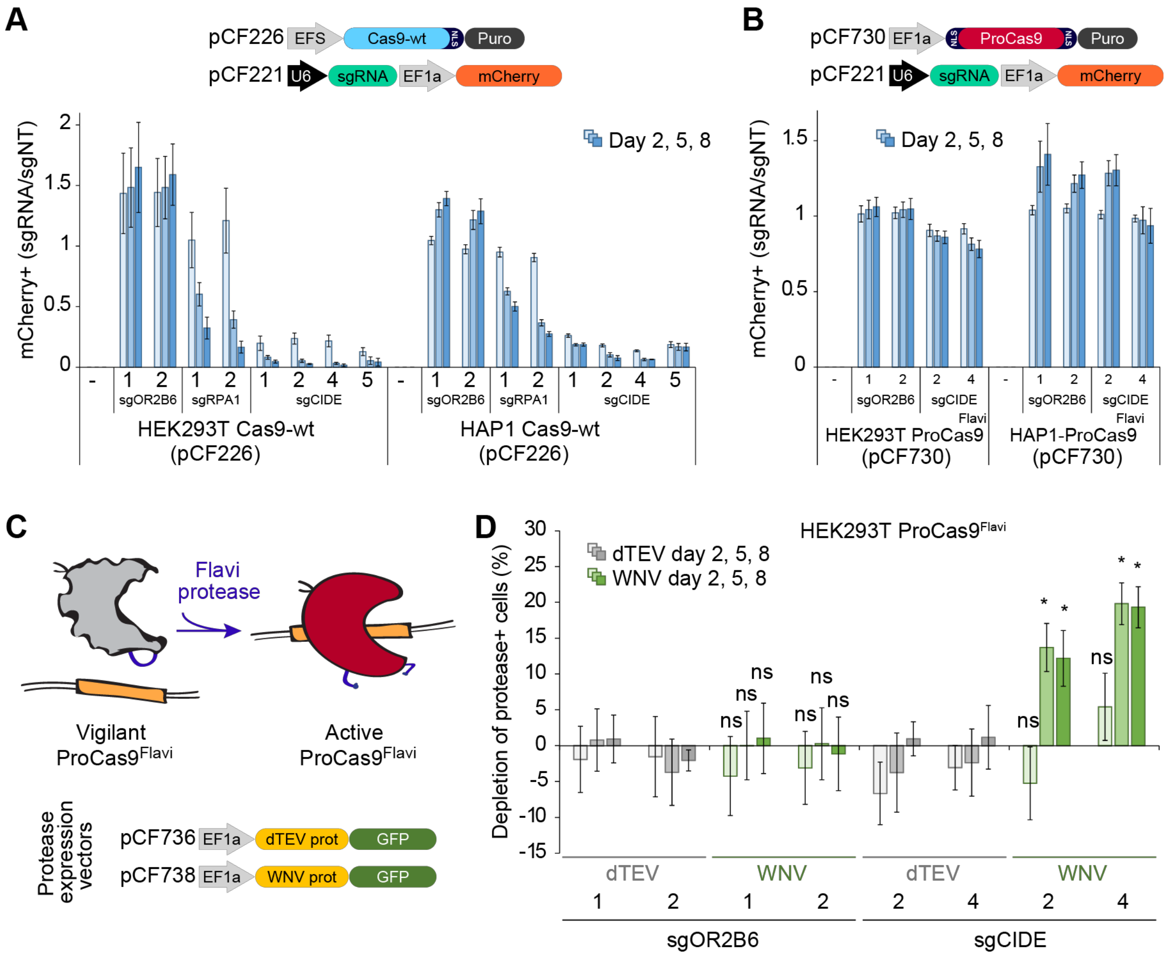


Figure 4-11. ProCas9 enables genomically encoded programmable response systems.

(**A**) CRISPR-Cas-programmed cell depletion. HEK293T and HAP1 cells expressing Cas9 WT were transduced with mCherry-tagged sgRNAs. After mixing with parental cells, the fraction of mCherry-positive cells was quantified over time. Different sgRNAs targeting a neutral gene (sgOR2B6), an essential gene (sgRPA1), >100,000 genomic loci (sgCIDE), and a non-targeting control (sgNT) were compared. Error bars represent the standard deviation of triplicates. (**B**) Competitive proliferation assay analogous to (A), conducted in HEK293T and HAP1 cells expressing the ProCas9$^{\text{Flavi}}$ system. Note that sgCIDE-positive cells show little or no depletion because the ProCas9$^{\text{Flavi}}$ is in its inactive, vigilant state. (**C**) ProCas9$^{\text{Flavi}}$ activation by *Flavivirus* proteases expressed from genomically integrated lentiviral vectors. (**D**) Competitive proliferation assay in HEK293T ProCas9$^{\text{Flavi}}$ cells expressing the indicated mCherry-tagged sgRNAs or a non-targeting control (sgNT) used for normalization. Cells were partially transduced with lentiviral vectors expressing a GFP-tagged dTEV or WNV protease and cell depletion quantified by flow cytometry. Note that the WNV protease leads to protective cell death (altruistic defense) in sgCIDE-expressing cells through activation of the ProCas9$^{\text{Flavi}}$ system. Error bars represent the SD of triplicates. Significance was assessed by comparing each sample to its respective dTEV control (unpaired, two-tailed t test, n = 3, $*p < 0.05$; ns, not significant).

4.3.8. Genomic ProCas9 can sense *Flavivirus* proteases and mount an altruistic defense

CIDE as an output constrains the performance of ProCas9. The system must remain off to minimize genomic damage, yet be vigilant to respond to a stimulus. To develop this protease-induced altruistic defense platform, we assessed whether stable expression of the best CIDE guide RNAs (sgCIDE-2, sgCIDE-4) in conjunction with a genomically integrated ProCas9Flavi cassette was viable in the absence of a stimulus (**Figure 4-11B**). Competitive proliferation assays analogous to the ones run with WT Cas9 showed that in the presence of ProCas9Flavi only minimal amounts of cell depletion were observed.

Finally, we tested induction of this stably integrated altruistic defense system by Flavivirus proteases (**Figure 4-12A**). Using the same cell lines (expressing ProCas9Flavi) as above, we observed that stable transduction with vectors expressing either a control (dTEV) or Flavivirus (WNV) protease led to specific cell depletion only when both the WNV protease was present and the system was programmed with one of the two CIDE sgRNAs (**Figure 4-11CD**, **Figure 4-12B**). Hence, these results confirmed that our Flavivirus ProCas9 system can be stably integrated into the genome of a host cell to detect predefined protease activity and mount a programmed defense, only in the presence of a specific stimulus of interest.

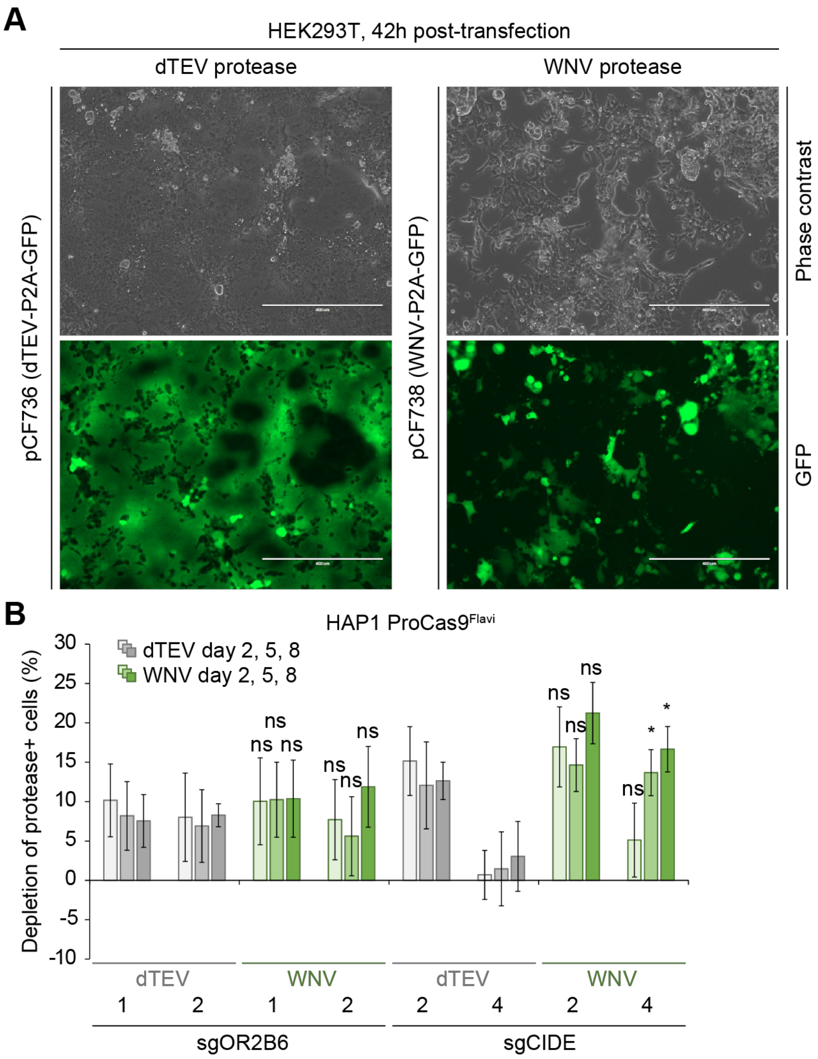


Figure 4-12. ProCas9-based altruistic defense systems.

(**A**) Transfection of protease expression vectors in virus packaging cell lines. GFP fluorescence imaging in HEK293T cells 42 hr after transfection of the indicated lentiviral plasmids expressing viral proteases. Lentiviral helper plasmids were co-transfected in each case. Scale bar: 400 μ m. (**B**) Competitive proliferation assay in HAP1 ProCas9^{Flavi} (pCF730) cell lines expressing the indicated mCherry-tagged controls (sgOR2B6-1, sgOR2B6-2) or guide RNAs targeting highly repetitive sequences (sgCIDE-2, sgCIDE-4), or a non-targeting control (sgNT) used for normalization. The cell lines were partially transduced with lentiviral vectors expressing a GFP-tagged dTEV (pCF736) or WNV (pCF738) protease, and cell depletion quantified by flow cytometry. Shown is the normalized (sgRNA/sgNT) depletion of protease-expressing (GFP+) cells among the sgRNA-positive (mCherry+) population. Error bars indicate the standard deviation of triplicates. Significance was assessed by comparing each sample to its respective dTEV control (unpaired, two-tailed t test, $n = 3$, $* = p < 0.05$, ns = not significant).

4.4. Discussion

Here we demonstrate that the large, multi-domain, and highly allosteric enzyme Cas9 is amenable to circular permutation via protein engineering, without apparent loss of its functions. By systematically creating and testing the sequence of Cas9 for circular permutation, we identified a comprehensive suite of novel variants that are efficient at genome binding and cleavage, with the added benefit of redistributed new N and C termini across Cas9's topology (**Figure 4-13**). Additionally, we show that Cas9 circular permutants can be rewired into molecular recording devices, termed ProCas9s, that can sense proteases—including those from Flaviviruses and Potyviruses—to stably record their activity in the genome or mount a pre-programmed defense. Importantly, the modularity of this system enables simple redesign of the ProCas9 sensing activity by swapping of the CP linker and, as such, could respond to any exogenous or endogenous sequence-specific protease. Thus, the system may be used to sense and report cell-intrinsic pathway activity, e.g., for molecular screening and drug discovery, or serve as a means for cell-type-specific Cas activation after general delivery of an editing complex to a target tissue or organ.

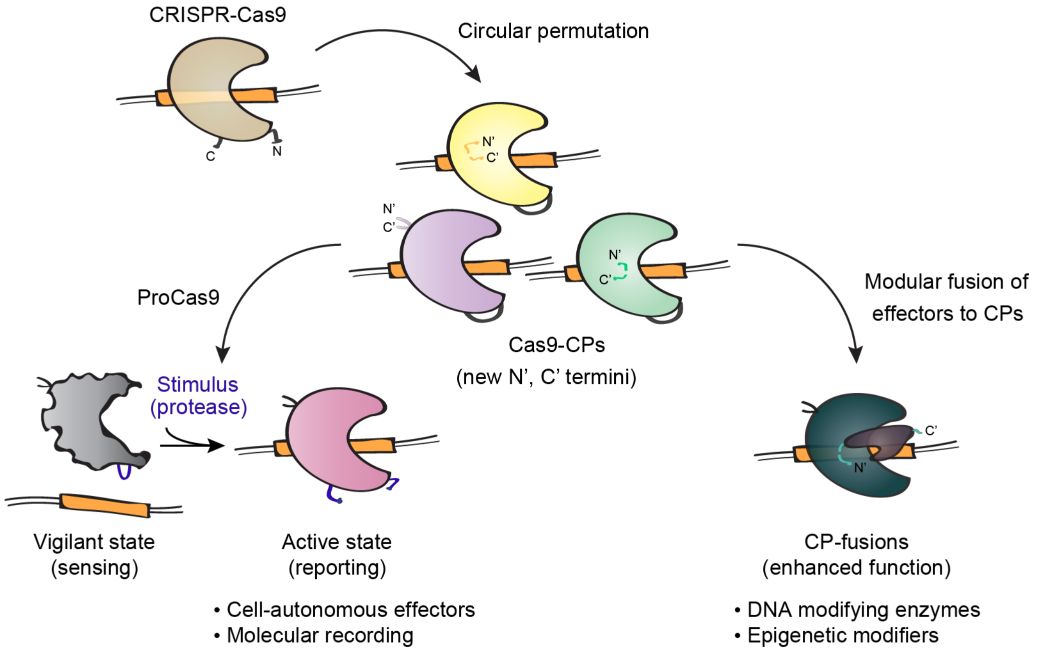


Figure 4-13. Application of Cas9 circular permutants.

Diagram showing various uses of Cas9 circular permutants (Cas9-CPs) as single-molecule sensor effectors for protease tracing and molecular recording, or as optimized scaffolds for modular CP-fusion proteins with novel and enhanced functionalities.

The ProCas9s ability to serve as a detector of pathogen activity is intriguing as it could enable their use as a fully modular, genomically encoded immune system with both a designable input and programmable output. For example, many plants are known to contain protease-gated transcription factors that activate a protective hypersensitivity response when cleaved by a pathogen protease[149,150] and one of these proteins has even been adapted for the recognition of a non-native protease[151]. Nevertheless, this system's output is constrained by the DNA-binding specificity of the transcription factor. In contrast, ProCas9s are a simple effector with a designable input and programmable output that should work in every organism CRISPR proteins have been shown to operate in. As one example, here we show how ProCas9s can be tuned to serve as an altruistic defense system to protect a population of human cells by self-elimination of the few cells expressing the Flavivirus protease, which mimics the infection. Thus, we have demonstrated an initial proof-of-concept for a fully synthetic and customizable resistance gene. Hence, it should be straightforward to transition this self-targeting system into a platform that can induce expression or suppression of genes to mount a systemic immune response, or to activate a synthetic cellular program to track pathogen invasion. Such a strategy for pathogen detection is broadly applicable, as many pathogens express proteases during host infection[152-154].

Others have recently adapted constitutively expressed CRISPR systems to target pathogenic viruses directly[155-157]. However, these systems utilize a fully active nuclease gated only by sequence recognition. The sustained expression of Cas9 both increases the risks of off-target effects and promotes evolution of the targeted viruses. Indeed, recent reports[156,158] highlight this phenomenon, which may represent an unintended consequence of utilizing CRISPR systems in a pathogen-directed fashion. In contrast, the ProCas9 system allows programming a response to a viral infection akin to innate immunity, where a self-directed response can be activated to minimize the opportunity for evasive viral hypermutation and resistance.

Additionally, the Cas9-CPs serve as a diverse set of protein scaffolds for advanced CRISPR-Cas fusion proteins. The natural N and C termini are fixed for all proteins. Our work paves the way for making a new class of CP-based CRISPR tools with optimized N and C termini for fusions. In Cas9, for example, the native termini are ~50 Å apart, requiring long linkers for fusions that seek access to the DNA[115,117,118,138]. The dearth of options when attempting to build new Cas9 fusions may explain the relative lack of activity or undesired side effects of many compound constructs. Indeed, dCas9 activators need numerous domains (up to 24) [127,128] or combinations of guide RNAs for high activity[116]. dCas9-FokI fusions are not as efficient at indel induction as Cas9 itself[115,118], and the base editing cytidine deaminase fusions, which result in strong C to T editing within a 12-bp target window, may also cause deamination up to 15 bp outside of the Cas9 target sequence[117]. Circular permutation of Cas9 yields a new class of scaffolds with N and C termini within 5 Å of the bound target or non-target strand, which may remedy current steric limitations.

Taken together, a more holistic approach to Cas9 tool building—one that includes engineering of both the fusion scaffold and fusion domain—enables a more proficient generation of modular and customizable CRISPR-Cas9 effectors. Our work lays the foundation for this process by providing both a blueprint for the circular permutation of Cas9, as well as by providing a resource of functionally active Cas9-CPs for advanced fusion proteins. Additionally, we present the concept of ProCas9 variants that can be enzymatically activated by sequence-specific proteases to serve as molecular recorders or tissue-specific effectors.

4.5. Materials and Methods

Table 4-4. Key resources.

Category	Reagent or resource	Source	Identifier
Antibodies	Anti-Flag-M2, clone M2	Sigma-Aldrich	Cat#1804
Antibodies	DYKDDDDK Tag (Anti-Flag) antibody	Cell Signaling Technology	Cat#2044
Antibodies	C-Cas9 Anti-SpyCas9, clone 10C11-A12	Sigma-Aldrich	Cat#SAB4200751
Antibodies	N-Cas9 Anti-SpyCas9, clone 7A9-3A3	Novus Biologicals	Cat#NBP2-36440
Antibodies	HRP-conjugated Anti-Beta-Actin, clone C4	Santa Cruz Biotechnology	Cat#sc-47778
Bacterial and Virus Strains	GFP/RFP expressing *E. coli* MG1655	Qi et al 2013	N/A
Deposited Data	Cas9-CP sequencing data	This paper	Accession code PRJNA505363
Experimental Models: Cell Lines	HEK293T	Thermo Fisher Scientific	Cat#R70007
Experimental Models: Cell Lines	HepG2	ATCC	Cat#HB-8065
Experimental Models: Cell Lines	A549	ATCC	Cat#CCL-185
Experimental Models: Cell Lines	HAP1	Jan Carette, Stanford; Carette et al., 2011	N/A
Oligonucleotides	sgGFP1 (CCTCGaaCTTCACCTCGGCG)	Oakes et al., 2016	N/A
Oligonucleotides	sgGFP2 (CaaCTACaaGACCCGCGCCG)	Oakes et al., 2016	N/A
Oligonucleotides	sgGFP9 (CCGGCaaGCTGCCCGTGCCC)	This paper	N/A
Oligonucleotides	sgCIDE-1 (TGTaaTCCCAGCACTTTGGG)	This paper	N/A
Oligonucleotides	sgCIDE-2 (TCCCaaAGTGCTGGGATTAC)	This paper	N/A
Oligonucleotides	sgCIDE-4 (CGCCTGTaaTCCCAGCACTT)	This paper	N/A
Oligonucleotides	sgCIDE-5 (CCTCGGCCTCCCaaAGTGCT)	This paper	N/A
Software and Algorithms	Cas9-CP analysis scripts	This paper	https://github.com/SavageLab/cpCas9

4.5.1. Bacterial strains and media

For in-vivo E. coli screening, fluorescence measurements, and cell proliferation assays, we used MG1655 with a chromosomally integrated and constitutively expressed GFP and RFP[25,137]. EZ-rich defined growth medium (EZ-RDM, Teknoka) was used for all liquid culture assays and plates were made using 2xYT. Plasmids used were based on a 2 plasmid system as reported previously[25,126,137] containing Cas9 and variants on a selectable CmR marker and plasmids with sgRNAs and proteases with AmpR markers, representative sequences of which can be found in **Table 4-1**. The antibiotics were used to verify transformation and to maintain plasmid stocks. No blinding or randomization was done for any of the experiments reported.

4.5.2. Mammalian cell culture

All mammalian cell cultures were maintained in a 37°C incubator, at 5% CO2. HEK293T (293FT; Thermo Fisher Scientific, #R70007) human kidney cells and derivatives thereof were grown in Dulbecco's Modified Eagle Medium (DMEM; Corning Cellgro, #10-013-CV) supplemented with 10% fetal bovine serum (FBS; Seradigm, #1500-500), and 100 Units/ml penicillin and 100 μ g/ml streptomycin (100-Pen-Strep; GIBCO #15140-122). HepG2 human liver cells (ATCC, #HB-8065) and derivatives thereof were cultured in Eagle ' s Minimum Essential Medium (EMEM; ATCC, #30-2003) supplemented with 10% FBS and 100-Pen-Strep. A549 human lung cells (ATCC, #CCL-185) and derivatives thereof were grown in Ham's F-12K Nutrient Mixture, Kaighn's Modification (F-12K; Corning Cellgro, #10-025-CV) supplemented with 10% FBS and 100-Pen-Strep. HAP1 cells (kind gift from Jan Carette, Stanford) and derivatives thereof were grown in Iscove's Modified Dulbecco's Medium (IMDM; GIBCO #12440-053 or HyClone #SH30228.01) supplemented with 10% FBS and 100-Pen-Strep. HAP1 cells had been derived from the near-haploid chronic myeloid leukemia cell line KBM7[159]. Karyotyping analysis demonstrated that most cells (27 of 39) were fully haploid, while a smaller population (9 of 39) was haploid for all chromosomes except chromosome 8, like the parental KBM7 cells. Less than 10% (3 of 39) were diploid for all chromosomes except for chromosome 8, which was tetraploid.

A549 cells were authenticated using short tandem repeat DNA profiling (STR profiling; UC Berkeley Cell Culture/DNA Sequencing facility). STR profiling was carried out by PCR amplification of nine STR loci plus amelogenin (GenePrint 10 System; Promega #B9510), fragment analysis (3730XL DNA Analyzer; Applied Biosystems), comprehensive data analysis (GeneMapper software; Applied Biosystems), and final verification using supplier databases including American Type Culture Collection (ATCC) and Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ).

HEK293T, HEK-RT1, HEK-RT6, HepG2, A549, and HAP1 cells were tested for absence of mycoplasma contamination (UC Berkeley Cell Culture facility) by fluorescence microscopy of methanol fixed and Hoechst 33258 (Polysciences #09460) stained samples.

4.5.3. Transposon library construction

To begin, a dCas9 flanked by BsaI restriction enzyme sites was inserted into a pUC19 based plasmid. A modified transposon with R1 and R2 sites based on[135] (**Table 4-1**), containing a chloramphenicol antibiotic resistance marker, p15A origin of replication, TetR and TetR/A promoter, was built using custom oligos and standard molecular biology techniques. It was then cleaved from a plasmid using HindIII and gel purified. This linear transposon product was used in overnight in vitro reactions (0.5 molar ratio transposon to 100 ng dCas9-Puc19 plasmid) using 1 μ L of MuA Transposase (F-750, Thermo Fisher) in 10 replicates. The transposed DNA was purified and recovered, Plasmids were electroporated into custom made electrocompetent MG1655 *E. coli*[137] using a BTX Harvard apparatus ECM 630 High Throughput Electroporation System and titered on carbenicillin (Carb) and chloramphenicol (CM) to ensure > 100x coverage of the library size (13,614). These cells were then outgrown for 12 hours and selected for via Carb and CM markers to ensure growth of transposed members. After isolating transposed plasmids via miniprep (QIAGEN), the original Puc19 backbone was removed via BsaI cleavage and dCas9 proteins transposed with a new plasmid backbone were selected via a 0.7% TAE agarose gel. The linear fragments were then ligated overnight with annealed and phosphorylated oligos coding for GGS linkers of 5, 10, 15 and 20 aa using a BsaI Golden Gate reaction. Completed libraries were purified, electroporated into the Mg1655 RFP and GFP screening strain containing a RFP-repressing sgRNA and titered on Carb and CM to ensure > 5x coverage of the library size (8,216).

4.5.4. Screening for Cas9 circular permutants (Cas9-CPs)

Screens were performed in a similar manner to previous reports[126,137], briefly biological duplicates of Cas9-CP libraries with a RFP guide RNA were transformed (at > 5x library size) into MG1655 with genetically integrated and constitutively expressed GFP and RFP. Cells were grown overnight in EZ-RDM + Carb, CM and 200 nM Anhydrotetracycline (aTc) inducer. *E. coli* were then sorted based on gates for RFP but not GFP repression, collected, and resorted immediately to further enrich for functional Cas9-CPs (**Figure 4-2C**). Double sorted libraries were then grown out and DNA was collected for sequencing. This DNA was also re-transformed onto plates and individual clones were picked for further analysis.

4.5.5. Deep sequencing library preparation

This method was modified from previous Tnseq protocols[160]. Briefly, the transposed plasmids were sheared to ~300bp using a S220 Focused-ultrasonicator (Covaris) and purified in between each of the following steps using Agencourt AMPure XP beads (Beckman Coulter). Following shearing, fragments were end-repaired and A-tailed according to NEB manufacturers protocols and then universal adapters were ligated on in a 50 ul quick ligase reaction at RT. Finally fragments from each library were amplified in a 20-cycle reaction with Indexed Illumina primers that annealed upstream of the new CP start codon and in the universal adaptor (**Table 4-1**). PCRs were cleaned again and analyzed for primer dimers via an Agilent Bioanalyzer DNA 1000 chip. Sequencing was performed at the QB3 Vincent J. Coates Genomics Sequencing Laboratory on a HiSeq2500 in a 100 bp run.

4.5.6. Deep sequencing analysis

Demultiplexed reads from the HiSeq2500 were assessed using FastQC to check basic quality metrics. Reads for each sample were then trimmed using a custom python script. The trimmed sequences were mapped to the dCas9 nucleic acid sequence using BWA via a custom python wrapper script to determine the amino acid position in dCas9 corresponding to the starting amino acid position in the dCas9-CP permutant. The resulting alignment files were then processed using a custom python script to calculate the abundance of each dCas9-CP permutant in a given library sample. Fold-changes for each dCas9-CP permutant between pre- and post-library sorts along with significance values for each enrichment were calculated (**Table 4-2**) using the DESeq package in R[161]. Due to ambiguity in transposon sequence, insertion site calls in **Table 4-2** are one greater (sites: n+1) than the variants named in **Table 4-3**. As per the DESeq guidelines, count data from technical sequencing replicates were summed to create one unique replicate before running through the DESeq pipeline. All relevant sequencing data and Cas9-CP analysis scripts are available at https://github.com/SavageLab/cpCas9.

4.5.7. *E. coli* CRISPRi GFP repression assay

Assays were performed similar to previous descriptions[126]. To measure the ability of a circular permutant to bind to and repress DNA expression, cells were co-transformed with a Cas9 permutant plasmid with aTc inducible promoter and a single guide RNA plasmid for RFP or GFP that, in the case of the ProCas9 assays, also contained the active or inactive proteases on an IPTG-inducible promoter. Endpoint Assay: Cells were picked in biological triplicate into 96 well plates containing 500 uL EZ MOPS plus Carb and CM. Plates were grown in 37°C shakers for twelve hours. Next, cells were diluted 1:1000 in 500 uL EZ MOPS plus Carb, CM, IPTG and aTc. 200 nM aTc was used to induce Cas9-CPs or ProCas9s and 50 uM IPTG levels was used to induce the proteases in a 2mL deep well blocks and shaken at 750 rpm at 37°C. After an eight-twelve hr induction and growth period, 20 uL of cells were added to 80 uL of water and put into a 96-well microplate reader (Tecan M1000) at 37°C and read immediately. Each well was measured for optical density at 600 nm and GFP or RFP fluorescence. GFP expression was normalized by dividing it with OD600. In the case of the time course assays (**Figure 4-4BC**), 150 uL of the 1:1000 dilution was used and placed into a black walled clear bottom plate (3631-Corning) and directly into the Tecan M1000 for a 130x 600 s kinetic cycle of reading. For *E. coli* single cell analysis (**Figure 4-4C**), cells from the endpoint time course were run on a Sony SH800 to capture 100,000 events per sample.

4.5.8. *E. coli* genomic cleavage assay

Assays were performed as previously described[126] *E. coli* containing sgRNA plasmids targeting a genomically integrated GFP were made electrocompetent and transformed with 10 ng of the the various Cas9-CP plasmids or controls using electroporation. After recovery in 1 mL SOC media for 1 hr, cells were plated in technical triplicate of tenfold serial dilutions onto 2xYT agar plates with antibiotics selection for both plasmids and aTc induction at 200 nM. Plates were grown at 37°C overnight and CFU/mL was determined. A reduction in CFUs indicated genomic cleavage and cell death.

4.5.9. *E. coli* western blotting

After CRISPRi repression assays for TEV linker Pro-Cas9s, 40 uL of cell culture was pelleted and resuspended in SDS loading buffer for further analysis. SDS samples were loaded into 4%–20% acrylamide gels (BioRad) for electrophoresis. After transfer to membranes (Trans-Blot Turbo- BioRad), blots were washed 3x with 1xTBS + 0.01% Tween 20, blocked with 5% milk for 1.5 hr and then a 1:1000 of HRP-conjugated DYKDDDDK Tag (Anti-Flag) antibody (Cell Signaling Technology, #2044) was incubated for twenty-four hours at 4°C. Antibodies were washed away with 3x TBST and detected using Pierce ECL Western Blotting Substrate (Thermo Fisher).

4.5.10. NIa protease cleavage sites

NIa protease cleavage sites – i.e., the CP linkers – were identified from previous reports[151] (TuMV, 7 aa), by using the sequence between the P3 and 6KI genes annotated in NCBI (PPV, PVY, CBSV), or from previously identified Potyvirus protease consensus sequences[141].

4.5.11. Lentiviral vectors

A lentiviral vector referred to as pCF204, expressing a U6 driven sgRNA and an EFS driven Cas9-P2A-Puro cassette, was based on the lenti-CRISPR-V2 plasmid[162], by replacing the sgRNA with an enhanced Streptococcus pyogenes Cas9 sgRNA scaffold[113]. The pCF704 and pCF711 lentiviral vectors, expressing a U6-sgRNA and an EFS driven ProCas9 variant, were derived from pCF204 by swapping wild-type Cas9 for the respective ProCas9 variant. The pCF712 and pCF713 vectors were derived from pCF704 and pCF711, respectively, be replacing the EF1a-short promoter (EFS) with the full-length EF1a promoter. The lentiviral vector pCF732 was derived from pCF712 by removal of the ProCas9's nuclear localization sequences (NLSs). Vectors not containing a guide RNA, including pCF226 (Cas9-wt) and pCF730 (ProCas9Flavi), were derived from pCF204 and pCF712, respectively, through KpnI/NheI-based removal of the U6-sgRNA cassette and blunt ligation. The guide RNA-only vector pCF221, encoding a U6-sgRNA cassette and an EF1a driven mCherry marker, is loosely based on the pCF204 backbone and guide RNA cassette. Lentiviral vectors expressing viral proteases, including pCF708 expressing an EF1a driven mTagBFP2-tagged dTEV protease, pCF709 expressing an EF1a driven mTagBFP2-tagged ZIKV NS2B-NS3 protease, and pCF710 expressing an EF1a driven mTagBFP2-tagged WNV protease, are all based on the pCF226 backbone. The GFP-tagged protease vectors pCF736 and pCF738 are derived from pCF708 and pCF710, respectively, by swapping mTagBFP2 with GFP. All vectors were generated using custom oligonucleotides (IDT), gBlocks (IDT), standard cloning methods, and Gibson assembly techniques and reagents (NEB). Vector sequences are provided (**Table 4-1**).

4.5.12. Design of sgRNAs

Standard sgRNA sequences were either designed manually, using CRISPR Design (crispr.mit.edu), or using GuideScan[163]. When editing endogenous genes, sgRNAs were often designed to target evolutionarily conserved regions in the 5′ proximal third of the gene of interest. The following sequences were used: sgGFP1 (CCTCGaaCTTCACCTCGGCG), sgGFP2 (CaaCTACaaGACCCGCGCCG), sgGFP9 (CCGGCaaGCTGCCCGTGCCC), sgOR2B6-1 (CATTATTCTAGTGTCACGCC), sgOR2B6-2 (GGGTATGaaGTTTGGTGTCC), sgPCSK9-4 (CCGGTGGTCACTCTGTATGC), sgPuro5 (TGTCGAGCCCGACGCGCGTG), sgPuro6 (GCTCGGTGACCCGCTCGATG), sgRPA1-1 (ACaaaaGTCAGATCCGTACC), sgRPA1-2 (TACCTGGAGCaaCTCCCGAG). All sgRNAs were designed with a G preceding the 20 nucleotide guide for better expression from U6 promoters.

To enable rapid CRISPR-Cas controlled cell depletion, through a strategy that we termed Cas-induced death by editing or ‘CIDE’, we designed sgRNAs (sgCIDEs) directed again highly repetitive sequences in the human genome. In brief, using GuideScan[163] we identified the most frequently occurring Streptococcus pyogenes Cas9 sgRNA target sites (5′ -NGG-3′ PAM) in the hg38 assembly (Genome Reference Consortium Human Build 38) of the human genome. From this list we eliminated sequences containing extended homomeric stretches (> 4 A/T/C/G), and empirically validated two sequences with slightly over 125,000 target loci (sgCIDE-4, CGCCTGTaaTCCCAGCACTT; sgCIDE-5, CCTCGGCCTCCCaaAGTGCT) and two sequences with approximately 300,000 target loci (sgCIDE-1, TGTaaTCCCAGCACTTTGGG; sgCIDE-2, TCCCaaAGTGCTGGGATTAC). All four sgCIDEs led to rapid cell depletion when expressed in presence of active Cas9.

4.5.13. Lentiviral transduction

Lentiviral particles were produced in HEK293T cells using polyethylenimine (PEI; Polysciences #23966) based transfection of plasmids. HEK293T cells were split to reach a confluency of 70%–90% at time of transfection. Lentiviral vectors were co-transfected with the lentiviral packaging plasmid psPAX2 (Addgene #12260) and the VSV-G envelope plasmid pMD2.G (Addgene #12259). Transfection reactions were assembled in reduced serum media (Opti-MEM; GIBCO #31985-070). For lentiviral particle production on 10 cm plates, 8 μ g lentiviral vector, 4 μ g psPAX2 and 2 μ g pMD2.G were mixed in 2 mL Opti-MEM, followed by addition of 42 μ g PEI. After 20-30 min incubation at room temperature, the transfection reactions were dispersed over the HEK293T cells. Media was changed 12 hr post-transfection, and virus harvested at 36-48 hr post-transfection. Viral supernatants were filtered using 0.45 μ m cellulose acetate or polyethersulfone (PES) membrane filters, diluted in cell culture media if appropriate, and added to target cells. Polybrene (5 μ g/ml; Sigma-Aldrich) was supplemented to enhance transduction efficiency, if necessary.

Transduced target cell populations (HEK293T, A549, HAP1, HepG2 and derivatives thereof) were usually selected 24-48 hr post-transduction using puromycin (InvivoGen #ant-pr-1; HEK293T, A549 and HepG2: 1.0 μ g/ml, HAP1: 0.5 μ g/ml) or hygromycin B (Thermo Fisher Scientific #10687010; 200-400 μ g/ml).

4.5.14. Rapid mammalian genome editing reporter assay

To establish a rapid and quantitative way to reliably assess genome editing efficiency from various CRISPR-Cas constructs in mammalian cells, we decided to build a fluorescence-based reporter assay. Assays leveraging editing-based disruption of a constitutively expressed fluorescence marker have been built before. However, such assays show a long detection lag time as the genetic disruption of a locus coding for the fluorescent marker will not immediately lead to a reduction in the fluorescence signal, due to the remaining presence of intact transcripts and protein half-life. To quantify this effect, we stably transduced HEK293T cells with a retroviral vector (LMP-Pten.1524) constitutively expressing GFP[164], and established monoclonal derivatives. The best performing cell line was termed HEK-LMP-10. When editing this reporter line with a vector (pX459, Addgene #48139) expressing wild-type Streptococcus pyogenes Cas9 and guide RNAs targeting the reporter (sgGFP1, sgGFP2), or a non-targeting control (sgNT), the editing detection lag – defined as the time between introduction of an editing reagent and complete loss of fluorescence signal in edited cells – was up to eight days (**Figure 4-3A**). Hence, this type of assay is inconvenient for rapid quantification of editing efficiency. Conversely, assays relying on frameshift mutations to activate a fluorescence reporter often require specific guide RNA sequences and only get activated with the faction of edits that lead to the required frameshift, thus introducing a quantification bias.

To overcome this limitation, we decided to build an inducible genome editing reporter cell line comprising a fluorescence marker that is not expressed in the default state but can be induced following a defined time of potential genome editing (**Figure 4-3B**). In this scenario, unedited cells will rapidly turn positive, while non-edited cells remain fluorophore negative. Specifically, inducible monoclonal HEK293T-based genome editing reporter cells, referred to as “HEK-RT1,” were established in a two-step procedure. In the first step, puromycin resistant monoclonal HEK-RT3-4 reporter cells were generated[165]. In brief, HEK293T human embryonic kidney cells were transduced at low-copy with the amphotropic pseudotyped RT3GEPIR-Ren.713 retroviral vector[164], comprising an all-in-one Tet-On system enabling doxycycline-controlled GFP expression. After puromycin (2.0 μ g/ml) selection of transduced HEK239Ts, 36 clones were isolated and individually assessed for (i) growth characteristics, (ii) homogeneous morphology, (iii) sharp fluorescence peaks of doxycycline (1 μ g/ml) inducible GFP expression, (iv) relatively low fluorescence intensity to favor clones with single-copy reporter integration, and (v) high transfectability. HEK-RT3-4 cells are derived from the clone that performed best in these tests.

Since HEK-RT3-4 are puromycin resistant, in the second step, monoclonal HEK-RT1 and analogous sister reporter cell lines were derived by transient transfection of HEK-RT3-4 cells with a pair of vectors encoding Cas9 and guide RNAs targeting puromycin (sgPuro5, sgPuro6), followed by identification of monoclonal derivatives that are puromycin sensitive. In total, eight clones were isolated and individually assessed for (i) growth characteristics, (ii) homogeneous morphology, (iii) doxycycline (1 μ g/ml) inducible and reversible GFP fluorescence, and (iv) puromycin and hygromycin B sensitivity. The monoclonal HEK-RT1 and HEK-RT6 cell lines performed best in these tests and were further evaluated in a doxycycline titration experiment (**Figure 4-3C**), showing that both reporter lines enable doxycycline concentration-dependent induction of the fluorescence marker in as little as 24-48 hours. The HEK-RT1 cell line was chosen as rapid mammalian genome editing reporter system for all further assays.

4.5.15. Genome editing analysis using mammalian HEK-RT1 reporter assay

When employing the HEK-RT1 genome editing reporter assay to quantify WT Cas9 (Cas9-wt) and ProCas9 variant activity following stable genomic integration, HEK-RT1 reporter cells were transduced with the indicated Cas-wt/ProCas9 and sgRNA lentiviral vectors (**Table 4-1**) and selected on puromycin. A guide RNA targeting the GFP fluorescence reporter (sgGFP9) was compared to a non-targeting control (sgNT). We used the non-targeting control in all assays for normalization, in case not all non-edited cells turned GFP positive upon doxycycline treatment, though usual reporter induction rates were above 95%. GFP expression in HEK-RT1 reporter cells was induced for 24-48 hr using doxycycline (1 μg/ml; Sigma-Aldrich), at the indicated days post-editing. Percentages of GFP-positive cells were quantified by flow cytometry (Attune NxT, Thermo Fisher Scientific), routinely acquiring 10,000-30,000 events per sample. When quantifying ProCas9 activation by mTagBFP2-tagged proteases, GFP fluorescence was quantified in mTagBFP2-positive cells. In all cases, editing efficiency was reported as the difference in percentage of GFP-positive cells between samples expressing a non-targeting guide (sgNT) and samples expressing the sgGFP9 guide targeting the GFP reporter. For ProCas9 GFP disruption assays following transfection of the tested components (**Figure 4-6F**), transfection-based plasmids were designed and cloned using standard molecular biology techniques to express either ProCas9-T2A-mCherry and a single guide RNA, or the protease of interest-P2A-mTagBFP2 (**Table 4-1**). Transient assays were performed as follows: in triplicate the reporter cell line HEK-RT1 was seeded at 20-30 thousand cells per well into 96-well plates and transfected using 0.5 uL of Lipofectamine 2000 (Thermo Fisher Scientific), 12.5 ng of the WT Cas9 or ProCas9 plasmid and 14 ng of the Protease plasmid (2x molar ratio), following the manufacturer's protocol. 24 hours later the media was changed and doxycycline was added to induce GFP expression. 48 hours following induction the cells were gated for mCherry (WT Cas9, ProCas9) expression and analyzed using flow cytometry for GFP depletion. At least 10,000 events were collected for each sample.

4.5.16. Mammalian flow cytometry and fluorescence microscopy

Flow cytometry (Attune Nxt Flow Cytometer, Thermo Fisher Scientific) was used to quantify the expression levels of fluorophores (mTagBFP2, GFP/EGFP, mCherry) as well as the percentage of transfected or transduced cells. For the HEK-RT1 genome editing reporter cell line, flow cytometry was used to quantify the percentage of GFP-negative (edited) cells, 24-48 hr after doxycycline (1 μg/mL) treatment to induce GFP expression. Phase contrast and fluorescence microscopy was carried out following standard procedures (EVOS FL Cell Imaging System, Thermo Fisher Scientific), routinely at least 48 hr post-transfection or post-transduction of target cells with fluorophore expressing constructs.

4.5.17. Mammalian immunoblotting

HEK293T (293FT; Thermo Fisher Scientific) were co-transfected with the indicated plasmids expressing Cas9-wt or ProCas9-Flavi and plasmids expressing dTEV or WNV protease. HEK293T cells were split to reach a confluency of 70%–90% at time of transfection. For transfections in 6-well plates, 1 μg Cas9-sgRNA vector and 0.75 μg protease vector (if applicable) were mixed in 0.4 mL Opti-MEM, followed by addition of 5.25 μg polyethylenimine (PEI; Polysciences #23966). After 20-30 min incubation at room temperature, the transfection reactions were dispersed over the HEK293T cells. Media was changed 12 hr post-transfection. At 36 hr post-transfection, HEK293T were washed in ice-cold PBS and scraped from the plates. Cell pellets were lysed in Laemmli buffer (62.5 mM Tris-HCl pH 6.8, 10% glycerol, 2% SDS, 5% 2-mercaptoethanol). Equal amounts of protein were separated on 4%-20% Mini-PROTEAN TGX gels (Bio-Rad, #456-1095) and transferred to 0.2 μm PVDF membranes (Bio-Rad, #162-0177). Blots were blocked in 5% milk in TBST 0.1% (TBS + 0.01% Tween 20) for 1 hr; all antibodies were incubated in 5% milk in TBST 0.1% at 4° C overnight; blots were washed in TBST 0.1%. The abundance of β-actin (ACTB) was monitored to ensure equal loading (**Figure 4-10B**). Immunoblotting was performed using the antibodies: mouse monoclonal Anti-Flag-M2 (Sigma-Aldrich, #1804, clone M2, 1:500; https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma/Bulletin/f1804bul.pdf), mouse monoclonal C-Cas9 Anti-SpyCas9 (Sigma-Aldrich, #SAB4200751, clone 10C11-A12, 1:500; https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma/Datasheet/10/sab4200751dat.pdf), mouse monoclonal N-Cas9 Anti-SpyCas9 (Novus Biologicals, #NBP2-36440, clone 7A9-3A3, 1:500; https://www.novusbio.com/PDFs2/NBP2-36440.pdf), HRP-conjugated mouse monoclonal Anti-Beta-Actin (Santa Cruz Biotechnology, #sc-47778 HRP, clone C4, 1:250;

https://datasheets.scbt.com/sc-47778.pdf), and HRP-conjugated sheep Anti-Mouse (GE Healthcare Amersham ECL, #NXA931; 1:5000; https://es.vwr.com/assetsvc/asset/es_ES/id/9458958/contents). Blots were exposed using Amersham ECL Western Blotting Detection Reagent (GE Healthcare Amersham ECL, #RPN2209) and imaged using a ChemiDoc MP imaging system (Bio-Rad). Protein ladders were used as molecular weight reference (Bio-Rad, #161-0374).

4.5.18. Mammalian competitive proliferation assay

For assessment of CRISPR-Cas programmed cell depletion using guide RNAs targeting an essential gene (RPA1) or sgCIDEs targeting hundreds of thousands of loci within the genome, cells were stably transduced with a lentiviral vector expressing Cas9-wt (pCF226) or ProCas9Flavi (pCF730), and selected on puromycin. Subsequently, these cell lines were further stably transduced with vectors expressing various mCherry-tagged sgRNAs and analyzed as follows: (1) After mixing sgRNA expressing populations with parental cells, the fraction of mCherry-positive cells was quantified over time. Different sgRNAs targeting a neutral gene (sgOR2B6), an essential gene (sgRPA1), > 100,000 genomic loci (sgCIDE) and a non-targeting control (sgNT) were compared. (2) Alternatively, the cell lines were partially transduced with lentiviral vectors expressing a GFP-tagged dTEV (pCF736) or WNV (pCF738) protease, and cell depletion quantified by flow cytometry. We quantified depletion of protease-expressing (GFP+) cells among the sgRNA-positive (mCherry+) population.

4.5.19. Statistical analysis

Specific statistical tests used are indicated in all cases. Propagation of uncertainty was taken into consideration when reporting data and their uncertainty (standard deviation) as functions of measurement variables. Unless otherwise noted, error bars indicate the standard deviation of triplicates, and significance was assessed by comparing samples to their respective controls using unpaired, two-tailed t tests (alpha = 0.05). Genome editing quantification using TIDE was carried out as recommended[146]. In brief, indels ranging from −10 to +10 nucleotides were quantified. Parental cells were used as reference for normalization. When reporting TIDE editing efficiencies, only indels with p values < 0.01 in at least one replicate were considered true.

4.5.20. Data and software availability

To identify functional Cas9 circular permutants (Cas9-CPs), fold-changes for each dCas9-CP between pre- and post-library sorts along with significance values for each enrichment were calculated (Table S2). Cas9-CP analysis scripts are available at https://github.com/SavageLab/cpCas9. All relevant sequencing data have been deposited in the National Institutes of Health (NIH) Sequencing Read Archive (SRA) at https://www.ncbi.nlm.nih.gov/bioproject/PRJNA505363 under ID code 505363, Accession code PRJNA505363.

Chapter 5. References

1. Venturelli, O. S., Egbert, R. G. & Arkin, A. P. Towards Engineering Biological Systems in a Broader Context. *J. Mol. Biol.* **428,** 928–944 (2016).
2. Goodman, A. L. *et al.* Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* **6,** 279–289 (2009).
3. Wu, M. *et al.* Genetic determinants of in vivo fitness and diet responsiveness in multiple human gut Bacteroides. *Science* **350,** aac5992 (2015).
4. Cole, B. J. *et al.* Genome-wide identification of bacterial plant colonization genes. *PLoS Biol* **15,** e2002860 (2017).
5. Mao, N., Cubillos-Ruiz, A., Cameron, D. E. & Collins, J. J. Probiotic strains detect and suppress cholera in mice. *Sci Transl Med* **10,** eaao2586 (2018).
6. Kurtz, C. B. *et al.* An engineered E. coli Nissle improves hyperammonemia and survival in mice and shows dose-dependent exposure in healthy humans. *Sci Transl Med* **11,** eaau7975 (2019).
7. Egbert, R. G. *et al.* A versatile platform strain for high-fidelity multiplex genome editing. *Nucleic Acids Res* **11,** 367–13 (2019).
8. Oakes, B. L. *et al.* CRISPR-Cas9 Circular Permutants as Programmable Scaffolds for Genome Modification. *Cell* **176,** 254–267.e16 (2019).
9. Schnoes, A. M., Brown, S. D., Dodevski, I. & Babbitt, P. C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. *PLoS Comput Biol* **5,** e1000605–13 (2009).
10. Clark, W. T. & Radivojac, P. Analysis of protein function and its prediction from amino acid sequence. *Proteins* **79,** 2086–2096 (2011).
11. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat Meth* **10,** 221–227 (2013).
12. Nichols, R. J. *et al.* Phenotypic Landscape of a Bacterial Cell. *Cell* **144,** 143–156 (2011).
13. Deutschbauer, A. *et al.* Evidence-Based Annotation of Gene Function in Shewanella oneidensis MR-1 Using Genome-Wide Fitness Profiling across 121 Conditions. *PLoS Genet* **7,** e1002385–17 (2011).

14. Deutschbauer, A. *et al.* Towards an informative mutant phenotype for every bacterial gene. *J. Bacteriol.* **196,** 3643–3655 (2014).

15. Price, M. N. *et al.* The genetic basis of energy conservation in the sulfate-reducing bacterium Desulfovibrio alaskensis G20. *Front. Microbiol.* **5,** 577 (2014).

16. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology* **2,** 1–11 (2006).

17. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460,** 894–898 (2009).

18. Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B. A. & Gill, R. T. Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. *Nat Biotechnol* **28,** 856–862 (2010).

19. Freed, E. F. *et al.* Genome-Wide Tuning of Protein Expression Levels to Rapidly Engineer Microbial Traits. *ACS Synth. Biol.* **4,** 1244–1253 (2015).

20. Garst, A. D. *et al.* Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat Biotechnol* **35,** 1–12 (2016).

21. van Opijnen, T., Bodi, K. L. & Camilli, A. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat Meth* **6,** 767–772 (2009).

22. Langridge, G. C. *et al.* Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. *Genome Research* **19,** 2308–2316 (2009).

23. Wetmore, K. M. *et al.* Rapid Quantification of Mutant Fitness in Diverse Bacteria by Sequencing Randomly Bar-Coded Transposons. *mBio* **6,** e00306–15–15 (2015).

24. Price, M. N. *et al.* Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature* **557,** 503–509 (2018).

25. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression. *Cell* **152,** 1173–1183 (2013).

26. Peters, J. M. *et al.* A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. *Cell* **165,** 1–39 (2016).

27. Liu, X. *et al.* High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae. *Molecular Systems Biology* **13,** 931–18 (2017).

28. Wang, T. *et al.* Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. *Nature Communications* **9,** 1–15 (2018).

29. Rousset, F., Cui, L., Siouve, E., Depardieu, F. & Bikard, D. Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. 1–31 (2018). doi:10.1101/308916

30. de Wet, T. J., Gobe, I., Mhlanga, M. M. & Warner, D. F. CRISPRi-Seq for the Identification and Characterisation of Essential Mycobacterial Genes and Transcriptional Units. 1–24 (2018). doi:10.1101/358275

31. Lee, H. H. *et al.* Functional genomics of the rapidly replicating bacterium Vibrio natriegens by CRISPRi. *Nature Microbiology* **4,** 1105–1113 (2019).

32. Gottesman, S. & Storz, G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology* **3,** a003798–a003798 (2011).

33. Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & van Oudenaarden, A. Multistability in the lactose utilization network of Escherichia coli. *Nature* **427,** 737–740 (2004).

34. Somvanshi, V. S. *et al.* A single promoter inversion switches Photorhabdus between pathogenic and mutualistic states. *Science* **337,** 88–93 (2012).

35. Oren, Y. *et al.* Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 16112–16117 (2014).

36. Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354,** 769–773 (2016).

37. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell* **66,** 285–299.e5 (2017).

38. Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549,** 111–115 (2017).

39. Zhu, S. *et al.* Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat Biotechnol* **34,** 1279–1286 (2016).

40. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355,** eaah7111–16 (2017).

41. Joung, J. *et al.* Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature* **548,** 343–346 (2017).

42. Kato, J.-I. & Hashimoto, M. Construction of consecutive deletions of the Escherichia coli chromosome. *Molecular Systems Biology* **3,** 966–7 (2007).

43. Osterman, A. L. & Gerdes, S. Y. *Microbial Gene Essentiality: Protocols and Bioinformatics*. **416,** (Humana Press, 2008).

44. Gerdes, K. & Maisonneuve, E. Bacterial Persistence and Toxin-Antitoxin Loci. *Annu. Rev. Microbiol.* **66,** 103–123 (2012).

45. Kint, C. I., Verstraeten, N., Fauvart, M. & Michiels, J. New-found fundamentals of bacterial persistence. *Trends in Microbiology* **20,** 577–585 (2012).

46. Verstraeten, N. *et al.* Obg and Membrane Depolarization Are Part of a Microbial Bet-Hedging Strategy that Leads to Antibiotic Tolerance. *Molecular Cell* **59,** 9–21 (2015).

47. Pedersen, K. & Gerdes, K. Multiple hok genes on the chromosome of Escherichia coli. *Mol. Microbiol.* **32,** 1090–1102 (1999).

48. Fuchs, J. A. & Karlström, H. O. Mapping of nrdA and nrdB in Escherichia coli K-12. *J. Bacteriol.* **128,** 810–814 (1976).

49. Garriga, X. *et al.* nrdD and nrdG genes are essential for strict anaerobic growth of Escherichia coli. *Biochem. Biophys. Res. Commun.* **229,** 189–192 (1996).

50. Haft, D. H. *et al.* TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res* **41,** D387–D395 (2012).

51. Klumpp, S., Zhang, Z. & Hwa, T. Growth Rate-Dependent Global Effects on Gene Expression in Bacteria. *Cell* **139,** 1366–1375 (2009).

52. Deutschbauer, A. M. Mechanisms of Haploinsufficiency Revealed by Genome-Wide Profiling in Yeast. *Genetics* **169,** 1915–1925 (2005).

53. Ardell, D. H. & Kirsebom, L. A. The Genomic Pattern of tDNA Operon Expression in E. coli. *PLoS Comput Biol* **1,** e12–14 (2005).

54. Couturier, E. & Rocha, E. P. C. Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. *Mol. Microbiol.* **59,** 1506–1518 (2006).

55. Hunter, W. N. The non-mevalonate pathway of isoprenoid precursor biosynthesis. *Journal of Biological Chemistry* **282,** 21573–21577 (2007).

56. Wong, B. G., Mancuso, C. P., Kiriakov, S., Bashor, C. J. & Khalil, A. S. Precise, automated control of conditions for high-throughput growth of yeast and bacteria with eVOlVer. *Nature Publishing Group* **67,** 1–15 (2018).

57. Yan, B., Boitano, M., Clark, T. A. & Ettwiller, L. SMRT-Cappable-seq reveals complex operon variants in bacteria. *Nature Communications* **9,** 318–11 (2018).

58. Rousset, F. *et al.* Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. *PLoS Genet* **14,** e1007749–28 (2018).

59. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nature Protocols* **8,** 2180–2196 (2013).

60. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337,** 816–821 (2012).

61. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).

62. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40,** 4288–4297 (2012).

63. Khalil, A. S. & Collins, J. J. Synthetic biology: applications come of age. *Nat Rev Genet* **11,** 367–379 (2010).

64. Weber, W. & Fussenegger, M. Emerging biomedical applications of synthetic biology. *Nat Rev Genet* **13,** 21–35 (2011).

65. Tyo, K. E. J., Ajikumar, P. K. & Stephanopoulos, G. Stabilized gene duplication enables long-term selection-free heterologous pathway expression. *Nat Biotechnol* **27,** 760–765 (2009).

66. Friehs, K. Plasmid copy number and plasmid stability. *Adv Biochem Eng Biotechnol* **86,** 47–82 (2004).

67. Bassalo, M. C. *et al.* Rapid and Efficient One-Step Metabolic Pathway Integration in E. coli. *ACS Synth. Biol.* **5,** 561–568 (2016).

68. Lee, J. W. *et al.* Creating Single-Copy Genetic Circuits. *Molecular Cell* **63,** 329–336 (2016).

69. Esvelt, K. M. & Wang, H. H. Genome-scale engineering for systems and synthetic biology. *Molecular Systems Biology* **9,** 641–641 (2013).

70. Bryant, J. A., Sellars, L. E., Busby, S. J. W. & Lee, D. J. Chromosome position effects on gene expression in Escherichia coli K-12. *Nucleic Acids Res* **42,** 11383–11392 (2014).

71. Murphy, K. C. Use of bacteriophage lambda recombination functions to promote gene replacement in Escherichia coli. *J. Bacteriol.* **180,** 2063–2071 (1998).

72. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci USA* **97,** 6640–6645 (2000).

73. Ellis, H. M., Yu, D., DiTizio, T. & Court, D. L. High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proc Natl Acad Sci USA* **98,** 6742–6746 (2001).

74. Datta, S., Costantino, N. & Court, D. L. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379,** 109–115 (2006).

75. Sharan, S. K., Thomason, L. C., Kuznetsov, S. G. & Court, D. L. Recombineering: a homologous recombination-based method of genetic engineering. *Nature Protocols* **4,** 206–223 (2009).

76. Isaacs, F. J., Carr, P. A., Wang, H. H. & Lajoie, M. J. Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* (2011). doi:10.1126/science.1204763

77. Lajoie, M. J. *et al.* Genomically recoded organisms expand biological functions. *Science* **342,** 357–360 (2013).

78. Lajoie, M. J. *et al.* Probing the limits of genetic recoding in essential genes. *Science* **342,** 361–363 (2013).

79. Wang, H. H. *et al.* Genome-scale promoter engineering by coselection MAGE. *Nat Meth* **9,** 591–593 (2012).

80. Wang, H. H. *et al.* Multiplexed in vivo His-tagging of enzyme pathways for in vitro single-pot multienzyme catalysis. *ACS Synth. Biol.* **1,** 43–52 (2012).

81. Zeitoun, R. I. *et al.* Multiplexed tracking of combinatorial genomic mutations in engineered cell populations. *Nat Biotechnol* **33,** 1–10 (2015).

82. Zeitoun, R. I., Pines, G., Grau, W. C. & Gill, R. T. Quantitative Tracking of Combinatorially Engineered Populations with Multiplexed Binary Assemblies. *ACS Synth. Biol.* **6,** 619–627 (2017).

83. Glickman, B. W. & Radman, M. Escherichia coli mutator mutants deficient in methylation-instructed DNA mismatch correction. *Proc Natl Acad Sci USA* **77,** 1063–1067 (1980).

84. Schaaper, R. M. & Dunn, R. L. Spectra of spontaneous mutations in Escherichia coli strains defective in mismatch correction: the nature of in vivo DNA replication errors. *Proc Natl Acad Sci USA* **84,** 6220–6224 (1987).

85. Sawitzke, J. A. *et al.* Recombineering: in vivo genetic engineering in E. coli, S. enterica, and beyond. *Meth. Enzymol.* **421,** 171–199 (2007).

86. Wang, H. H., Xu, G., Vonner, A. J. & Church, G. Modified bases enable high-efficiency oligonucleotide-mediated allelic replacement via mismatch repair evasion. *Nucleic Acids Res* **39,** 7336–7347 (2011).

87. Nyerges, Á. *et al.* Conditional DNA repair mutants enable highly precise genome engineering. *Nucleic Acids Res* **42,** e62–e62 (2014).

88. Nyerges, Á. *et al.* A highly precise and portable genome engineering method allows comparison of mutational effects across bacterial species. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 2502–2507 (2016).

89. Bubnov, D. M., Yuzbashev, T. V., Vybornaya, T. V., Netrusov, A. I. & Sineoky, S. P. Development of new versatile plasmid-based systems for λRed-mediated Escherichia coli genome engineering. *Journal of Microbiological Methods* **151,** 48–56 (2018).

90. Lennen, R. M. *et al.* Transient overexpression of DNA adenine methylase enables efficient and mobile genome engineering with reduced off-target effects. *Nucleic Acids Res* **44,** e36–e36 (2016).

91. Sergueev, K., Yu, D., Austin, S. & Court, D. Cell toxicity caused by products of the p(L) operon of bacteriophage lambda. *Gene* **272,** 227–235 (2001).

92. Lawther, R. P. *et al.* Molecular basis of valine resistance in Escherichia coli K-12. *Proc Natl Acad Sci USA* **78,** 922–925 (1981).

93. Lawther, R. P. *et al.* DNA sequence fine-structure analysis of ilvG (IlvG+) mutations of Escherichia coli K-12. *J. Bacteriol.* **149,** 294–298 (1982).

94. Tedin, K. & Norel, F. Comparison of ΔrelA strains of Escherichia coli and Salmonella enterica serovar Typhimurium suggests a role for ppGpp in attenuation regulation of branched-chain …. *J. Bacteriol.* (2001). doi:10.1128/JB.183.21.6184-6196.2001

95. Lajoie, M. J., Gregg, C. J., Mosberg, J. A., Washington, G. C. & Church, G. M. Manipulating replisome dynamics to enhance lambda Red-mediated multiplex genome engineering. *Nucleic Acids Res* **40,** e170–e170 (2012).

96. Mosberg, J. A., Gregg, C. J., Lajoie, M. J., Wang, H. H. & Church, G. M. Improving lambda red genome engineering in Escherichia coli via rational removal of endogenous nucleases. *PLoS ONE* **7,** e44638 (2012).

97. Glascock, C. B. & Weickert, M. J. Using chromosomal lacIQ1 to control expression of genes on high-copy-number plasmids in Escherichia coli. *Gene* **223,** 221–231 (1998).

98. Choi, Y. J. *et al.* Novel, versatile, and tightly regulated expression system for Escherichia coli strains. *Appl. Environ. Microbiol.* **76,** 5058–5066 (2010).

99. Khlebnikov, A., Skaug, T. & Keasling, J. D. Modulation of gene expression from the arabinose-inducible araBAD promoter. *J Ind Microbiol Biotechnol* **29,** 34–37 (2002).

100. Casini, A. *et al.* R2oDNA designer: computational design of biologically neutral synthetic DNA sequences. *ACS Synth. Biol.* **3,** 525–528 (2014).

101. Gama-Castro, S. *et al.* RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res* **44,** D133–43 (2016).

102. Bipatnath, M., Dennis, P. P. & Bremer, H. Initiation and velocity of chromosome replication in Escherichia coli B/r and K-12. *J. Bacteriol.* **180,** 265–273 (1998).

103. Reynolds, T. S. & Gill, R. T. Quantifying Impact of Chromosome Copy Number on Recombination in Escherichia coli. *ACS Synth. Biol.* **4,** 776–780 (2015).

104. Sauer, C. *et al.* Effect of Genome Position on Heterologous Gene Expression in Bacillus subtilis: An Unbiased Analysis. *ACS Synth. Biol.* **5,** 942–947 (2016).

105. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium Escherichia coli as determined by whole-genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **109,** E2774–83 (2012).

106. Stringer, A. M. *et al.* FRUIT, a scar-free system for targeted chromosomal mutagenesis, epitope tagging, and promoter replacement in Escherichia coli and Salmonella enterica. *PLoS ONE* **7,** e44841 (2012).

107. Sarkar, S., Ma, W. T. & Sandri, G. H. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica* **85,** 173–179 (1992).

108. Ma, W. T., Sandri, G. H. & Sarkar, S. Analysis of the Luria–Delbrück distribution using discrete convolution powers. *Journal of Applied Probability* 255–267 (1992).

109. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).

110. Fellmann, C., Gowen, B. G., Lin, P.-C., Doudna, J. A. & Corn, J. E. Cornerstones of CRISPR–Cas in drug discovery and therapy. *Nat Rev Drug Discov* **16,** 89–100 (2016).

111. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2,** e00471 (2013).

112. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).

113. Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155,** 1479–1491 (2013).

114. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159,** 1–15 (2014).

115. Guilinger, J. P., Thompson, D. B. & Liu, D. R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat Biotechnol* **32,** 577–582 (2014).

116. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat Biotechnol* **33,** 510–517 (2015).

117. 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,** 1–17 (2016).

118. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat Biotechnol* **32,** 569–576 (2014).

119. Richter, F. *et al.* Switchable Cas9. *Current Opinion in Biotechnology* **48,** 119–126 (2017).

120. Kim, K. *et al.* Genome surgery using Cas9 ribonucleoproteins for the treatment of age-related macular degeneration. *Genome Research* **27,** 419–426 (2017).

121. Staahl, B. T. *et al.* Efficient genome editing in the mouse brain by local delivery of engineered Cas9 ribonucleoprotein complexes. *Nat Biotechnol* **35,** 431–434 (2017).

122. Zuris, J. A. *et al.* Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nat Biotechnol* **33,** 73–80 (2015).

123. Roybal, K. T. *et al.* Precision Tumor Recognition by T Cells With Combinatorial Antigen-Sensing Circuits. *Cell* **164,** 770–779 (2016).

124. Davis, K. M., Pattanayak, V., Thompson, D. B., Zuris, J. A. & Liu, D. R. Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nature Chemical Biology* **11,** 316–318 (2015).

125. Hemphill, J., Borchardt, E. K., Brown, K., Asokan, A. & Deiters, A. Optical Control of CRISPR/Cas9 Gene Editing. *J. Am. Chem. Soc.* **137,** 5642–5645 (2015).

126. Oakes, B. L. *et al.* Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nat Biotechnol* 1–8 (2016). doi:10.1038/nbt.3528

127. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nature Publishing Group* **12,** 326–328 (2015).

128. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A Protein-Tagging System for Signal Amplification in Gene Expression and Fluorescence Imaging. *Cell* **159,** 1–12 (2014).

129. Yu, Y. & Lutz, S. Circular permutation: a different way to engineer enzyme structure and function. *Trends in Biotechnology* **29,** 18–25 (2011).

130. Beernink, P. T. *et al.* Random circular permutation leading to chain disruption within and near alpha helices in the catalytic chains of aspartate transcarbamoylase: effects on assembly, stability, and function. *Protein Sci.* **10,** 528–537 (2001).

131. Mehta, M. M., Liu, S. & Silberg, J. J. A transposase strategy for creating libraries of circularly permuted proteins. *Nucleic Acids Res* **40,** e71 (2012).

132. Qian, Z. & Lutz, S. Improving the catalytic activity of Candida antarctica lipase B by circular permutation. *J. Am. Chem. Soc.* **127,** 13466–13467 (2005).

133. Whitehead, T. A., Bergeron, L. M. & Clark, D. S. Tying up the loose ends: circular permutation decreases the proteolytic susceptibility of recombinant proteins. *Protein Eng. Des. Sel.* **22,** 607–613 (2009).

134. Plainkum, P., Fuchs, S. M., Wiyakrutta, S. & Raines, R. T. Creation of a zymogen. *Nat. Struct. Biol.* **10,** 115–119 (2003).

135. Jones, A. M. *et al.* The Structure of a Thermophilic Kinase Shapes Fitness upon Random Circular Permutation. *ACS Synth. Biol.* **5,** 415–425 (2016).

136. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 1–16 (2014).

137. Oakes, B. L., Nadler, D. C. & Savage, D. F. in *The Use of CRISPR/Cas9, ZFNs, and TALENs in Generating Site-Specific Genome Alterations* **546,** 491–511 (Elsevier, 2014).

138. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551,** 464–471 (2017).
139. Johnson, R. J., Lin, S. R. & Raines, R. T. A ribonuclease zymogen activated by the NS3 protease of the hepatitis C virus. *FEBS Journal* **273,** 5457–5465 (2006).
140. Butler, J. S., Mitrea, D. M., Mitrousis, G., Cingolani, G. & Loh, S. N. Structural and thermodynamic analysis of a conformationally strained circular permutant of barnase. *Biochemistry* **48,** 3497–3507 (2009).
141. Seon Han, J., Kim, D.-H. & Yong Choi, K. in *Handbook of Proteolytic Enzymes* 2427–2432 (Elsevier, 2013). doi:10.1016/B978-0-12-382219-2.00542-1
142. Skern, T. in *Handbook of Proteolytic Enzymes* 2396–2402 (Elsevier, 2013). doi:10.1016/B978-0-12-382219-2.00535-4
143. Tomlinson, K. R., Bailey, A. M., Alicai, T., Seal, S. & Foster, G. D. Cassava brown streak disease: historical timeline, current knowledge and future prospects. *Molecular Plant Pathology* **19,** 1282–1294 (2017).
144. Bera, A. K., Kuhn, R. J. & Smith, J. L. Functional characterization of cis and trans activity of the Flavivirus NS2B-NS3 protease. *Journal of Biological Chemistry* **282,** 12883–12892 (2007).
145. Kümmerer, B. M., Amberg, S. M. & Rice, C. M. in *Handbook of Proteolytic Enzymes* 3112–3120 (Elsevier, 2013). doi:10.1016/B978-0-12-382219-2.00687-6
146. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res* **42,** e168 (2014).
147. Ramanathan, M. P. *et al.* Host cell killing by the West Nile Virus NS2B-NS3 proteolytic complex: NS3 alone is sufficient to recruit caspase-8-based apoptotic pathway. *Virology* **345,** 56–72 (2006).
148. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350,** 1096–1101 (2015).
149. Ade, J., DeYoung, B. J., Golstein, C. & Innes, R. W. Indirect activation of a plant nucleotide binding site-leucine-rich repeat protein by a bacterial protease. *Proc Natl Acad Sci USA* **104,** 2531–2536 (2007).
150. Chisholm, S. T. *et al.* Molecular characterization of proteolytic cleavage sites of the Pseudomonas syringae effector AvrRpt2. *Proc Natl Acad Sci USA* **102,** 2087–2092 (2005).

151. Kim, S. H., Qi, D., Ashfield, T., Helm, M. & Innes, R. W. Using decoys to expand the recognition specificity of a plant disease resistance protein. *Science* **351,** 684–687 (2016).

152. Alfano, J. R. & Collmer, A. Type III secretion system effector proteins: double agents in bacterial disease and plant defense. *Annu Rev Phytopathol* **42,** 385–414 (2004).

153. Gao, M. *et al.* The protease of herpes simplex virus type 1 is essential for functional capsid formation and viral growth. *J. Virol.* **68,** 3702–3712 (1994).

154. Hartmann, S. & Lucius, R. Modulation of host immune responses by nematode cystatins. *Int. J. Parasitol.* **33,** 1291–1302 (2003).

155. Baltes, N. J. *et al.* Conferring resistance to geminiviruses with the CRISPR–Cas prokaryotic immune system. *Nature Plants 2015 1:10* **1,** 15145 (2015).

156. Chaparro-Garcia, A., Kamoun, S. & Nekrasov, V. Boosting plant immunity with CRISPR/Cas. *Genome Biology* **16,** 254 (2015).

157. Kennedy, E. M. *et al.* Inactivation of the human papillomavirus E6 or E7 gene in cervical carcinoma cells by using a bacterial CRISPR/Cas RNA-guided endonuclease. *J. Virol.* **88,** 11965–11972 (2014).

158. Mehta, D., Stürchler, A., Hirsch-Hoffmann, M., bioRxiv, W. G.2018. CRISPR-Cas9 interference in cassava linked to the evolution of editing-resistant geminiviruses. *biorxiv.org* doi:10.1101/314542

159. Carette, J. E. *et al.* Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* **477,** 340–343 (2011).

160. Coradetti, S. T. *et al.* Functional genomics of lipid metabolism in the oleaginous yeast Rhodosporidium toruloides. *Elife* **7,** 283 (2018).

161. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106 (2010).

162. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nature Publishing Group* **11,** 783–784 (2014).

163. Perez, A. R. *et al.* GuideScan software for improved single and paired CRISPR guide RNA design. *Nat Biotechnol* **35,** 347–349 (2017).

164. Fellmann, C. *et al.* An optimized microRNA backbone for effective single-copy RNAi. *CellReports* **5,** 1704–1713 (2013).

165. Park, H. M. *et al.* Extension of the crRNA enhances Cpf1 gene editing in vitro and in vivo. *Nature Communications* **9,** 3313 (2018).