Minimization of Cas9 and Perspectives on Genetically Engineered Microorganisms and Their Regulation

By

Arik Shams

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor David Savage, Chair
Professor Susan Marqusee
Assistant Adjunct Professor Ross Wilson
Associate Professor John Dueber

Fall 2022

Abstract

Minimization of Cas9 and Perspectives on Genetically Engineered Microorganisms and Their Regulation

By
Arik Shams

Doctor of Philosophy in Molecular and Cell Biology
University of California, Berkeley

Professor David Savage, Chair

CRISPR technology has revolutionized the way biology is conducted. In a few years, scientists have broken through barriers that have hamstrung the field for decades, thanks to the conceptually simple ability to target specific sequences in an organism's genome. That ability alone has led to enormous leaps in our understanding of complex traits, biological pathways, disease, and evolution. Additionally, genome editing with CRISPR has ushered in a new age of therapeutics and genetic engineering. As new applications of CRISPR technology emerge, we are beginning to push the boundaries of what it is capable of, and new modalities of CRISPR are required to overcome new challenges. For example, despite its reputation of being the "flagship" CRISPR molecule, the *S. pyogenes* Cas9 (a.k.a. SpCas9), is still too large to be genetically encoded and delivered by adeno-associated viruses (AAVs). As AAVs are one of the predominant delivery mechanisms for gene therapies, precluding SpCas9 from its repertoire is a significant deficiency.

SpCas9 has been studied extensively to date. However, certain aspects of the molecule and its mechanisms are still unknown and evade our understanding even as we expand our gene-editing toolkit to include new Cas9s, new CRISPR systems, and new platforms. One of the questions associated with SpCas9 is its multi-domain architecture, which is atypically complex and large for bacterial proteins. While the general molecular mechanism of Cas9 is understood thanks to structural and kinetics studies, the evolution of the overall protein and its domains are less understood. This is especially interesting considering how little sequence identity Cas9s share among orthologs while still possessing a similar three-dimensional architecture.

Additionally, SpCas9's combination of stability, precision, and versatility makes it a prime candidate for protein engineering. Researchers have tried adding new functions to SpCas9, improving its performance, and even making it more suitable for certain use cases using rational design principles. However, making the protein smaller has been an under-utilized concept with limited success. Minimizing SpCas9 while still retaining its DNA targeting function has two main functions: a) understand the essentiality of its domains for DNA targeting, and b) develop a novel protein scaffold that is smaller and more feasible for AAV and other forms of cellular delivery. In Chapter 2 of

this thesis, I discuss a project to probe the amino acid deletion landscape of SpCas9, in an effort to biochemically characterize its domains, and also to arrive at a minimal DNA-binding protein module.

CRISPR-Cas9 has undoubtedly changed the biotechnology world by advancing genetic engineering and breaking through technical barriers that have plagued the field for decades. Simultaneously, the maturation of several other synthetic biology tools has also converged into a new era of biotechnology. Genome editing, sequencing, bioinformatics, gene synthesis, etc. have brought upon a revolution of new bioengineered products to the market. Leading the charge are genetically engineered microorganisms, or GEMs. GEMs are being developed for the purpose of making biofuels, commodity chemicals, materials, food and food additives, pesticides, and fertilizers. Many of these products have promise as more efficient, eco-friendly, and overall more beneficial alternatives to conventional industries.

However, adoption and deployment of these technologies are not as simple as building them. Understandably, bioengineered products require regulatory oversight to ensure safety and efficacy. In the U.S., biotechnology products are regulated by a tri-agency framework consisting of the Food and Drug Administration (FDA), U.S. Department of Agriculture (USDA), and the Environmental Protection Agency (EPA). These agencies have distinct roles and evaluation criteria when assessing new biotech products for market approval. However, one of the key issues with the current regulatory landscape in the U.S. is its complicated and circuitous nature that often delay and/or drive up costs of product development. To maximize the impact of all the innovation and benefit of scientific advances happening in the laboratory, regulation needs to be more streamlined without compromising safety.

In Chapter 3 of this thesis, I describe the current state of GEMs in the U.S., from the development and regulation perspectives. I discuss the product areas in which GEMs are emerging as feasible and scalable alternatives to conventional industrial processes, such as fuels, food, and agriculture. I also attempt to explain the current tri-agency regulatory framework as set by the FDA, USDA, and EPA. Finally, I discuss ways in which both GEM regulators and product developers must work hand in hand to enact large-scale solutions to major modern problems like climate change and food insecurity.

Acknowledgments

It has taken me six long years to complete my PhD at UC Berkeley. I can confidently say that I am not the same person as I was when I first arrived at Berkeley in August 2016. Looking back, it is clear to me today how I am not only a product of the MCB program, but probably more so the people I have met and had relationships with over the years. I would like to first express my gratitude to the people to whom I owe a career in science. First, Dr. Anthony Bell, who took me under his wing as an untrained, inexperienced undergraduate student in his biochemistry lab at the University of Southern Mississippi. I owe him my love of research, even though I didn't know it at the time, and I was convinced I was still going to medical school. He saw in me potential I genuinely didn't believe I had and pushed me to seek out research as a postbac. He is a supporter of mine to this day (in fact, at the time of writing this, Dr. Bell had just texted me a few hours prior to ask me how I was doing). Next, I want to thank Dr. Jonathan Bird, who mentored me as a postbac at the NIDCD-LMG, and from whom I learned a tremendous amount of biochemistry. I also learned from him many things that are required to be a good scientist beyond just the bench, namely organization, confidence in my own process, collaboration, and patience. He is also an active supporter of mine to this day, and I hope for many more years to come. I also want to thank Dr. Tom Friedman, LMG Lab Chief, who took a chance on hiring me as a postbac and passionately believed in training students to be excellent scientists.

I want to thank Dr. Dave Savage, my PhD advisor at Berkeley. There are many, many positive qualities about Dave I could expound upon, including his kindness, generosity, passion, sense of humor, etc. etc. but I'll settle for this: I think young trainee scientists end up being the type of scientist(s) who mentored them, and I would be happy and proud to be anything like Dave.

On the topic of mentors, I want to thank Dr. Melinda Kliegman, who has been incredibly helpful and supportive in my transition from academic science to science policy. I truly could not have asked for a better guide and resource as I take this leap in my career.

Next, I want to thank the people in the Savage Lab, who are the best lab in the whole world. I truly enjoyed spending all these years with you, and I deeply cherish the culture we have cultivated and fiercely protect. Emeric (I'm so glad I joined the Savage Lab), Rob, Avi, Jack, Julia, Noam, Evan, Shin, Andrew, Maria, Brittney, Naiya, Eli, Sean, Tom, Luke, Luis, Jorge, Rachel, Flora, David, Emily, Dylan, and many others over the years, I love and appreciate you all.

Next, I want to thank my friends and fellow MCBers at Berkeley. I honestly would not have survived this far without all of you. Tess, Justin, Paige, Dan, Josh, Maya, Ellie, Gabriella, Phil, Josh Tworig, Huntly, Matt, Michael, Samira, Laura, Mathew, Davis, Shally, George, Amy, Perri, Camila, and everyone else in the MCB (and MCB-adjacent) grad student community: I love you all so much and hope we remain friends forever. Each of you has helped me in ways I can't even describe, but collectively I'm here because of you.

Finally, I want to thank my family. This PhD is for you and because of you. Maa and Baba: I can't really express in words how grateful I am to you both for the sacrifices you've made so I could come this far but know that everything I've done the past six years was possible because of you, and I'll always believe that. Apu: thank you for being there always when I needed you. Just knowing I could count on you when I needed you was enough for me to forge ahead so many times in my life. To my dulabhai Anthony: I love you and appreciate you, and I am so happy to have you as part of my family. Thank you for supporting Apu and all of us. To my yet unborn niece Maya: I can't wait to meet you and be the best uncle. To be honest I'm terrified I might not be, but I know you're coming into the most loving family. I also want to acknowledge my family members who couldn't see me come this far today: Dadu and Nanu, whom I miss dearly; Boro Fupi, Chhoto Fupi, Chhoto Chacha, Khalamoni, whom I know would be so incredibly proud of me. I hope I continue to make you proud the rest of my life.

Table of Contents

Chapter 1

Introduction

1.1 The basics of CRISPR engineering

The invention of CRISPR genome-editing technology in 2012 has accelerated our understanding and manipulation of biology and ushered in a new biotechnology revolution in areas like human therapeutics, diagnostics, genetic and organismal engineering, sequencing, and basic research. A simple and precise, as well as efficient and accessible, tool to make targeted and programmable double-strand breaks (DSBs) in DNA has been an ambitious goal among biologists for many decades, and the invention of the Cas9 endonuclease coupled with a single guide-RNA (sgRNA) has manifested exponentially more tools to manipulate an organism's genome. In its basic, natural function, a Cas9 is activated by binding to the trans-activating CRISPR RNA (tracrRNA), which then anneals to the "homing" component of the complex, the CRISPR RNA (crRNA). Together, they form the guide-RNA (gRNA) that "homes in" on a target DNA sequence. Jinek et. al. showed in 2012 that the functions of the two RNA molecules – tracrRNA and crRNA – can be combined into a single-guide RNA (sgRNA), that not only simplifies the use of Cas9 by reducing it to a two-component system, but also shows enhanced efficacy 1. At its core, CRISPR gene editing is simple in principle: making DSBs allows endogenous cellular DNA repair machinery to delete or insert bases, thereby knocking out gene function or knocking in new sequences, respectively 2. Templated repair pathways enable insertion of sequences into the DSB locus, which can be artificially provided in many forms, e.g. plasmids, linear oligomers, single-stranded DNA, or even RNA. This has led to the discovery and characterization of complex DNA-repair pathways in multiple organisms, which have been further linked to studies of cancer, aging, and other genetic conditions 3,4.

Because of ease of use, CRISPR has also been used to manipulate the genome of previously recalcitrant non-model organisms, boosting the diversity of biological systems available for study. The utility of CRISPR is not limited to its function as "molecular scissors," as the ability of Cas molecules to target specific sequences of DNA alone has massively boosted the invention of new applications of the technology. By fusing transcription factors to a catalytically-dead Cas9 (dCas9), scientists have been able to modulate transcription of a gene via CRISPR-interference (CRISPRi) or CRISPR-activation (CRISPRa) [5]. Tethering epigenetic factors to Cas molecules have also allowed the precise perturbation of the epigenome [6]. In recent years, CRISPR has opened up new genome editing modalities like base-editing, where single-nucleotide changes can be made by fusing and targeting deaminases to specific sequences without the need for DNA breaks. Prime-editing is another example where a reverse transcriptase is targeted to a locus using an attenuated Cas9 molecule, and new DNA bases are inserted in templated fashion complementary to a provided RNA template, circumventing DSBs and allowing more precise insertion of new sequences [7,8].

1.1.1 Overview of Cas9

The advancement of CRISPR technology and its applications have been very rapid in the laboratory in the past several years, and soon began to push the limits of the original Cas9 molecule derived from the *Streptococcus pyogenes* bacterium. New homologs of Cas9, as well as entirely new phage defense systems in bacteria were quickly discovered and adopted, diversifying the genome editing toolkit even further. Even among Cas9 orthologs, there is a remarkable amount of interspecies variability. For example, the two common Cas9 molecules, from *S. pyogenes* (SpCas9) and *Staphylococcus aureus* (SaCas9), share less than 30% amino acid sequence identity, despite having very similar mechanisms and structural elements [9]. Another important difference between the two are their cognate sgRNAs, which are of different size, sequence, and secondary structure. However, they are both important and distinct in their usage in biology: SpCas9 requires a NGG protospacer-adjacent motif (PAM) to identify its targets to then initiate gRNA-mediated binding to DNA, while SaCas9 requires a NNGRRT PAM. This means that they can be used to target different genomic loci, and can also be exploited selectively for applications where multiple targets need to be targeted simultaneously. Other Cas9 orthologs from type II-C CRISPR systems have since been discovered and characterized in their editing efficacy, a select few of which include those from *Francisella novicida* (FnCas9), *Neisseria meningitidis* (Nme1Cas9 and Nme2Cas9), and *Geobacillus stearothermophilus* (GeoCas9) [10–12].

1.1.2 Engineering Cas9

Although there are many scenarios where alternative Cas9s are useful, so far SpCas9 has remained the benchmark for its combination of efficacy, precision, stability, and simplicity. However, there are aspects of the molecule that restrict its applications in some cases. One of these aspects is its size: SpCas9 is composed of 1368 amino acids, is ~160 kDa, and has a relatively large hydrodynamic radius of 10 nm when complexed with its gRNA [13]. SpCas9 is encoded by 4.1 kbp of DNA, putting it just over the limit of adeno-associated virus (AAV) payloads – especially when including its single-guideRNA (100 bases) and associated promoters and terminators. As AAVs are one of the most common and tractable ways to deliver gene therapies and editing reagents into cells, this precludes SpCas9 from many applications. Its large, multidomain structure is also a potential issue when delivered in ribonucleoprotein (RNP) form, lowering passive uptake by cells via endocytosis. While many of these issues can be circumvented or compensated for by overengineering (e.g. increasing AAV capsid size, fusing cell-penetrating units to the protein, chemical transfection), altering the Cas9 protein molecule itself was largely unexplored [14].

In early studies of SpCas9, the structure and mechanisms of the molecule were characterized by rationally deleting its domains and measuring function. While these studies revealed plenty in a general sense, they were not systematic in exploring the roles of each of these domains and how they could be exploited to make Cas9 more

"streamlined" and more suited for specific applications. Briefly, SpCas9 is composed of two lobes known as the RECognition and NUClease lobes. The REC lobe consists of the REC1, REC2, REC3 and C-terminal domains (CTD), and is responsible for binding the sgRNA, recognizing and annealing to the DNA target (the "spacer"). The NUC lobe consists of the HNH and RuvC nuclease domains, which cut the target and non-target strands of dsDNA, respectively. Binding to the spacer triggers a conformational change in the REC lobe, and allosterically moves the nuclease domains into the catalytic center to make the DSB [2].

Previously, deletion of the REC2 domain found nominal to no effect on DNA binding ability, but did not explore the observation further or suggest a function for the presence of the domain, despite knowledge from structural studies that showed that the REC2 domain underwent conformational rearrangement upon target binding. Another hint to the ambiguity of REC2 was the fact that SaCas9, which operates under very similar molecular mechanisms according to structure-function studies, does not possess a REC2 domain, yet is able to recognize and bind DNA – albeit with significant differences from SpCas9. SpCas9 is considered a single-turnover enzyme, as it possesses relatively strong affinity for the spacer and remains bound many hours after initial binding and cleavage to a perfectly complementary target. SaCas9, however, is multiple-turnover, as it binds, cleaves, and releases its target many times in the same timescales. Compared to SpCas9, SaCas9 is observed to have a much higher dissociation rate (k_d) for its spacer [15]. In summary, all these clues pointed to an unknown function for the REC2 domain, and possibly others in SpCas9.

1.1.3 Minimization of Cas9

Through the lens of protein engineering, SpCas9 was a promising candidate for "minimization," which means paring down the protein to uncover the pieces that are essential for a particular function. Developing an approach to make protein size smaller while retaining (or in some cases, adding) a desired function is of particular interest in the case of Cas9, for the delivery challenges described above [16]. However, size is an underexplored parameter in the field of protein evolution, and experimental evidence for minimization is sparse mainly due to library complexity, lack of diversity, complicated cloning and screening, and low throughput [17–19].

The evolutionary unit of proteins is the domain, defined as functional subunits of the larger protein structure that are independently folded and have some role in the overall protein function [20,21]. Deleting an entire domain will expectedly have some consequence on the overall function or stability of a protein, but that may not be the case if the deletions were smaller in size, such as a single or few amino acids. Therefore, developing such a system to make every possible contiguous amino acid deletion in a protein of interest would be highly useful, not only to create novel, miniaturized proteins, but to probe the molecular function and evolutionary trajectories of proteins. A systematic deletion landscape of a protein sequence can reveal the essentiality of amino acids and domains under screening for a particular parameter. In the case of Cas9, the

most valuable parameter would be accurate DNA binding. In the first chapter of my thesis, I describe a project where I and co-author Dr. Sean Higgins, developed a technique called MISER (Minimization by Iterative Size-Exclusion and Recombination), and applied it to SpCas9 to make minimized versions of the protein and uncover potentially new roles for certain domains.

1.2 Genetically engineered microbes: their present, future, and regulation

One of the key applications unlocked by CRISPR genome editing is the rapid development of new, engineered microorganisms. Previously, precisely engineering an organism's traits involved specialized tools that were often unique to the organism being edited, such as using plasmids with organism-specific promoters in yeast and *E. coli*. CRISPR opened up the possibility of making persistent and reliable edits in genomes of previously intractable and non-model microorganisms [22]. Simultaneously, the emergence of orthogonal technologies for synthetic biology, such as bioinformatic data mining, deep sequencing of microbial communities, and massively parallel gene synthesis have blown open the field of microbial biotechnology [23–25].

The convergence of these technologies comes at a critical time, when we are actively experiencing the effects of climate change, extreme weather, food insecurity, loss of biodiversity, and energy shortages. Large-scale, existential problems like these require an equally large, total response from all levels of the scientific, policymaking, and civic communities. From a technology standpoint, biology has a lot to offer to mitigate some of these issues, especially pertaining to food, energy, materials, and remediation. Microorganisms have been exploited for human use since the beginning of civilization, and in this new era of biotechnology, may unlock new potential in the fight for survival, in the form of genetically engineered microorganisms, or GEMs. Many bacteria, algae, yeasts, and fungi hold the key to making new biofuels more efficiently and with far less harm to the environment, they may be used as food sources, additives, or supplements to agriculture industries like fertilizer or pesticides.

However, as the world moves towards a "green economy," many of these new products have to overcome policy and regulatory hurdles that delay their utility for public good. In the U.S, the regulatory landscape is convoluted and non-cohesive due to bureaucratic inertia. For science to be effective, these innovations must leave the lab and be deployed at scale, without compromising safety. For these reasons, there must be a drastic rethinking of the roles of regulatory agencies in the U.S., especially when it comes to GEMs.

1.2.1 The rapid evolution of GEMs

Microorganisms and microbiology have been exploited for human uses since the dawn of civilization, even before knowledge of their existence. As our understanding of them became more and more sophisticated, so too did our manipulation of them to make

food, medicine, fuels, materials, and also to improve their resilience, productivity, and safety. Traditional methods of microbial genetic engineering (GE) relied on applying selective pressure and random mutagenesis. Although imprecise, these methods have been quite successful historically, and even today many modern products are made with microorganisms that were randomly mutagenized (e.g. penicillin) [26]. However, modern biotechnology has rendered these approaches obsolete thanks to our ever-improving understanding of the genetic basis of traits, and our ability to make more precise and predictable changes to an organism's genome. Newer technologies like high-throughput sequencing and CRISPR have enormously accelerated the development of new microbes.

Modern problems require modern solutions, and with the climate crisis endangering food and energy security, many private and public entities are looking towards genetically engineered microorganisms, or GEMs, with renewed interest. GEMs have the potential to usher in a "green revolution" in the food, fuel, and materials industries. In the United States, companies are actively using modern GE technologies to rapidly develop and characterize microbially-produced products that are more environmentally friendly, safer, less resource-intensive, and efficient. Prominent ones include LanzaTech, who are developing bacteria that can use industrial waste-gases as feedstock for biofuels; bioMason, who are developing GEMs for making cement through environmental carbon capture; and Pivot Bio, who are making bacteria-based biofertilizer for nitrogen uptake in plants.

1.2.2 The slow regulation of GEMs

Bringing the products to market and implementing them is a different story, however, and one which is quickly becoming less of a scientific and technical barrier and more of a policy-based one. U.S. regulation of GE products is done by three agencies (aka tri-agency): the Food and Drug Administration (FDA), the U.S. Department of Agriculture (USDA), and the Environmental Protection Agency (EPA). While these agencies have been vigilant in evaluating all recombinant DNA products since the adoption of the Coordinated Framework in 1986, the scientific and technological advancement of genetic engineering has far outpaced these agencies. The Coordinated Framework for the Regulation of Biotechnology is a policy framework that specified the roles of the tri-agency system, and laid out governing principles for how new biotechnology products would be evaluated for the market [27]. The guidelines generally state that regulation of GE products would be based solely on the characteristics of the final product and not the process by which they were made, and any regulatory burden placed on the stakeholder would be based on scientific evidence and rationale.

While this approach may appear logically sound and fair for stakeholders and consumers, the problem lies not with the standards set by the regulatory agencies but the process by which they each individually operate. The tri-agency regulatory landscape is inherently fragmented, with each agency serving their own mandate for ensuring public and economic safety. Briefly, the USDA is charged with ensuring the protection of the

U.S.'s natural and agricultural resources, such as its commercial supply of livestock, poultry, eggs, etc. Plant and animal pests are of particular concern to the USDA with regard to GEMs, which are therefore assessed with an eye towards pathogenesis in commercially-important resources. The FDA regulates all food and food additives that may reach a consumer, ensuring purity, safety, and efficacy of biological products. The FDA may also issue guidelines related to proper labeling of food products on the market. The EPA is responsible for ensuring the safety of products with regards to environmental exposure and safety. This includes GEMs that are released into the environment, substances produced by GEMs in the environment, and exposure to hazards in the processing and manufacture of GEMs [28].

Even within the three agencies, there are multiple offices that are responsible for different aspects of the evaluation process. These definitions of the agencies' roles were laid out in the 2017 update of the Coordinated Framework. It is obvious from the descriptions of the agencies' roles that there is a huge amount of overlap in regulatory standards for any particular GEM or GEM-derived product. However, the agencies operate separately with little to no collaboration between them when a new product is presented for approval. A developer that has a new GEM intended for market has to petition each agency separately for approval, and the burden is on the developer to assess the safety and quality of their product and present supportive data to each agency.

For example, a developer – defined as a stakeholder person or company who is creating a new product – may invent a new bacterial strain that can be sprayed directly onto wheat grown in farmland in the U.S. that can protect against fungal infection. The developer would have to seek approval from the USDA to ensure the bacterium was not harmful to wheat or a threat to any other crop. They would have to seek approval from the EPA to ensure that the bacteria were not an environmental hazard and produced no harmful byproducts. They may also have to petition the FDA to assure them any bacteria that persisted in the final wheat product would be safe for consumption. It is up to the developer to find and understand the guidelines for each agency, provide supporting data, and pursue an approval or exemption separately. They would likely have to employ multiple science and policy experts to navigate each agency's rules. Any one agency could request additional information, or simply be slow to process the application, which in turn could slow down the entire process for the developer and likely incur more costs. These bureaucratic and political challenges severely clog the laboratory-to-market pipeline for new products and slow down the actual implementation of scientific breakthroughs.

1.2.3 The future landscape of GEMs

Without question, the safety of the public and the environment are paramount and cannot be compromised. However, the costs associated with a complex regulatory environment fall on more than just the company. Slowing down approval and adoption of GE products precludes the public from their benefits, as well. In 2012, the

Obama White House released the National Bioeconomy Blueprint, which recognized the importance and rapid progress of biotechnology in the U.S. and the world. The document proposed several goals for the government to support innovation in biotech, which included streamlining the regulatory process for GE products. Since then little has changed, and in 2022 the Biden administration released a more comprehensive proposal called the National Biotechnology and Biomanufacturing Initiative (NBBI), which contained a substantial investment of $2 billion to support the bioeconomy in various capacities. The need for streamlining the regulatory process was included yet again in the NBBI ten years later. At the time of this writing, there is not even a unified website that may direct a developer of a GEM product to the applicable guidelines and/or agency office among the EPA, USDA, and FDA.

As mentioned previously, the ongoing climate and population crises are existential threats that require an extremely broad and diverse response from all facets of society, including private and public sectors. There are many potential biology-based stopgaps and solutions that are yet to be deployed at scale, especially in areas of food, biofuels, and carbon sequestration. Specifically, climate adaptation in plants [29], nitrogen uptake efficiency by crops [30], and water use efficiency [31] are all feasible and scalable technologies that should be deployed immediately with as little regulatory burden as possible without compromising safety. As the bioeconomy grows, the regulatory process must keep up not only to ensure safety of the products but to also maximize their public impact.

1.3 References

1. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
2. Jiang, F. & Doudna, J. A. CRISPR-Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.* **46**, 505–529 (2017).
3. Zhang, H. *et al.* Application of the CRISPR/Cas9-based gene editing technique in basic research, diagnosis, and therapy of cancer. *Mol. Cancer* **20**, 126 (2021).
4. Doudna, J. A. The promise and challenge of therapeutic genome editing. *Nature* **578**, 229–236 (2020).
5. Singh, V. Chapter 1 - An introduction to genome editing CRISPR-Cas systems. in *Genome Engineering via CRISPR-Cas9 System* (eds. Singh, V. & Dhar, P. K.) 1–13 (Academic Press, 2020). doi:10.1016/B978-0-12-818140-9.00001-5.
6. Nakamura, M., Gao, Y., Dominguez, A. A. & Qi, L. S. CRISPR technologies for precise epigenome editing. *Nat. Cell Biol.* **23**, 11–22 (2021).
7. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
8. Scholefield, J. & Harrison, P. T. Prime editing - an update on the field. *Gene Ther.* **28**, 396–401 (2021).
9. Nishimasu, H. *et al.* Crystal Structure of Staphylococcus aureus Cas9. *Cell* **162**, 1113–1126 (2015).
10. Acharya, S. *et al. Francisella novicida* Cas9 interrogates genomic DNA with very high specificity and can be used for mammalian genome editing. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20959–20968 (2019).
11. Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15644–15649 (2013).
12. Harrington, L. B. *et al.* A thermostable Cas9 with increased lifetime in human plasma. *Nat. Commun.* **8**, 1424 (2017).
13. Wei, T., Cheng, Q., Min, Y.-L., Olson, E. N. & Siegwart, D. J. Systemic nanoparticle delivery of CRISPR-Cas9 ribonucleoproteins for effective tissue specific genome editing. *Nat. Commun.* **11**, 3232 (2020).
14. Wang, D., Zhang, F. & Gao, G. CRISPR-Based Therapeutic Genome Editing: Strategies and In Vivo Delivery by AAV Vectors. *Cell* **181**, 136–150 (2020).
15. Yourik, P., Fuchs, R. T., Mabuchi, M., Curcuru, J. L. & Robb, G. B. Staphylococcus aureus Cas9 is a multiple-turnover enzyme. *RNA* **25**, 35–44 (2019).
16. Ma, D., Peng, S., Huang, W., Cai, Z. & Xie, Z. Rational Design of Mini-Cas9 for Transcriptional Activation. *ACS Synth. Biol.* **7**, 978–985 (2018).
17. Simm, A. M., Baldwin, A. J., Busse, K. & Jones, D. D. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 beta-lactamase. *FEBS Lett.* **581**, 3904–3908 (2007).
18. Morelli, A., Cabezas, Y., Mills, L. J. & Seelig, B. Extensive libraries of gene truncation variants generated by in vitro transposition. *Nucleic Acids Res.* **45**, e78 (2017).

19. Arpino, J. A. J., Reddington, S. C., Halliwell, L. M., Rizkallah, P. J. & Jones, D. D. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure* **22**, 889–898 (2014).
20. Teichmann, S. A., Park, J. & Chothia, C. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14658–14663 (1998).
21. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300**, 1701–1703 (2003).
22. Shapiro, R. S., Chavez, A. & Collins, J. J. CRISPR-based genomic tools for the manipulation of genetically intractable microorganisms. *Nat. Rev. Microbiol.* **16**, 333–339 (2018).
23. Harbola, A., Negi, D., Manchanda, M. & Kesharwani, R. K. Chapter 27 - Bioinformatics and biological data mining. in *Bioinformatics* (eds. Singh, D. B. & Pathak, R. K.) 457–471 (Academic Press, 2022). doi:10.1016/B978-0-323-89775-4.00019-5.
24. Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J. Clin. Invest.* **132**, (2022).
25. Song, L.-F., Deng, Z.-H., Gong, Z.-Y., Li, L.-L. & Li, B.-Z. Large-Scale de novo Oligonucleotide Synthesis for Whole-Genome Synthesis and Data Storage: Challenges and Opportunities. *Front Bioeng Biotechnol* **9**, 689797 (2021).
26. Adrio, J. L. & Demain, A. L. Genetic improvement of processes yielding microbial products. *FEMS Microbiol. Rev.* **30**, 187–214 (2006).
27. The Unified Website for Biotechnology Regulation. https://usbiotechnologyregulation.mrp.usda.gov/biotechnologygov/resources.
28. Modernizing the Regulatory System for Biotechnology Products: Final Version of the 2017 Update to the Coordinated Framework for the Regulation of Biotechnology. Preprint at https://usbiotechnologyregulation.mrp.usda.gov/2017_coordinated_framework_update.pdf (2017).
29. Karavolias, N. G., Horner, W., Abugu, M. N. & Evanega, S. N. Application of Gene Editing for Climate Change in Agriculture. *Frontiers in Sustainable Food Systems* **5**, (2021).
30. Wen, A. *et al.* Enabling Biological Nitrogen Fixation for Cereal Crops in Fertilized Fields. *ACS Synth. Biol.* **10**, 3264–3277 (2021).
31. Hatfield, J. L. & Dold, C. Water-Use Efficiency: Advances and Challenges in a Changing Climate. *Front. Plant Sci.* **10**, 103 (2019).

Chapter 2

Comprehensive deletion landscape of CRISPR-Cas9 identifies minimal RNA-guided DNA-binding modules*

* The work presented in this chapter is adapted from a previously published article with permission:

Shams A, Higgins SA, Fellmann C, Laughlin TG, Oakes BL, Lew R, Kim S, Lukarska M, Arnold M, Staahl BT, Doudna JA, Savage DF. Comprehensive deletion landscape of CRISPR-Cas9 identifies minimal RNA-guided DNA-binding modules. *Nat Commun.* 2021 Sep 27;12(1):5664. doi: 10.1038/s41467-021-25992-8. PMID: 34580310; PMCID: PMC8476515.

2.1 Abstract

Proteins evolve through the modular rearrangement of elements known as domains. It is hypothesized that extant, multidomain proteins are the result of domain accretion, but there has been limited experimental validation of this idea. Here, we introduce a technique for genetic minimization by iterative size-exclusion and recombination (MISER) for comprehensively making all possible deletions of a protein. Using MISER, we generated a deletion landscape for the CRISPR protein Cas9. We found that the catalytically-dead Streptococcus pyogenes Cas9 can tolerate large single deletions to the REC2, REC3, HNH, and RuvC domains, while still functioning *in vitro* and *in vivo*, and that these deletions can be stacked together to engineer minimal, DNA-binding effector proteins. In total, our results demonstrate that extant proteins retain significant modularity from the accretion process and, as genetic size is a major limitation for viral delivery systems, establish a general technique to improve genome editing and gene therapy-based therapeutics.

2.2 Introduction

Domains are the fundamental unit of protein structure[1–3]. Domains are also the unit of evolution in proteins, accumulating incremental mutations that change their function and stability, as well as being recombined within genomes to create new proteins via insertions, fusions, or deletions[4–7]. Extant multidomain proteins are thus thought to have evolved via the continuous accretion of domains to gain new function[4,8,9]. Additionally, eukaryotic proteome diversity is vastly increased by alternative splicing, which tends to insert or delete protein domains[10]. The phenomenon of domain modularity in proteins has been exploited synthetically to rearrange and expand the architecture of a protein, enabling new functionality[11–13]. For example, the programmable DNA nuclease Cas9 can be converted into a ligand-dependent allosteric switch using advanced molecular cloning, similar to other domain insertions dictated by allostery[13,14]. Although there are several methods for comprehensively altering protein topology[15,16], no method has been demonstrated for domain deletion.

Rationally constructed protein deletions have long been essential to elucidating functional and biochemical properties but are generally limited to a handful of truncations. Moreover, protein engineering can make use of deletions to alter enzyme substrate specificity[17], enable screens for improved activity and thermostability[18], or minimize protein size[19]. Early approaches to protein deletion libraries resulted in the deletion of single amino acids using an engineered transposon[20,21]. Other methods utilize direct PCR[22], random nuclease digestion[23], or random *in vitro* transposition followed by a complicated cloning scheme[24] to achieve deletion libraries containing a variety of lengths and reading frames. These techniques are low in throughput and/or require complex molecular techniques which poorly capture library diversity; in contrast to protein insertions where library size grows linearly with target length, deletion libraries grow as the square.

A simple and efficient method for building protein deletions coupled with a selection strategy would provide the ability to comprehensively query and delineate the function of domains or motifs in complex and multi-domain proteins. Such a technique could be used to identify crucial functions within multidomain proteins or splicing variants in a manner akin to how deep mutational scanning can be used to identify the effects of single nucleotide polymorphisms on functionality[25]. Moreover, with sufficient modularity, the evolutionary path of domain accretion could be explored through iterative combining, or 'stacking,' of domain deletions to isolate a minimal, core protein for a defined function[7–9].

One attractive target for such a strategy is *Streptococcus pyogenes* Cas9 (SpCas9), the prototypal RNA-guided DNA endonuclease used for genome editing[26]. SpCas9 is an excellent model protein for a comprehensive deletion study because of its multi-domain architecture and availability of high-throughput assays for either DNA cutting or binding[27]. Functionally, SpCas9 targets and cleaves DNA in a multi-step process. First, an apo Cas9 molecule forms a complex with a guide RNA (gRNA), containing a 19-22 bp variable "spacer" sequence that is complementary to a DNA target locus. The primed ribonucleoprotein (RNP) complex then surveils genomic DNA for a protospacer-adjacent motif (PAM) – 5'-NGG-3' in the case of SpCas9, where N is any nucleobase – that initiates a transient interaction with the protein to search for an adjacent ~20-bp target sequence. If a target is present, the double-stranded DNA (dsDNA) helix is unwound, allowing the gRNA to anneal to the DNA and form a stable RNA-DNA hybrid structure called an R-loop (see illustration in Fig. S8). Formation of a complete 20-bp R-loop triggers a conformational change in Cas9 to form the catalytically-active complex[28–30].

SpCas9 has a bi-lobed architecture consisting of the RECognition lobe, responsible for recognizing and binding DNA sequences, and the NUClease lobe, which possesses HNH and RuvC domains that cut the target and non-target strands of DNA, respectively. Cas9 is postulated to have evolved via domain accretion from a progenitor RuvC domain[9,31]. As a consequence, Cas9 orthologs possess manifold architectures. For example, the SpCas9 REC lobe possesses three domains (REC1, REC2 and REC3) while the *Staphylococcus aureus* Cas9 (SaCas9) has a contiguous REC domain without REC2[32,33]. The function of REC2 is ambiguous but is thought to act as a conformational switch to trigger DNA cleavage[34,35], raising the question of how SaCas9 accomplishes the effect[36]. Thus, the multi-domain, multi-functional nature of Cas9s make them an excellent model system for exploring domain deletions. Relatedly, Cas9's large size also complicates its delivery using viral vectors. Knowledge of functional deletions may thus facilitate the delivery of genome editing therapeutics.

Here, we introduce genetic minimization by iterative size-exclusion and recombination (MISER), a technique for systematically exploring in-frame deletions within a protein. Application of MISER to the catalytically dead SpCas9 (dCas9) identified regions of the protein which can be deleted with minimal consequence to binding function. Furthermore, we stacked individual deletions to engineer novel CRISPR Effector (CE) proteins that are less than 1000 amino acids in length. CRISPRi and biochemical assays demonstrated that these variants remain competent for target DNA-

binding but are less functional than single deletion variants. Finally, to understand the structural consequence of deletion, we used single-particle cryo-electron microscopy to solve a 6.2 Å structure of the smallest, 874 amino acid CE. This structure surprisingly revealed an overall conformation that preserves essential functions of SpCas9—emphasizing the concept of domains as independent modules—even though the quaternary structure is severely modified.

2.3 Results

2.3.1 MISER reveals the comprehensive deletion landscape of SpCas9

The general concept of MISER is to create a pool of all possible contiguous deletions of a protein and analyze them in a high-throughput fitness assay. The process can then be iterated to stack deletions together. We created such a library by: i) systematically introducing two distinct restriction enzyme sites, each once, across a gene on an episomal plasmid, ii) excising the intervening sequence using the restriction enzymes and iii) re-ligating the resulting fragments (Fig. 1A). In the instantiation here, two separate restriction enzymes (NheI and SpeI) with compatible sticky ends are used. Cleavage, removal of intervening sequence, and ligation thus results in a two-codon scar site (encoding either Ala-Ser or Thr-Ser) not recognized by either enzyme, thereby increasing efficiency of cloning and enabling iteration of the entire process (Fig. S1).

The MISER library was made for nuclease-dead Cas9 (dCas9) as follows. First, single NheI or SpeI sites were systematically introduced into a dCas9 gene with flanking BsaI sites using a targeted oligonucleotide library and recombineering (Fig. S1)[37,38]. Second, these plasmid libraries were isolated, digested respectively with BsaI and either NheI or SpeI, and then ligated together (Fig. S1B). The resulting ligation of gene fragments produces deletions, as well as duplications, such that a MISER library has a triangular distribution, with near-wildtype (WT) length proteins most frequent and the largest deletions least frequent (Fig. 1C). To empirically determine the size range of functional deletions, the dCas9 MISER library was separated on an agarose gel and divided into six sublibrary slices of increasing deletion size. The sublibraries were then independently cloned into expression vectors and assayed for bacterial CRISPRi GFP repression via flow cytometry (Fig. 1B, S2)[39,40]. Sublibrary Slice 4 (ranging from 3.2-3.5 kb) was the most stringent (i.e. smallest) library with detectable repression, and functional variants became more frequent in slices possessing smaller deletions, as expected.

Fluorescence-activated Cell Sorting (FACS) and sequencing of MISER variants identified dCas9 deletion variants competent for DNA-binding. To focus sequencing on functional variants, Slice 4 and Slice 5 were sequenced pre- and post-FACS sorting, and the enrichment or depletion of individual variants was quantified (Fig. S3). Four large deletion regions were independently identified in both libraries. Although the libraries target different size ranges, their overlapping data were significantly correlated (Fig. S3). These data were normalized and combined to generate a comprehensive landscape of

functional dCas9 deletions (Fig. 1C). 80% of sequencing depth was focused on deletions from 150 to 350 amino acids in length (Slice 4), and 51.4% (115,530/224,718) of these deletions were detected. Overall, this landscape includes data for 27.5% of all possible dCas9 deletions (257,737/936,396). The four large deletion regions roughly corresponded to the REC2, REC3, HNH, and RuvC-III domains. While larger deletions are bounded between domain termini, small deletions and insertions (~10 amino acids) are tolerated in much of the structure (Fig. S4), a finding that has been previously observed in other proteins[17,22]. Two clear exceptions are the mechanistically essential 'bridge helix'[35], which orders and stabilizes the R-loop[41,42], and the 'phosphate lock loop'[43], which interacts with the PAM-proximal target strand phosphate to enable gRNA strand invasion. It should be noted that the enrichment data presented here is somewhat sparse and only a relative measurement of CRISPRi function; the larger-scale features of acceptable domain and sub-domain level deletions were therefore extensively validated with further *in vivo* and biochemical assays.

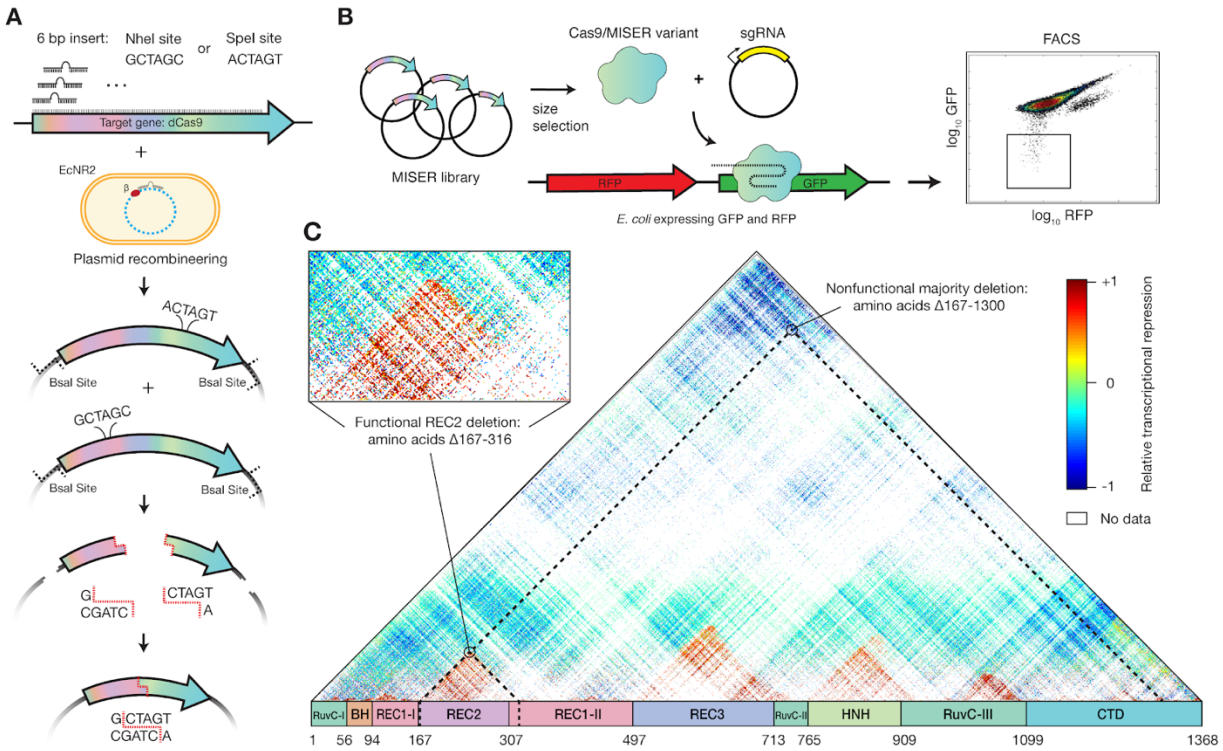


Figure 2.1: Minimization by Iterative Size Exclusion and Recombination (MISER). A) MISER library construction. A 6-bp SpeI or NheI recognition site is inserted separately into a dCas9-encoding plasmid flanked by BsaI sites using plasmid recombineering. The resultant libraries are digested with BsaI and either SpeI or NheI, and the two fragment pools are combined and ligated together to generate a library of dCas9 ORFs possessing all possible deletions. **B)** The MISER library is cloned into a vector and co-transformed in *E. coli* expressing RFP and GFP with an sgRNA targeting GFP. The library products are expressed, functional variants bind to the target, and repress the fluorophore. Repression activity *in vivo* is measured by flow cytometry. **C)** Enrichment map of the MISER deletion landscape of *S. pyogenes* dCas9. A single pixel within the map represents an individual variant that contains a deletion beginning

where it intersects with the horizontal axis moving to the left (N) and ends where it intersects with the axis moving to the right (C). Larger deletions are at the top, with some deletions almost spanning the whole protein. The heatmap shows relative repression activity of variants from two FACS sorts of a single replicate. The map is a composite of Slice 4 and Slice 5 in Figure S3 A-B, which present variant ratios post- versus pre-FACS sorting.

2.3.2 Cas9 tolerates large domain deletions while retaining DNA-binding function

To validate the deletion profile, individual variants from each of the four large deletion regions were either isolated from the library (Fig. S5) or constructed via PCR and assayed individually. Representative variants from these regions could be identified that exhibited bacterial CRISPRi nearly as effectively as full-length dCas9 (Fig. 2A, S5). Intriguingly, there are regions within our identified deletions that have been previously tested based on rational design, providing additional insight into the biochemical mechanisms lost with the removal of each domain[35,44]. The most obvious of the acceptable deletions is of the HNH domain that is responsible for cleaving the target strand and gating cleavage by the RuvC domain; it was thus of little surprise that deletions of HNH were tolerated in a molecule that is required to bind but not cleave DNA. In fact, Sternberg et al. previously demonstrated that a HNH-deleted (Δ768-919) Cas9 is competent for nearly WT levels of binding activity, but is unable to cleave[45]. In contrast, we also uncovered a deletion in the RuvC-III domain that has never been observed. Modeling this deletion on the previously determined structure of SpCas9 bound to a DNA-target (PDB ID 5Y36)[46] revealed that it removes a large set of loops, an alpha helix and two antiparallel beta sheets (Fig. S7). This deletion does not seem to overlay with a known functional domain and thus may serve as a module that further stabilizes the RuvC domain as a whole. Additionally, this deletion abuts the nontarget and target strand DNA (distance of ~4-6 Å) and may provide a highly useful site to replace with accessory fusions, such as deaminases suitable for base editing the non-target strand, as was engineered with circularly permuted base-editors[16,47].

Our observations for the REC2 and REC3 domains likewise expand upon two rationally engineered deletions. Chen et al. previously demonstrated that the REC3 domain gates the rearrangement of the HNH cleavage by sensing the extended RNA:DNA duplex[44]. Deletion of this domain (Δ497-713) ablated cleavage activity while maintaining full binding affinity. Nishimasu et. al. also previously deleted the REC2 domain because they postulated that it was unnecessary for DNA cleavage, as it is poorly conserved across other Cas9 sequences and lacks significant contact to the bound guide:target heteroduplex in the structure; however, the deletion mutant was found to have reduced activity[35].

To further validate the function and potential deficits of these single whole-domain deletions, we biochemically analyzed representative deletions of each of the REC2, REC3, HNH, and RuvC domains (Fig. 2B). These single-deletion constructs are henceforth referred to as ΔREC2 (residues 180-297 deleted), ΔREC3 (Δ503-708), ΔHNH (Δ792-897), and ΔRuvC (Δ1010-1081). Deletion variants were expressed, purified (see Supporting Information and Fig. S9 for purification data), and assayed for DNA binding

activity using bio-layer interferometry (BLI) (Fig. 2C, Fig. S10). Binding assays revealed that the REC2 deletion confers a defect in binding to a fully-complementary double-stranded DNA target (dsDNA) when complexed with a single-guide RNA (sgRNA). Interestingly, the defect is almost fully rescued upon the addition of a 3-bp mismatch bubble between the target and nontarget DNA strands adjacent to the PAM. DNA unwinding is initiated by Cas9 at the PAM-adjacent seed region, enabling the RNA-DNA R-loop hybrid to form. Rescue via seed bubble therefore suggests a potential role for the REC2 domain in unwinding dsDNA.

A similar phenomenon is observed with the ΔREC3 variant, although the binding defect is less pronounced than in ΔREC2. ΔREC3 is also unable to bind fully-complementary dsDNA – an effect that is rescued by the same PAM-adjacent 3-bp bubble in the dsDNA substrate, implying a similar DNA unwinding function by the REC3 domain. These results suggest that both the REC2 and REC3 domains are not essential for DNA binding by SpCas9 but may have evolved as "enhancer" domains to allow SpCas9 to more efficiently bind DNA inside the cell.

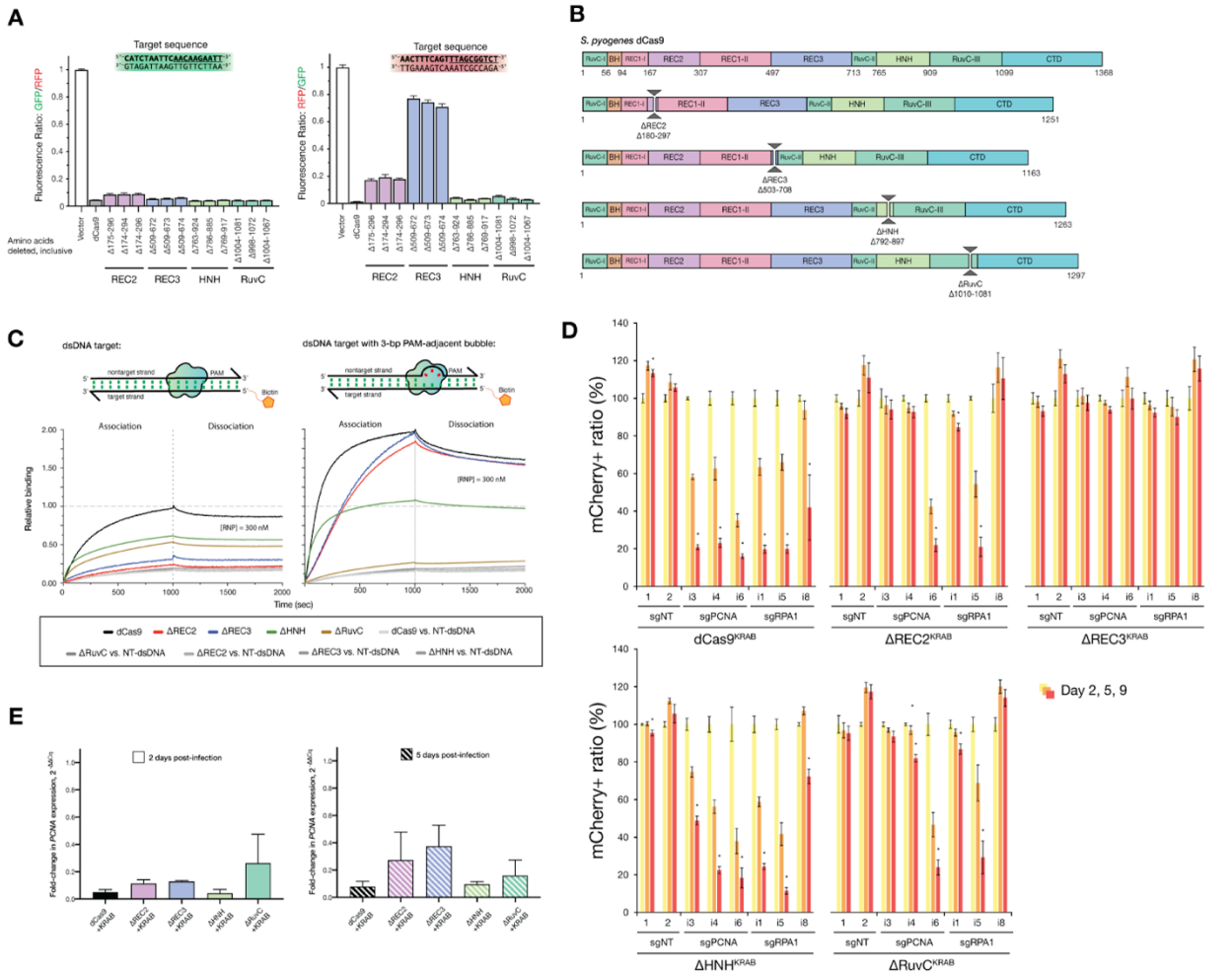
When measuring the repression activity of the ΔREC3 constructs *in vivo*, we also observed that the ΔREC3 appears to exhibit varying levels of repression between different gRNA sequences. Specifically, we found that a GFP-targeting gRNA repressed stronger than an RFP-targeting gRNA with ΔREC3, after controlling for cell growth and fluorophore maturity (Fig. 2A). This was unexpected, since the binding of WT Cas9 is generally thought to be gRNA sequence-agnostic[48]. One possibility is that the GC content of the targets in GFP and RFP could affect function, for example, a higher proportion of GC base-pairing in the "seed" region of a DNA target could present a greater energetic cost of unwinding to a deletion variant like ΔREC3[49]. Analysis of 16 additional spacer sequences and their repression activity relative to WT suggests this mechanism only moderately (R^2 = 0.2) explains the variance (Fig. S8). Further comprehensive analysis of the sequence-dependent variability is required to identify the precise energetic threshold the ΔREC3 construct overcomes to unwind DNA.

Similar binding experiments with ΔHNH and ΔRuvC showed that they possess activity intermediate to WT dCas9 and ΔREC2 or ΔREC3 (Fig. 2C). Surprisingly, adding a 3-bp mismatch bubble adjacent to the PAM does not seem to fully restore binding function. ΔHNH reaches ~50% binding upon addition of the bubble, performing worse than the ΔREC2 and ΔREC3 constructs upon addition of the bubble. The bubble also does not appear to increase ΔRuvC's binding to dsDNA (Fig. 2C). We speculate that the defect in binding may be due to the R-loop being destabilized by nuclease domain deletion but is stable enough for bulk repression of a fluorophore in culture.

To test whether the MISER constructs retain DNA binding activity in mammalian systems, we performed CRISPRi to knock down genes in a U-251 glioblastoma cell line. We transduced target cells with lentiviral vectors expressing our single-deletion MISER constructs (ΔREC2, ΔREC3, ΔHNH, and ΔRuvC) fused to the KRAB repressor domain, followed by selection on puromycin. Stable cell lines were then transduced with a secondary lentiviral vector expressing mCherry fluorescent protein and either an sgRNA targeting one of the essential genes *PCNA* (sgPCNA) or RPA1 (sgRPA1) or a control non-targeting sgRNA (sgNT). Transduced cells were mixed with the parental populations and

monitored for mCherry fluorescence by flow cytometry over several days. We observed that for dCas9 and three of the four single-deletion constructs (ΔREC2, ΔHNH, ΔRuvC), mCherry fluorescence is markedly lower at 5- and 9-days post-transduction, with multiple guides targeting PCNA and RPA1 (Fig. 2D). This suggested that the MISER-expressing mCherry-positive cell lines were repressing essential genes and were depleted from the population. The ΔREC3 construct exhibited little depletion, which is consistent with the BLI data (Fig. 2C) showing that ΔREC3 appears to have lower association compared to dCas9 and ΔREC2. Western blot data shows that the ΔREC3 is expressed at similar levels to the other single-deletion constructs (Fig. S11E), so it is unclear why this defect is observed in mammalian cells compared to bacterial repression (Fig. 2A). One possible explanation could be that the mammalian genome is packaged much differently from the bacterial genome, and DNA-targeting proteins have more difficulty accessing heterochromatin.

As the competition assay does not directly measure repression, RT-qPCR was used to quantitate expression of PCNA 2- and 5-days post-transduction. (Fig. S11). RT-qPCR of PCNA showed that after 2 days the ΔREC2, ΔREC3, ΔHNH and ΔRuvC constructs repress PCNA expression relative to a non-targeting gRNA (sgNT) (Fig. 2E; all measurements are averages ± S.D. from biological duplicates), with a mean fold-change of 0.11±0.03, 0.13±0.01, 0.04±0.03, and 0.26±0.21 in PCNA expression, respectively. At 5 days post-transfection, ΔREC2 and ΔREC3 appear to lose some repression activity (0.3±0.2 and 0.4±0.2 fold-change relative to sgNT, respectively), while the ΔHNH and ΔRuvC constructs are comparable to WT dCas9 at Day 5 (0.10±0.02 and 0.2±0.1 respectively) (see Supporting Information and Fig. S11 for more details on RT-qPCR). Thus, it appears that ΔREC3-KRAB is functional, but does not repress enough to generate a phenotype in our competition assay.



<u>Figure 2.2:</u> Cas9 tolerates whole-domain deletions while maintaining target-binding activity. A) *In vivo* transcription repression activity of MISER-dCas9 variants with specified amino-acids deleted, targeting either GFP (left) or RFP (right). dCas9s with REC2, REC3, HNH, or RuvC domain deletions have near-WT binding activity when targeted to GFP. When targeted to RFP, ΔREC2 and ΔREC3 show less robust binding activity. Data are normalized to vector-only control representing maximum fluorescence. Data are plotted as mean±SD from biological triplicates. **B)** Schema showing cloned MISER constructs with individual domain deletions corresponding to tolerated regions found in MISER screen. **C)** Bio-layer interferometry (BLI) assay of MISER constructs. ΔREC2 and ΔREC3 exhibit weak binding against a fully-complementary dsDNA target, while ΔHNH and ΔRuvC show intermediate binding. Binding is rescued to near-WT levels in ΔREC2 and ΔREC3, although at a slower rate, when the dsDNA contains a 3-bp bubble in the PAM-proximal seed region. Data are normalized to dCas9 binding to fully-complementary dsDNA. **D)** U-251 cells stably expressing the indicated MISER-dCas9 or WT-dCas9 KRAB fusion. Proteins were transduced with mCherry-tagged lentiviral vectors expressing sgRNAs targeting essential genes (sgPCNA, sgRPA1) or non-targeting controls (sgNT). At Day-2 post-transduction, cells were mixed with the respective parental population; mCherry fluorescence was monitored over time. Error bars indicate the SD of triplicates. Significance in cell depletion was assessed by comparing samples to their respective Day-2 controls using unpaired, two-tailed t-tests (alpha = 0.01). **E)** Measurement of CRISPRi efficacy of single-deletion MISER constructs in mammalian U-251 cells using RT-qPCR. U-251 cells were stably transduced with lentiviral vectors encoding dCas9 or MISER constructs fused with a KRAB repressor, along with lentivirus expressing sgRNA targeting *PCNA*. Cells were harvested 2 (left panel) or 5 (right) days post-

transduction of the sgRNA and assayed for PCNA expression. Bar graphs represent fold-change of PCNA expression relative to a non-targeting sgRNA. Error bars represent SD for at least 2 replicates.

2.3.3 Stacking MISER deletions results in minimal DNA-binding CRISPR proteins

Protein domains are accreted during the evolution of large proteins[3,4,50]. In principle, accretion could be experimentally reversed provided sufficient modularity is present to offset evolutionary divergence, epistasis, and other deleterious effects in 'stacked' deletions. To emulate this process, while also engineering a minimal Cas9-derived DNA-binding protein, we generated a library of constructs that consolidated the ΔREC2, ΔREC3, ΔHNH, and ΔRuvC deletions found by the MISER screen.

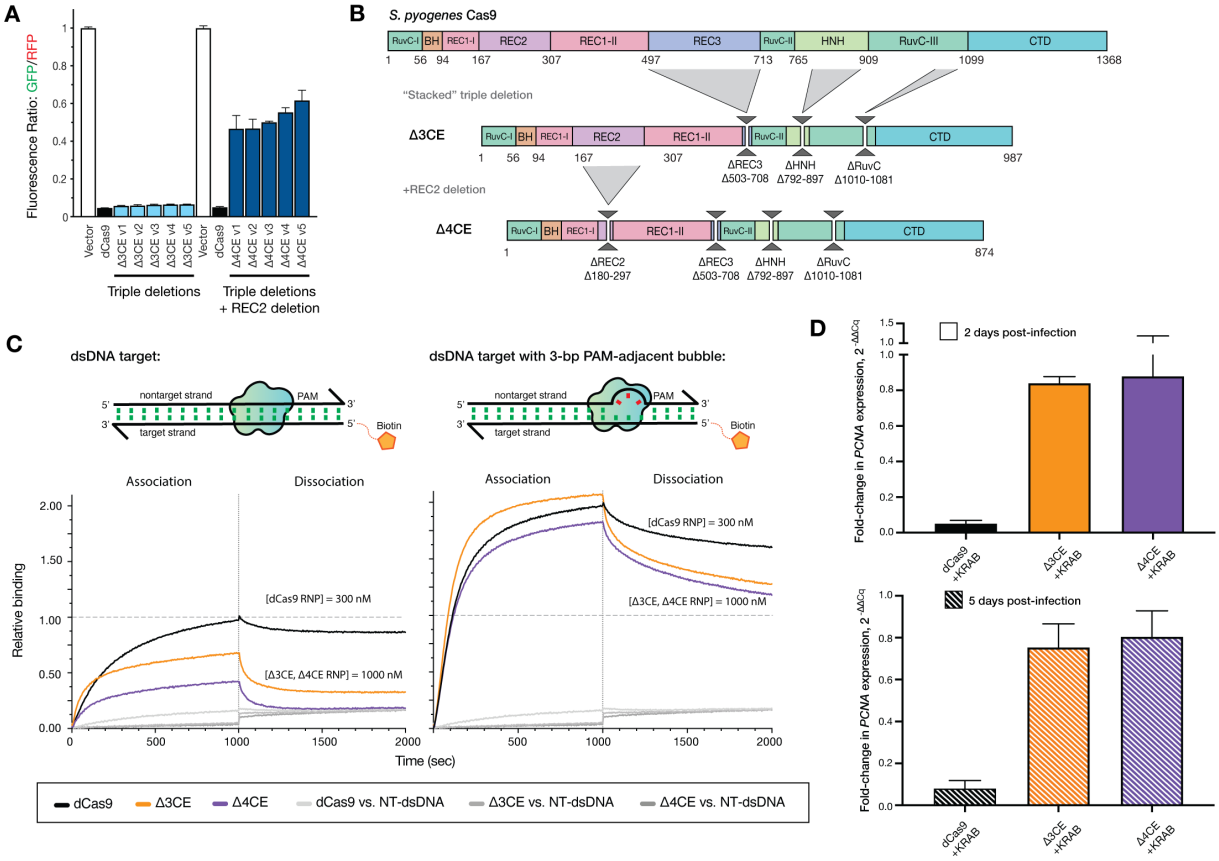
A library of multi-deletion variants, termed CRISPR Effectors (CE) due to their highly pared-down sequence relative to wild-type Cas9, were constructed as follows: individual sublibraries of deletions from REC2, REC3, and the HNH domains were isolated from the full MISER library. This was done by selecting against the full-length dCas9 sequence by targeting a pre-existing restriction site within each deleted region, so that only transformations of circular plasmids that had the respective deletion would be favored (Fig. 3B, S5). The RuvC deletion was an exception since it did not have a pre-existing restriction site; therefore, a manually constructed ΔRuvC variant (Δ1010-1081) was amplified and used as a starting point for further stacking.

The dCas9 gene was divided into four fragments spanning the major deletions and recombined using Golden Gate cloning (Fig. S6). The resulting library, CE Library 1, was assayed using bacterial CRISPRi, and functional variants were isolated by FACS, as above. A variety of functional CEs were obtained (Fig. 3A), although surprisingly, none of them possessed a REC2 deletion. We therefore generated a second library, CE Library 2, in which a library of triple-deletion variants was cross-bred with REC2 deletion variants to bias towards a deletion from this region (Fig. S6). Again, the most functional CE variants isolated by FACS did not contain REC2 deletions. Finally, in an attempt to force a minimal CE, the most active CE variant from CE Library 1 and 2, termed Δ3CE, was directly combined with a library of REC2 deletions and screened for activity. The resulting 'hard-coded' quadruple deletion CE variants all exhibited loss of function relative to WT (Fig. 3A), which explains why the REC2 deletion was lost in our functional variants. The most active variant (Δ4CE) possessed a deletion of Δ180-297 and was confirmed upon re-transformation to display ~50% the activity of WT dCas9 (Fig. 3A, 3C) in *E. coli*.

To validate the stacked deletion constructs biochemically, we expressed and purified the Δ3CE and Δ4CE variants from *E. coli* (Fig. 3B, S10). BLI experiments revealed that compared to the bacterial *in vivo* repression data, the DNA-binding abilities of both stacked deletion constructs were attenuated relative to dCas9 (Fig. 3C). To obtain reasonable kinetic profiles, the concentration of RNP for Δ3CE and Δ4CE was increased to 1000 nM, but even under these conditions both variants lag WT dCas9 at 300 nM. The PAM-interrogation ability of the two constructs appeared to be intact, as evidenced by the sharp drop-off in signal during the dissociation phase, but both Δ3CE and Δ4CE dissociated at a much higher rate compared to dCas9. The k_{on} was restored upon

addition of a 3-bp bubble, suggesting that these minimal Cas9s possess the kinetic defect in dsDNA binding inherent to both ΔREC2 and ΔREC3. The fact that these minimal constructs are still able to bind DNA in a sgRNA-targeted fashion is surprising, considering that the Δ3CE and Δ4CE constructs retain only ~72% and ~63%, respectively, of the original dCas9 protein primary sequence (Fig. 3B).

We assessed the DNA binding activity of the CE constructs in mammalian cells similarly to the single-deletion variants described earlier. As before, we performed CRISPRi against PCNA in U-251 cells, this time transducing the Δ3CE and Δ4CE KRAB fusions and sgRNA, followed by mixing with the parental cells and monitoring for mCherry fluorescence for up to 9 days. As expected from the minimal repression in bacteria, we did not observe functional depletion in the competition assay (Fig. S11H). We followed the fluorescence assay in mammalian cells with RT-qPCR 2- and 5-days post-transfection. Unlike the single-deletion variants, Δ3CE and Δ4CE do not repress nearly as well as dCas9, exhibiting a fold-change in PCNA expression relative to non-targeting sgRNA of 0.75±0.11 and 0.80±0.13, respectively, after five days post-transduction of the sgRNA (Fig. 3D). This result suggests that the Δ3CE and Δ4CE constructs are functional but severely defective in DNA binding in a mammalian system.



<u>Figure 2.3:</u> Stacking multiple domain deletions on Cas9 results in defective DNA-binding activity. A) *In vivo* transcription repression activity of MISER CRISPR effectors containing triple (Δ3CE) and quadruple (Δ4CE) deletion variants. Sublibraries of REC2, REC3, HNH, and RuvC were combined to build a library of

stacked deletions, and the resulting library was assayed for high-performing variants using FACS (light blue bars). As none of the variants contained a REC2 deletion (~Δ167-307), we named the highest-performing triple-deletion variant in this library (Library 2; see Fig S6) Δ3CE. To force a library containing REC2 deletions, a sublibrary of REC2 deletions was added to Δ3CE, resulting in a library of quadruple deletion variants that contain Δ3CE and a REC2 deletion (dark blue bars). Data are plotted as mean±SD from biological triplicates. **B)** Expression constructs for Δ3CE and Δ4CE, with specified deletions manually cloned in. **C)** BLI assay of CE constructs. Δ3CE and Δ4CE exhibit almost no binding against a fully-complementary dsDNA target at 300 nM RNP (see Fig. S10); and weak binding at 1000 nM RNP. Binding is rescued to near-WT levels when RNP concentration is 3.3x that of dCas9 if the dsDNA contains a 3-bp bubble in the PAM-proximal seed region. Data are normalized to 300 nM dCas9 binding to fully-complementary dsDNA. **D)** Measurement of CRISPRi efficacy of Δ3CE and Δ4CE in U-251 cells using RT-qPCR. Fold-change in PCNA expression levels is measured by RT-qPCR, 2- and 5-days after KRAB-Δ3CE and -Δ4CE expressing cell lines are transduced with a sgRNA targeting *PCNA*. Δ3CE and Δ4CE exhibit weak DNA binding and transcriptional repression activity compared to dCas9. Bars represent fold-change of PCNA expression relative to a non-targeting sgRNA. Error bars represent SD for at least 2 replicates.

2.3.4 The minimal Δ4CE construct has a similar structure as an ablated wild-type SpCas9

To understand the structural rearrangement accompanying domain deletion, we used single-particle cryo-EM to determine a reconstruction of the Δ4CE DNA-bound holocomplex (Fig. 4), to a resolution of 6.2 Å (Fig. S12). Remarkably, overlaying the density of the Δ4CE construct over the WT SpCas9 R-loop structure (PDB ID 5Y36)[46] as a rigid-body model shows that the minimal complex, consisting primarily of the REC1, RuvC, and C-terminal domains, possesses the same overall architecture as the WT holocomplex (Fig. 4A, S12). The double-helical dsDNA target and the stem-loop of the gRNA that are part of the R-loop can be resolved from the density and overlays almost exactly over the WT SpCas9 R-loop. This observation supports the hypothesis that the R-loop is a thermodynamically stable structure that drives the formation of the primed Cas9 RNP-DNA complex[51,52]. Although individual residues cannot be resolved, the remaining RuvC domain in the construct is linked to the C-terminus of the REC1 domain via a TS linker (MISER scar), thereby maintaining a bi-lobed complex reminiscent of WT SpCas9. The gRNA-interacting regions of the REC1 and CTD are also spatially conserved, consistent with their observed indispensability on the MISER enrichment map. This raises the question of how the minimal protein is able to form a stable R-loop despite lacking a large part of the REC lobe.

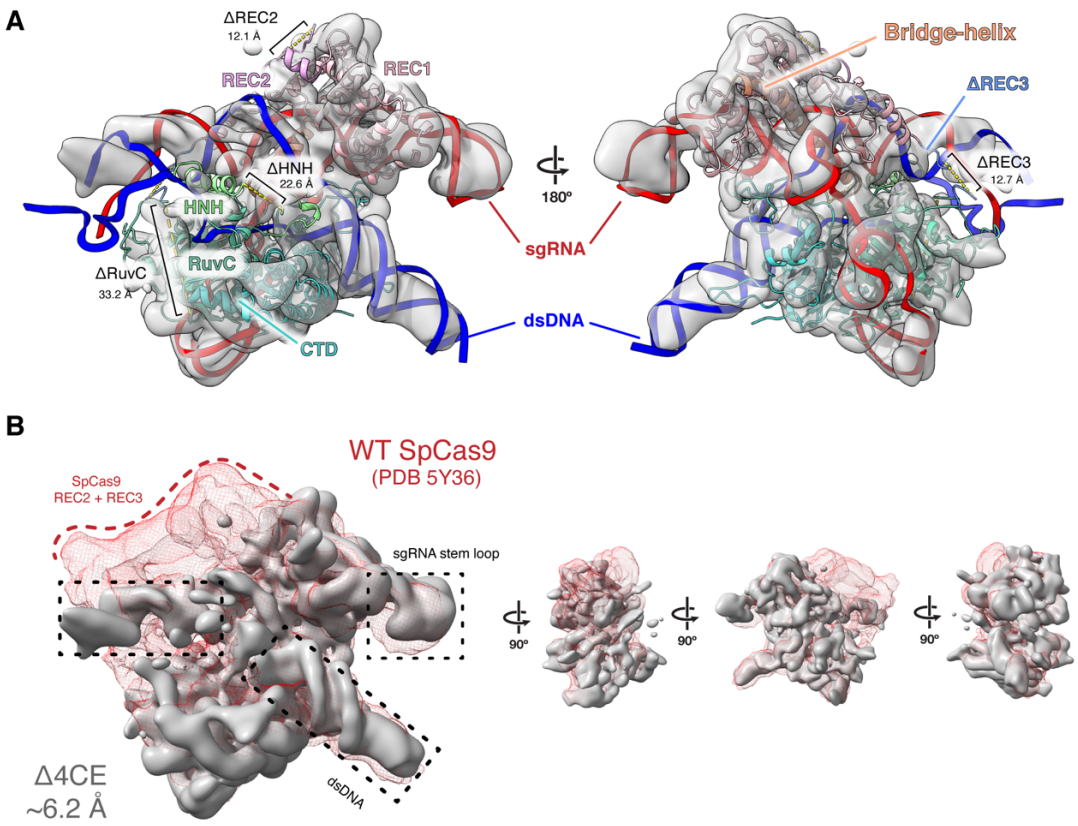


Figure 2.4: Density map of Δ4CE compared to WT SpCas9. A) Single-particle cryo-electron microscopy was used to obtain a density map of the dsDNA-bound RNP complex of the Δ4CE construct at an overall resolution of 6.2 Å (EMD-22518). Light grey volume shows the Δ4CE density overlaid onto RNA-DNA hybrid R-loop (red and blue) and structure of WT SpCas9 (PDB 5Y36). Cartoon model corresponds to the WT SpCas9 structure, showing only the remaining residues and corresponding domains after the REC2, REC3, HNH, and RuvC deletions from the Δ4CE construct are manually removed from the model. Deletion termini are labeled with the distances between termini. **B)** Density of Δ4CE cryo-EM overlaid with the WT SpCas9 clearly shows volumes representing dsDNA target and the sgRNA stem-loop (black boxes). The red mesh represents the total WT SpCas9 density from EMD-8236.

2.4 Discussion

Protein evolution takes large steps through sequence space using domain rearrangements, duplications, and indels[2,53]. While rearrangements, duplications, and insertions have been widely studied, domain deletions are largely under-investigated, due to limited experimental data and the difficulty in properly annotating deletions in protein sequence datasets[54]. Although deletion studies in proteins have been performed, they are limited in their scope in regard to the scale of deletions, complexity, and

generalizability. In this work, we present a technique that is versatile, comprehensive, and unbiased to probe the deletion landscape of virtually any protein, limited only by the fidelity and efficiency of a functional screen.

We have used SpCas9 as proof-of-concept to demonstrate the utility of MISER because it is a well-characterized, multi-domain protein, easy to assay, and its overall size poses a limit for therapeutic delivery. The wild-type SpCas9 gene is too large to be packaged into an adeno-associated viral vector (AAV), which has a maximum reported cargo size of <5 kb[55,56] when including the sgRNA sequence and necessary promoters. There are now smaller characterized CRISPR-Cas effectors suitable for AAV delivery by themselves[19,57], but an important need in both research and therapy is delivery of effectors fused to other domains, such as for base-editing and transcriptional activation or repression[58]. MISER may thus find utility in minimizing these much larger constructs. Additionally, immunogenicity is emerging as a major issue when developing SpCas9 as a therapeutic and deleting antigenic surface residues can potentially reduce the reactivity of the protein against the immune system[59,60].

We were surprised to discover the effect the deletion of the REC2 domain had on SpCas9 binding. Nishimasu et al. had previously reported that a REC2 deletion (Δ175-307) retained ~50% of editing activity and suggested that the attenuated activity might be due to poor expression or stability[35]. In contrast, our data suggest that the ΔREC2 variant folds and retains target recognition and binding function but loses DNA unwinding capability. The observation that ΔREC2 binding is restored upon addition of a 3-bp bubble adjacent to the PAM suggests that the poor binding is due to a kinetic defect. The specific nature of the defect requires further study, although we speculate that the REC2 domain interacts nonspecifically and transiently with the R-loop, perhaps stabilizing the DNA strands during hybridization (i.e., lowering the kinetic barrier) or stabilizing the final R-loop complex (i.e., lowering the energetic cost of unwinding and hybridization)[44].

We also note the observed difference in activity of the MISER constructs between bacterial in vivo repression experiments and the in vitro binding activity using BLI. We speculate that the MISER constructs are inherently defective for binding target DNA, but that sufficiently perturbed dsDNA in bacteria—such as during replication, transcription, or other rearrangements—presents enough opportunity in the form of dynamically un- and under-wound dsDNA, or stretches of single-stranded DNA, to allow the gRNA to anneal to the spacer sequence[52,61]. Additionally, abundant or overexpressed proteins, as is the case here, can often achieve concentrations exceeding 1 μM inside E. coli cells, so it is also possible that the overall high abundance of the MISER constructs in the bacterial repression assay is contributing to the binding signal[62].

The effect of the ΔHNH and ΔRuvC deletions was as expected in the bacterial repression assay; however, we were surprised to see that in the in vitro experiments the binding defect was not fully rescued upon addition of the 3-bp bubble in the dsDNA substrate. This suggests that while the REC domains might be conferring some kind of kinetically driven unwinding function to Cas9, the HNH and RuvC nuclease domains might instead have some role in stabilizing the overall DNA-bound complex. The

difference in the in vivo and in vitro conditions may be due to DNA dynamics inside a cell versus in solution.

Finally, in our cryo-EM structure of Δ4CE, we note the remarkable similarity of the protein to WT SpCas9, which underscores the inherent stability of the Cas9 R-loop complex. Previous studies have shown that formation and maintenance of the R-loop is the molecular "glue" that holds the DNA-RNA-protein complex together[51]. The similitude between the WT and Δ4CE structure also hints at the evolutionary history of SpCas9, suggesting that the "essential" function of the protein was to enable the formation of an R-loop upon a RuvC scaffold for DNA binding and cleavage, which was then tuned by accretion and interactions of other domains—such as those that comprise the REC lobe and the HNH domains[9,63]. Notably, this analysis ignores the role of the gRNA; future iterations of MISER could also be used to evaluate the deletion landscape of CRISPR-associated RNAs.

MISER facilitates the study of protein deletions with unprecedented versatility and efficiency. In this study we have explored domain modularity and essentiality of CRISPR-Cas9 domains, but MISER can be adapted to any application requiring a reduction in genetic size. AAV-based transgene delivery is subject to a <5 kb payload limit and is a prime target for MISER. Besides CRISPR proteins and their cognate gRNAs, there are numerous other therapeutic proteins limited by their size, such as CFTR (cystic fibrosis) and dystrophin (muscular dystrophy)[55,64]. Beyond threshold effects, even partially reducing the size of AAV genomes can provide a large advantage in packaging efficiency by improving capsid formation[55]. Finally, MISER also reveals small deletions tolerated within proteins, which suggests that this approach could be useful in the development of non-immunogenic biomolecules. Paring away antigenic residues may remove antigenic epitopes on a protein surface, thus allowing the molecule to function without eliciting an immune response[65,66].

2.4 Experimental Design

2.4.1 Molecular Biology

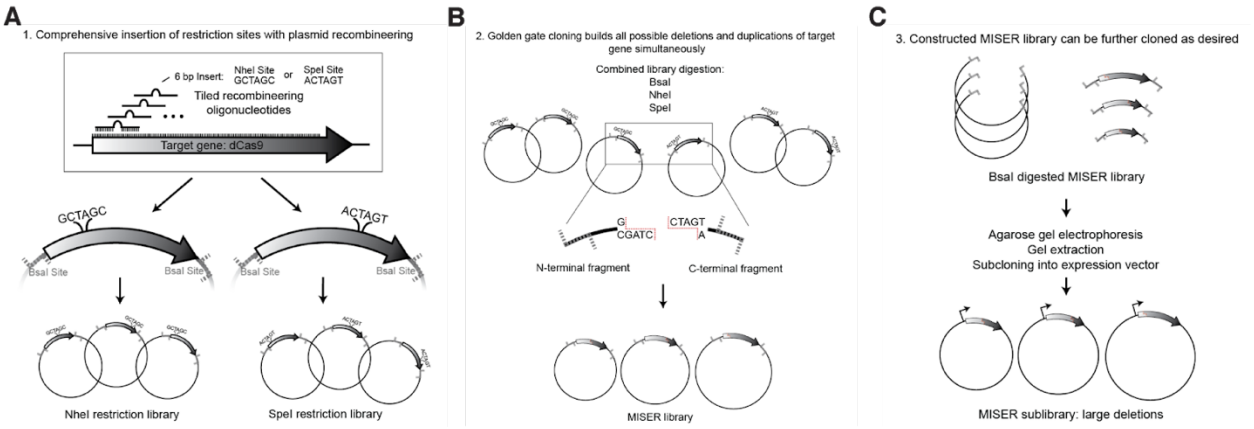
All restriction enzymes were ordered from New England Biolabs (NEB). Polymerase Chain Reaction (PCR) was performed using Q5 High-Fidelity DNA Polymerase from NEB. Ligation was performed using T4 DNA Ligase from NEB. Agarose gel extraction was performed using the Zymoclean Gel DNA Recovery kit, and PCR clean-up was performed using the 'DNA Clean & Concentrator', both from Zymo Research. Plasmids were isolated using the QIAprep Spin Miniprep Kit (Qiagen). All DNA-modifying procedures were performed according to the manufacturers' instructions.

2.4.2 MISER library construction: plasmid recombineering

Two sets of 1368 oligonucleotides were designed and ordered as Oligonucleotide Library Synthesis (OLS) from Agilent Technologies (Table S1). Oligonucleotides were designed to insert a six base-pair (bp) recognition sequence for either the restriction enzyme NheI or SpeI between every codon in dCas9 (Figure S1A). The full list of ordered oligonucleotides is available as Auxiliary Supplementary Materials - Recombineering Oligonucleotides. Internal priming sites were included in order to amplify NheI or SpeI specific oligonucleotide libraries. A modified amplification procedure was performed. In a 50 µL PCR reaction, 10 ng of template oligonucleotide library was amplified according to manufacturer's instructions, but with an extension time of only five seconds, and a total of only 15 cycles. 1.5% dimethyl sulfoxide (DMSO) was also included in the PCR reaction. These modifications were empirically determined in order to minimize undesirable higher order PCR products that were observed to be produced by amplification. These side products are likely the result of complementary oligonucleotides priming one another. Notably this phenomenon is likely inherent to amplification of a library of DNA tiled across a common sequence--in this case dCas9. PCR primers can be found in Table S6 and Auxiliary Supplementary Materials – Primer Sequences. 24 such reactions were typically performed in parallel and then combined, followed by concentration with Zymo DNA Clean & Concentrator. BsmbI restriction digestion was then used to remove priming ends, followed by a second concentration with Zymo DNA Clean & Concentrator, resulting in mature double-stranded recombineering-competent DNA.

Plasmid recombineering was performed as described in Higgins et al. 2017, using strain EcNR2 (Addgene ID: 26931) to generate MISER libraries in plasmid pSAH060. Plasmid sequences can be found in Auxiliary Supplementary Materials – Plasmid Sequences. Briefly, mature double-stranded recombineering-competent DNA at a final volume of 50 µL of 1 µM, plus 10 ng of pSAH060, was electroporated into 1 mL of induced and washed EcNR2 using a 1 mm electroporation cuvette (BioRad GenePulser). A Harvard Apparatus ECM 630 Electroporation System was used with settings 1800 kV, 200 Ω, 25 µF. Three replicate electroporations were performed, then individually allowed to recover at 30° C for 1 hr in 1 mL of SOC (Teknova) without antibiotic. LB (Teknova) and kanamycin (Fisher) at 60 µg/mL was then added to 6 mL final volume and grown overnight. A sample of recovered culture was diluted and plated on kanamycin to estimate the total number of transformants, typically $>10^7$. Cultures were miniprepped and combined the next day. Plasmid recombineering is relatively inefficient, and only a fraction of recovered plasmids contained successful NheI or SpeI insertions. In order to recover completely penetrant libraries, an intermediate cloning step was performed. A PCR product conferring resistance to chloramphenicol was cloned into both libraries of pSAH060 plasmids (Auxiliary Supplementary Materials - Chloramphenicol Selection). This PCR product contained either flanking NheI restriction sites or SpeI restriction sites, such that only modified pSAH060 plasmids (possessing NheI or SpeI restriction sites) could obtain chloramphenicol resistance through NheI/SpeI digestion and subsequent ligation. Libraries were then purified (Zymo) and transformed into XLI-Blue competent

cells for overnight selection in chloramphenicol (Amresco) at 25 µg/mL, followed by plasmid isolation the next day. Samples of recovered cultures were also plated on both kanamycin alone (native pSAH060 resistance) and chloramphenicol alone (resistance mediated by successful recombineering insertion) to estimate the fraction of modified plasmids and therefore the restriction library size. Recombineering efficiencies were observed at ~0.5% by this method, indicating restriction library sizes of ~50,000, well above the number of unique insertion sites per library (1,368). Finally, chloramphenicol resistant pSAH060 libraries were digested with either NheI or SpeI as appropriate, removing the chloramphenicol cassette. The libraries were run on an agarose gel, and the 5953 bp (5947 bp pSAH060 + 6 bp inserted restriction site) linear band corresponding to each library was gel extracted. To construct deletion variants composed of N- and C- terminal dCas9 fragments, one µg of each library was mixed and digested with BsaI, then cleaned up (Zymo). The resulting DNA mixture contained equimolar free dCas9 N- and C-terminal fragments, as well as equimolar pSAH060 vector backbone. This mixture was then ligated in the presence of SpeI and NheI, 'locking' dCas9 fragments together by one of two six bp scar sites not recognized by either enzyme (Figure S1B). The ligated MISER library was transformed into XL1-Blue, grown overnight and plasmids were isolated the next day. The MISER library of dCas9 is quite large, with 936,396 possible deletions (N(N + 1) / 2, N = 1368), and all cloning steps were performed with validation that $>10^7$ transformants were obtained.



<u>Figure 2.5:</u> Full cloning scheme for Minimization by Iterative Size-Exclusion and Recombination (MISER). The method can be considered in three parts. **A)** Plasmid recombineering generates two comprehensive libraries of restriction site insertions across the target gene. These restriction sites are both novel to the target plasmid and produce compatible sticky ends. Recombineering was performed similarly as in (Higgins 2017), where the target gene lacks a promoter and start codon to prevent growth biases during library construction and is flanked by BsaI sites for later Golden Gate cloning (here, plasmid pSAH060). Additionally, rather than mutagenic oligos, double stranded PCR product was used for recombineering, and another cloning step was introduced to remove unmodified plasmids. These modifications are described in Experimental Design. **B)** Modified golden gate cloning generates a library of ligated N- and C- terminal fragments of the target gene, comprehensively producing protein deletion variants as well as duplication variants. An equimolar mixture of the two plasmid libraries is mixed and fully digested to produce free N- and C- terminal fragments of the target gene. This fragment mixture is then re- ligated in the presence of NheI and SpeI. Successful ligation of an N- and C-terminal fragment

from differing libraries produces one of two possible 6 base-pair scar sequences. These novel scar sequences are not recognized by either NheI or SpeI, thus trapping the desired chimeric product as a final ligated vector. Because N- and C-terminal fragments are ligated randomly, these chimeric products produce both protein deletions and protein duplications. Ideally the library is both large enough and minimally biased to produce a large fraction of possible variants. The product of this step can be considered a MISER library of plasmid pSAH060. **C)** A final cloning step moves the MISER library into a desired context – i.e. an expression plasmid, here pSAH063. Step C also allows for size-based exclusion of undesired protein variants by extraction from an agarose gel (Figure 1 and Figure S2).

	SpeI Insertion	**NheI Insertion**
Recombineering Oligo: Insertion Site 1	AACACGTCCGTCCTAGAACTcgtctcatac gcaaAccgcctctccccgcgcgttggcggt ctcaatctATGactagtgataagaaatact caataggcttagctatcggcacaaatagcg tcgggagacgGCAAGCGGTACACTCAGATC AGTGTTGAGCGTAACCAAGT	AACACGTCCGTCCTAGAACTcgtctcatacgcaa Accgcctctccccgcgcgttggcggtctcaatct ATGgctagcgataagaaatactcaataggcttag ctatcggcacaaatagcgtcgggagacgGCAA C GGTACACTCAGATCAGTGTTGAGCGTAACCAAGT

Table 2.1: Example Oligo Library Synthesis (OLS) oligonucleotides used in this study. The full list of ordered oligonucleotides is available as 'Auxiliary Supplementary Materials - Recombineering Oligonucleotides'. All oligonucleotides were ordered from Agilent Technologies, Inc. Oligos were designed to incorporate 45 and 47 bp of homology upstream or downstream of the insertion site, respectively (lowercase). Six bp were inserted between dCas9 codons, beginning after the target codon. The above example targets the start codon, 'ATG' (bold uppercase). These six bp consisted of recognition sequences for either the restriction enzyme SpeI or NheI (underlined). Flanking primer sequences allowed the amplification of the entire OLS library (italics) using oligonucleotides SAH_284 and SAH_285 (Table S6). Specific libraries of SpeI recombineering oligonucleotides or NheI recombineering oligonucleotides were amplified using forward primer SAH_284 and either SAH_286 or SAH_287 reverse primers, respectively. After amplification, these dsDNA products can be 'matured' by cleavage with the restriction enzyme BsmbI (bold lowercase), which cleaves internally of its recognition site, thus removing all non-homologous priming sequence from the recombineering template.

2.4.3 MISER library construction: library size selection

The MISER library is theoretically composed of all possible N- and C-terminal fragments, including both duplications and deletions. To isolate deletions in a particular size range, the MISER library was digested with BsaI, in order to excise the dCas9 gene from the vector backbone and run on an agarose gel. Various slices of the MISER library were individually gel extracted (Fig. S2A), ligated into expression vector pSAH063 (Fig. S2B), and transformed into *E. coli*.

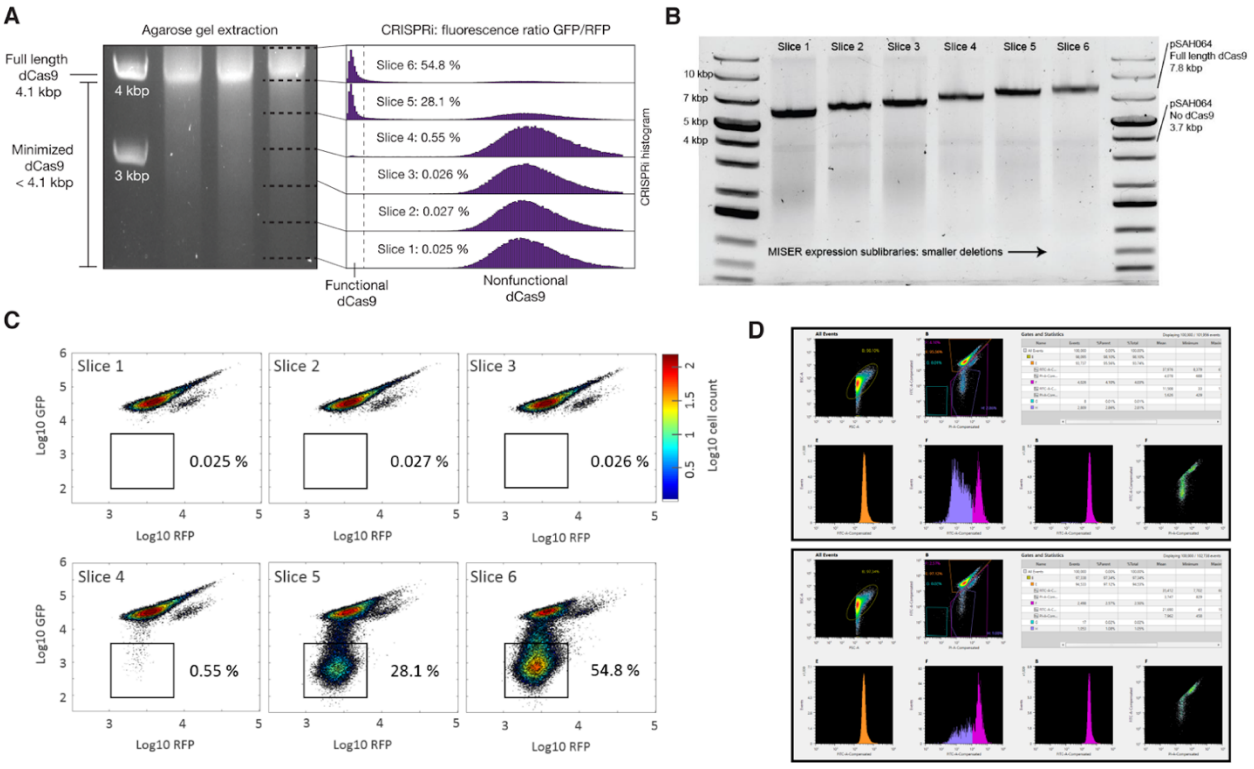


Figure 2.6: Size exclusion and flow cytometry identify the range of dCas9 deletion sizes exhibiting *in vivo* transcriptional repression. A) To empirically determine the size range of functional deletions, an agarose gel of the dCas9 MISER deletion library was sliced into six sub-libraries, independently cloned into expression vectors (B), and assayed for CRISPRi GFP repression via flow cytometry (C). Sublibrary Slice 4 was the most stringent library with detectable repression, with functional variants becoming more frequent in slices composed of smaller deletions as expected. B) The six gel slices in (A) were individually gel extracted and ligated into expression vector pSAH063, generating pSAH064 plasmids with dCas9 deletions. The resulting expression sub-libraries exhibit high precision in size ranges when assayed by agarose gel electrophoresis. C) Flow cytometry identifies Slice 4, 5, and 6 as expression sub-libraries containing functional dCas9 deletion variants. GFP repression CRISPRi was performed as described in Experimental Design. The region of phenotype defined as 'functional' is illustrated. The percent of functional hits is annotated. D) Screenshots from Sony Cell Sorter Software exemplifying the gating strategy, with upper panel showing full library sort and lower panel showing Slice 4. Gate H was used to sort cells containing repression-competent CE variants.

2.4.4 Fluorescence repression assays and flow cytometry

The catalytically dead dCas9 MISER variants were used to repress the transcription of genomically encoded fluorescent reporter genes in *E. coli* as previously described [1]. A sgRNA targeting Green Fluorescent Protein (GFP) was transcribed from plasmid pgRNA-bacteria (Addgene ID 44251) [1], which results in repression of constitutively expressed GFP, contingent on functional dCas9 expression from pSAH063 [2]. This repression was quantified relative to non-targeted Red Fluorescent Protein (RFP), which is expressed from the same genomic locus [1]. This assay yields robust repression detection (Fig. S2B), with at least an order of magnitude lower GFP signal after 8 hours of growth at 37° C with 750 rpm shaking in LB media + 1 nM Isopropyl β-D-1-thiogalactopyranoside (IPTG) induction of dCas9 from pSAH063. Assays and flow cytometry were conducted in either an M1000 plate reader (Tecan) or an SH800 Cell Sorter (Sony Biotechnology). For GFP/RFP ratiometric measurements (Fig. 2A, 3A) there was no significant difference between samples for the RFP fluorescence measurement.

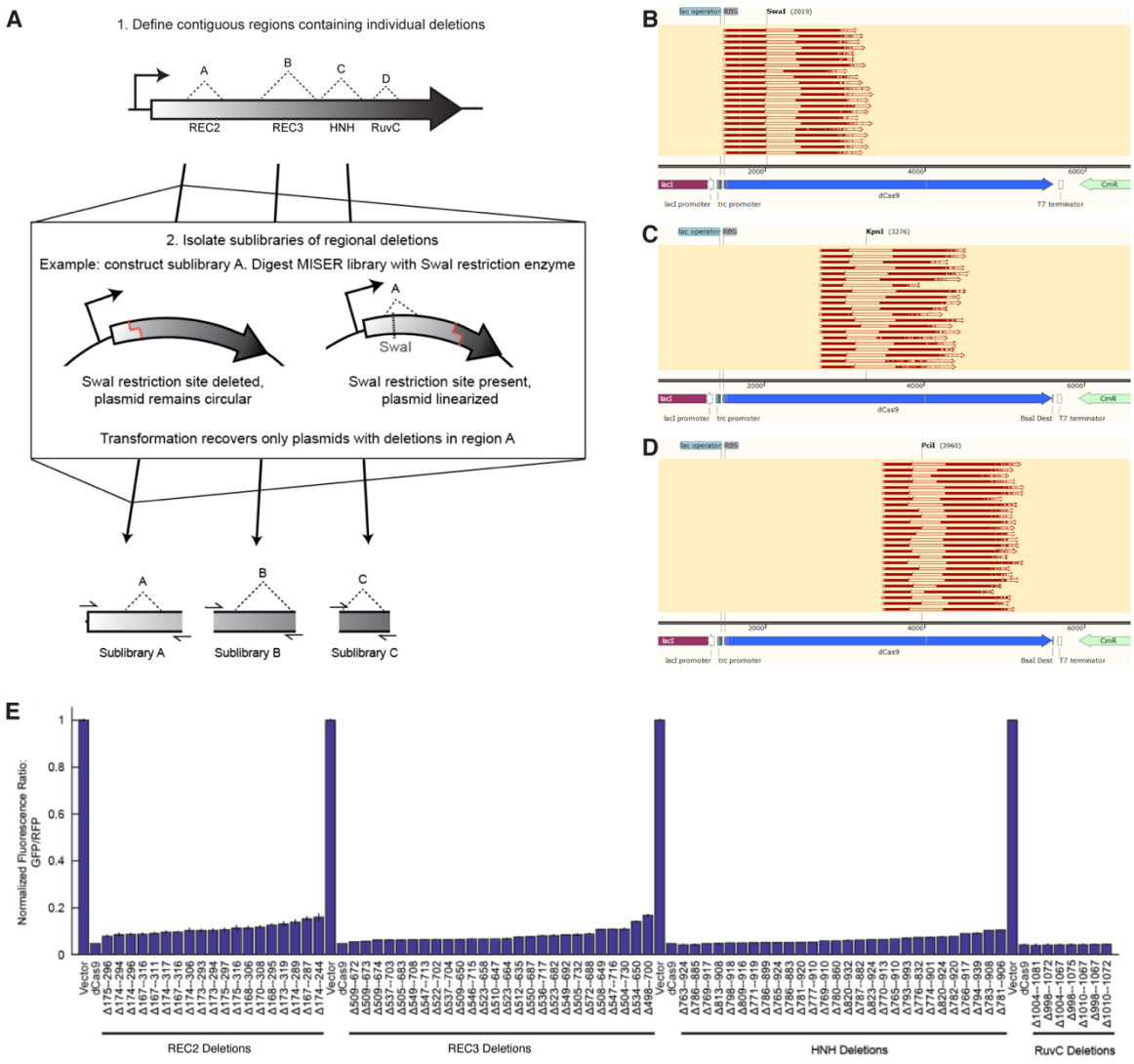


Figure 2.7: MISER sublibraries composed of specific deletions can be generated by restriction digestion. A) Digesting a MISER library with a restriction enzyme that has exactly one site within the plasmid will linearize the majority of plasmids, while plasmids with the site deleted will remain circular. This reaction can then be transformed in order to recover a sublibrary containing deletions from a specific region. **B)** For example, the restriction enzyme SwaI was used to isolate deletions in the REC2 region. The enzyme recognition site is shown mapped to the sequence of pSAH064, the dCas9 expression plasmid, illustrating the overlap with various sequenced deletions. **C)** The restriction enzyme KpnI was used to isolate deletions in the REC3 region, as in B. **D)** The restriction enzyme PciI was used to isolate deletions in the HNH region, as in B. **E**) Sublibraries containing regional individual deletion variants were re-transformed, and colonies were picked and assayed for CRISPRi activity. A subset of the most active clones was Sanger sequenced to identify the precise deletion. RuvC deletions could not be isolated by the sublibrary approach, and instead were cloned manually by PCR. Data are plotted as mean±SD from biological triplicates.

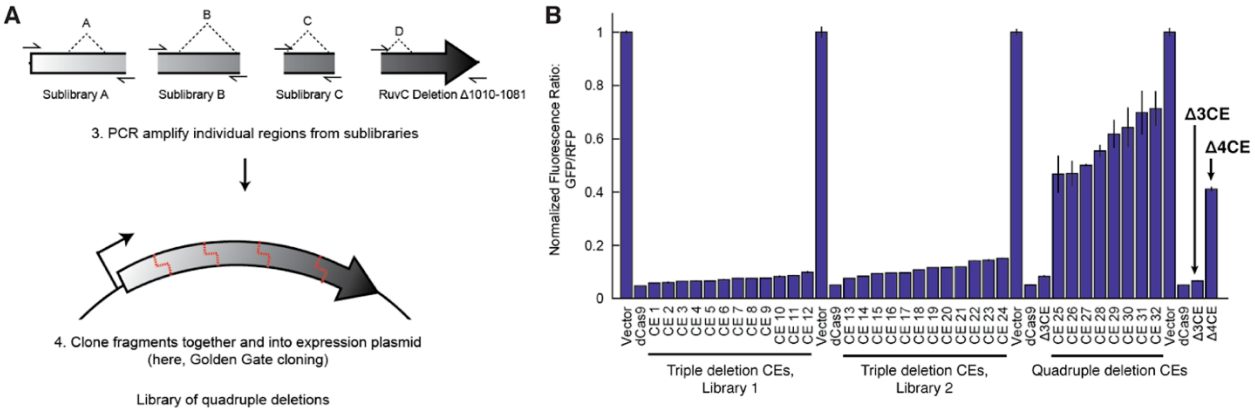


Figure 2.8: Golden Gate Cloning builds libraries of CRISPR Effector (CE) variants with multiple deletions. A) One highly functional RuvC deletion variant from Region D was PCR amplified, along with Sublibraries A, B, and C. PCR primers added Golden Gate compatible sticky ends, enabling Golden Gate cloning of individual fragments to form a library of CE deletion variants, Library 1. **B)** Flow cytometry was performed to isolate the most functional CE variants from the "stacked" library described in (A). All highly functional CE variants from Library 1 were found to lack REC2 deletions (sequences of CE variants selected for display on this plot can be found in Table S3). To verify this result, a second version of Sublibrary A was created, using a different strategy to isolate REC2 deletions as follows: the full MISER library was digested with the restriction enzyme BlpI, which cuts at amino acids 227-228 (instead of SwaI), and the resulting DNA was used directly as template for the PCR reaction (BlpI cuts pSAH064 three times and thus cannot be directly re-transformed to isolate the sublibrary). Library 2 thus contains all four deletion variants as in Library 1, except the sublibrary of REC2 deletions was entirely remade. However, once again functional CE variants isolated by FACS lacked REC2 deletions. The most functional variant in Library 2, CE 13, was named Δ3CE. Finally, to directly assay the effects of a REC2 deletion, the REC2 region of Δ3CE was replaced with a library of deletions from Sublibrary A. These quadruple deletion CE variants all exhibited vastly reduced CRISPRi activity compared to Δ3CE alone. The most functional variant assayed was named Δ4CE. Data are plotted as mean±SD from biological triplicates.

Deletion	Δ3CE v1	Δ3CE v2	Δ3CE v3	Δ3CE v4	Δ3CE v5	Δ3CE v6	Δ3CE v17	Δ3CE v21	Δ3CE v22	Δ3CE	Δ4CE
REC2	-	-	-	-	-	-	-	-	-	-	[180-297]
REC3	[511-716]	[498-699]	[500-688]	[497-700]	[501-664]	[512-721]	[509-650]	[508-649]	[508-646]	[503-708]	[503-708]
HNH	[813-909]	[813-908]	[811-898]	[786-882]	[804-893]	[809-916]	[776-923]	[768-900]	[786-923]	[792-897]	[792-897]
RuvC	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]	[1010-1081]

Table 2.2: Deletions present in selected MISER variants. Indicated numbers represent the first and last amino acid deleted from the protein.

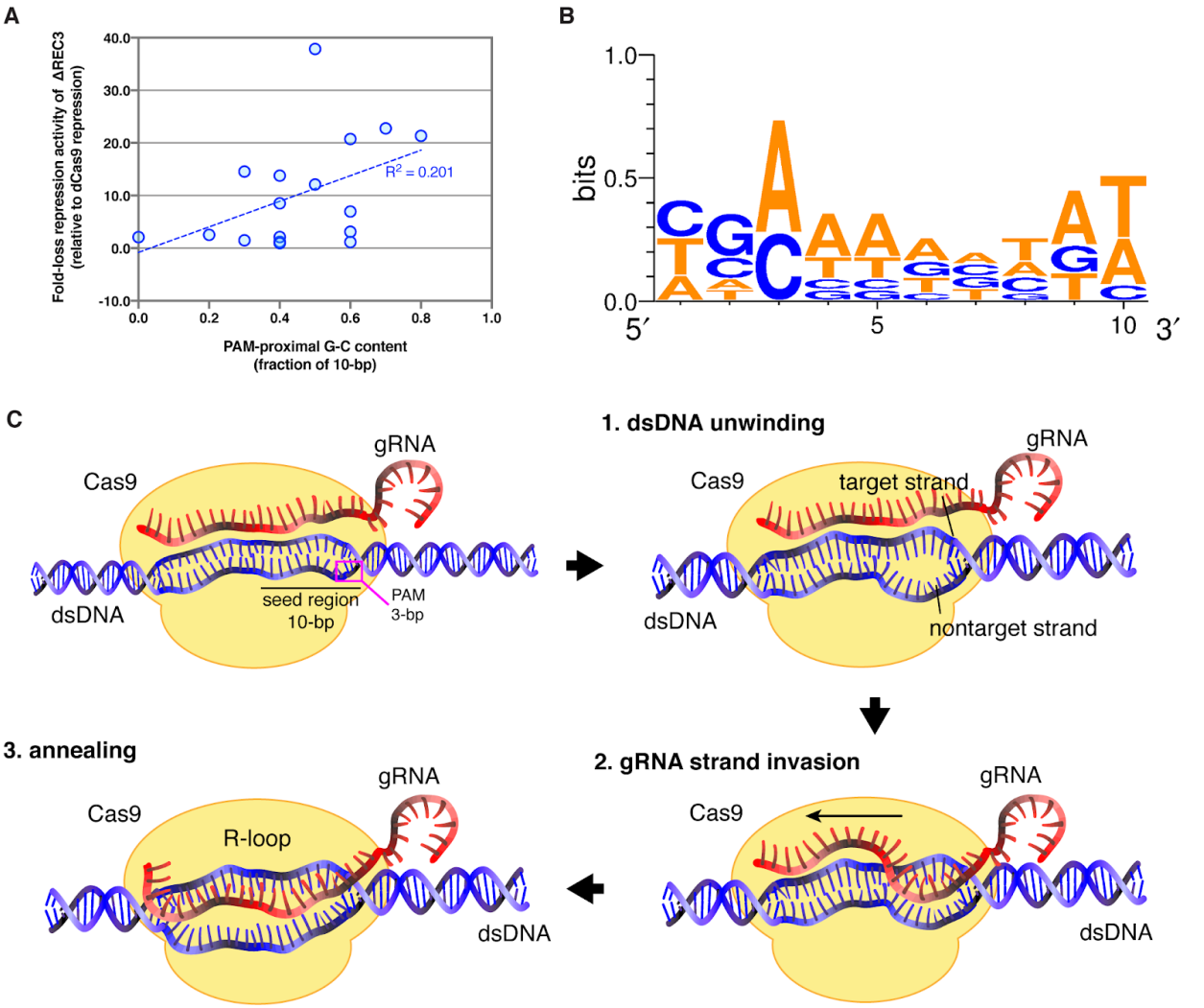


Figure 2.9: Spacer sequence-dependent variability in repression activity of ΔREC3. A) Plot showing fold-change in repression by ΔREC3 for different targets versus fraction of G-C content in seed region. Correlation between G-C content and repression is low and does not fully explain the variability in repression seen by the ΔREC3 construct across different target sequences. **B)** WebLogo showing spacer sequence variability for guides that exhibit at least a three-fold loss in repression by ΔREC3 compared to dCas9. **C)** Schematic showing the process of gRNA invasion into the dsDNA target leading to R-loop formation by Cas9. In Step 1, unwinding of the dsDNA double-helix is initiated at 1-2 bases adjacent to the PAM in the seed region, creating a destabilized region where the gRNA can invade, in Step 2. Hybridization of the gRNA to the target strand occurs in the seed region and proceeds in the PAM-distal direction (3'→5'), until the entire spacer sequence (~20bp) is annealed to the target strand, generating an RNA-DNA duplex called an R-loop (Step 3). RNA-DNA hybrid is shown as a 2-D representation for clarity instead of a helix.

Gene	Distance from RBS (bp)	PAM-proximal 10bp sequence (5'-3')	PAM-proximal G-C fraction	Fold loss	Std. dev.
GFP	**38**	**AACAAGAATT-NGG**	**0.2**	**2.54**	**0.23**
RFP	124	TTAGCGGTCT-NGG	0.5	37.84	3.78
GFP	**130**	**ATAAATTTAA-NGG**	**0.0**	**2.11**	**0.01**
GFP	**174**	**TGACAAGTGT-NGG**	**0.4**	**1.23**	**0.02**
GFP	**196**	**TGAACACCAT-NGG**	**0.4**	**2.14**	**0.10**
GFP	**225**	**TCATGTGATC-NGG**	**0.4**	**0.96**	**0.05**
GFP	262	CCTTCGGGCA-NGG	0.7	22.77	0.73
GFP	**316**	**CGCGTCTTGT-NGG**	**0.6**	**1.18**	**0.06**
GFP	**355**	**CGATTAACAA-NGG**	**0.3**	**1.50**	**0.06**
RFP	111	TACCTTCGTA-NGG	0.4	8.54	0.50
RFP	130	TTCAGTTTAG-NGG	0.3	14.56	0.77
RFP	**165**	**CCCAAGCGAA-NGG**	**0.6**	**3.13**	**0.06**
RFP	182	CTGCGGGGAC-NGG	0.8	21.35	0.71
RFP	197	GGAACCGTAC-NGG	0.6	6.98	0.23
RFP	208	ACGTAAGCTT-NGG	0.4	13.79	2.92
RFP	239	CAGGTAGTCC-NGG	0.6	20.74	4.25
RFP	248	GGACAGTTTC-NGG	0.5	12.10	0.60

Table 2.3: gRNA target loci and G-C content dependence of ΔREC3 repression. Spacer sequences highlighted in blue were used to generate the WebLogo in Figure S9A.

2.4.5 Deep sequencing

100 nucleotide single end reads were used to sequence the dCas9 Slice 4 and Slice 5 libraries. dCas9 open reading frames were amplified from pSAH064 libraries with primers SAH_356 and SAH_358. PCR products were further prepared for deep sequencing by the UC Berkeley Functional Genomics Laboratory. Sequencing was performed by the UC Berkeley Vincent J. Coates Genomics Sequencing Laboratory on an Illumina HiSeq4000. Samples were mixed at custom ratios as follows: Slice 5 Naïve Library – 10%; Slice 5 Sorted Library – 10%; Slice 4 Naïve Library – 40%; Slice 4 Sorted Library – 40%. Sequencing analysis was performed with custom MATLAB scripts available online at https://github.com/savagelab. Briefly, reads were analyzed for the novel presence of the two possible MISER scar sequences, 'GCTAGT' or 'ACTAGC'. The majority of reads were fully WT dCas9 sequences, as expected due to the fact that scar sequences can occur anywhere along dCas9. Once detected, reads containing 15 bp upstream and downstream of the scar (that exactly matched dCas9 sequence) were used to identify the location of a deletion. Sequencing statistics can be found in Table

S3. Enrichment ratios were calculated by taking the ratio of the frequency of each variant before and after selection [3]. To conservatively display variants only detected in one library, one artificial read was added to both datasets. The log base ten of these enrichment ratios were plotted (Figure S3 A and B) for each of the two libraries. For visualization, these two datasets were also normalized according to their Pearson Correlation (Figure S3 E), combined (the mean was calculated for those variants with two values), and rescaled for display (Figure 1C and S4 A). Variants with large deletions (>1000-bp) as shown in Figure S3 C and D are most likely "cheaters," i.e., small plasmids that are missing most of the dCas9 sequence and are therefore more easily replicated and less toxic to the cells.

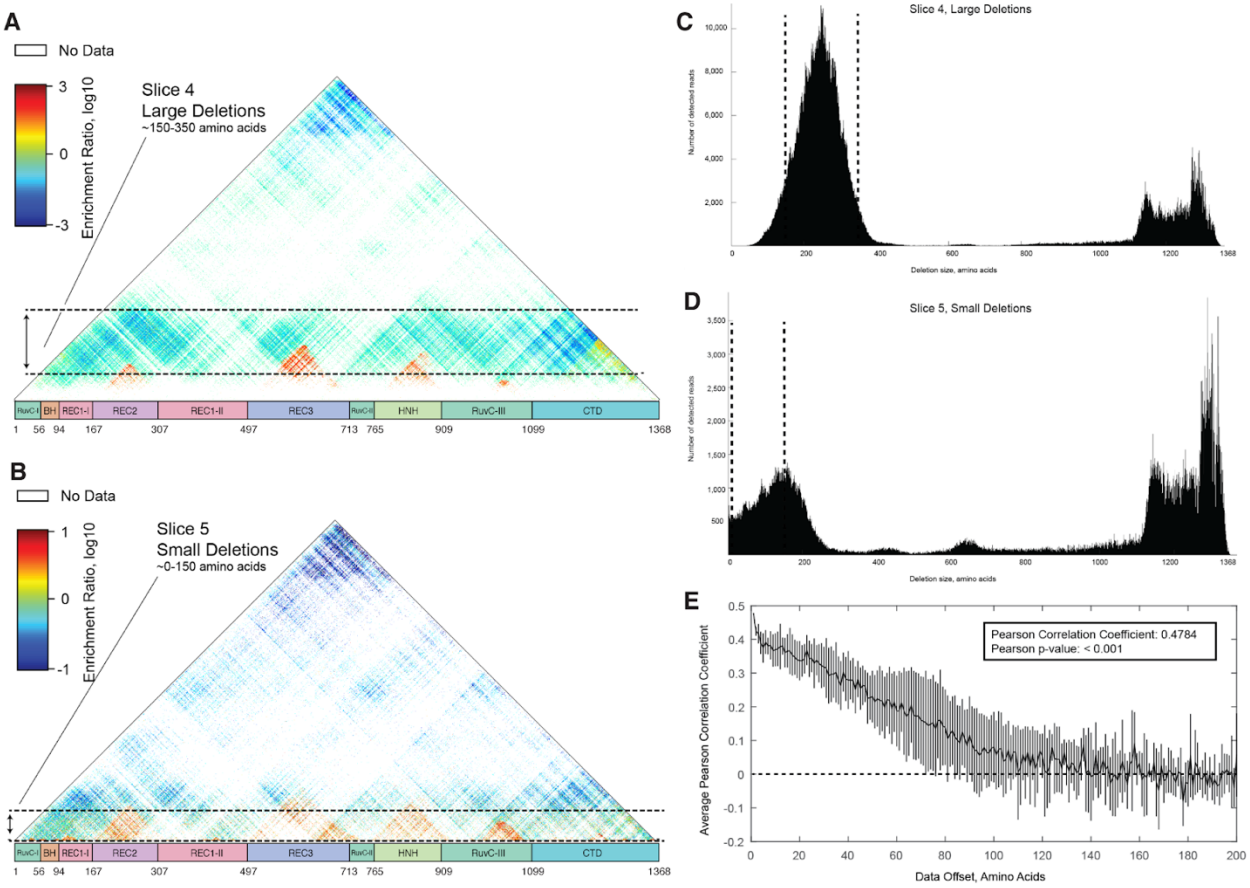


Figure 2.10: Deep sequencing of the sublibraries of Slice 4 and Slice 5 reveal deletion regions throughout dCas9. A) Raw enrichment map of Slice 4 sub-library. Each pixel represents a single deletion variant, whose start and end points are the axis intercepts when moving down and to the left or right, respectively, as described in the main text. Domain boundaries are labeled by amino acid number. The pixel color also denotes the degree of enrichment or loss following flow cytometry screening for transcriptional repression in vivo. Detailed calculations are described in the supplementary methods. Deletions corresponding to sizes within the gel slice are indicated by dashed lines. **B)** Raw enrichment map of Slice 5 sub-library, as in (A). Note the differing range of enrichment ratios. **C)** Histogram of deletion sizes in the naïve Slice 4 library. The hypothetical edges of the gel slice are indicated by dashed lines. **D)** Histogram of deletion sizes in the naïve Slice 5 library. The edges of the gel slice are indicated by dashed lines. **E)** Slices 4 and 5 independently replicate the same large functional deletion regions. The raw

enrichment maps of Slice 4 and Slice 5 contain many of the same variants, and the Pearson correlation for these variants is highly significant ($p < 0.001$). Furthermore, this correlation is progressively lost if the two enrichment maps are shifted relative to one another. The line plots the mean of four additional Pearson correlations where the data array has been offset – either up, down, left, or right – by the indicated number of amino acids. This analysis verifies that the two enrichment maps independently identify large-scale regions of dCas9 which can be deleted and validates the apparent visual correspondence between maps A and B. Error bars, standard deviation.

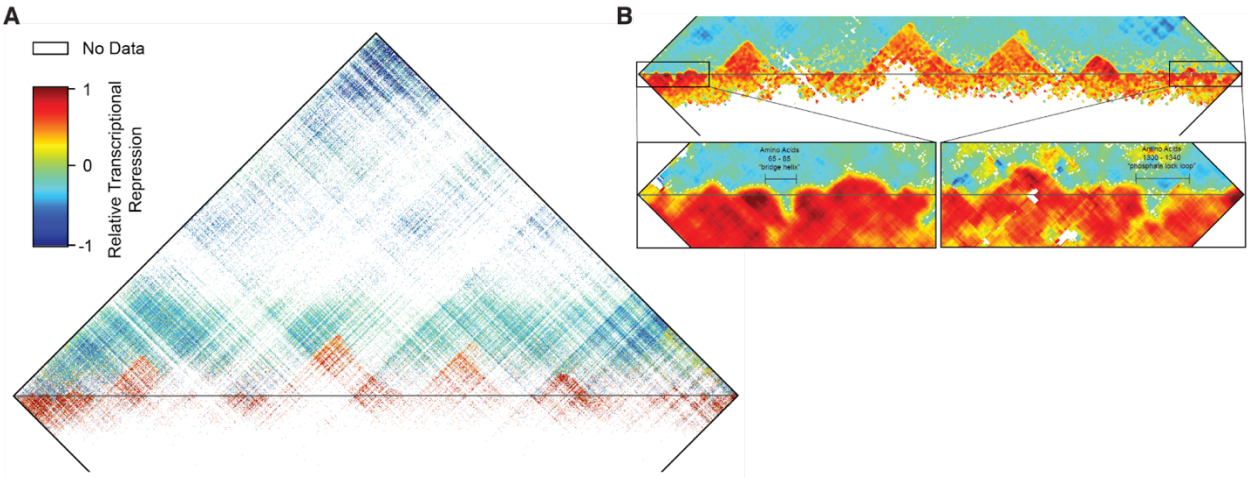


Figure 2.11: Key elements of dCas9 secondary structure are revealed by the functional impact of small deletions and insertions. A) The enrichment map of Figure 1C is presented in its entirety, including small duplications of dCas9 sequence. The horizontal grey line corresponds to the boundary between deletions (top) and tandem duplicate insertions (bottom). Note that in all cases a two amino acid MISER scar is also present (either Ala-Ser or Thr-Ser) which is not included in display or numbering. B) The combined enrichment map in (A) was interpolated to highlight the boundaries between functional and non- functional deletions, which are not clearly visible in the raw data. Pixels were replaced by the mean enrichment value of neighboring deletions/duplications, plus itself, in a square window 10 amino acids wide. Windows with fewer than five values were left white. Insets: The N- and C- terminal regions were particularly well resolved by this method, and elements of interest are annotated. The 'bridge helix' and 'phosphate lock loop' are two examples of secondary structure which strongly disallow small insertions.

	Total Reads	Deletions Sequenced	Unique Deletions	Enriched Unique Deletions	De-enriched Unique Deletions
Slice 4 Naïve	132,274,232	1,923,543	192,447		
Slice 4 Sorted	140,589,968	1,960,138	25,948	19,618	6,330
Slice 5 Naïve	37,873,068	590,859	111,438		
Slice 5 Sorted	35,016,326	290,947	51,462	31,794	19,668
Total	345,753,594	4,765,487	381,295	51,412	25,998

Table 2.4: Statistics for deep sequencing of MISER libraries Slice 4 and Slice 5.

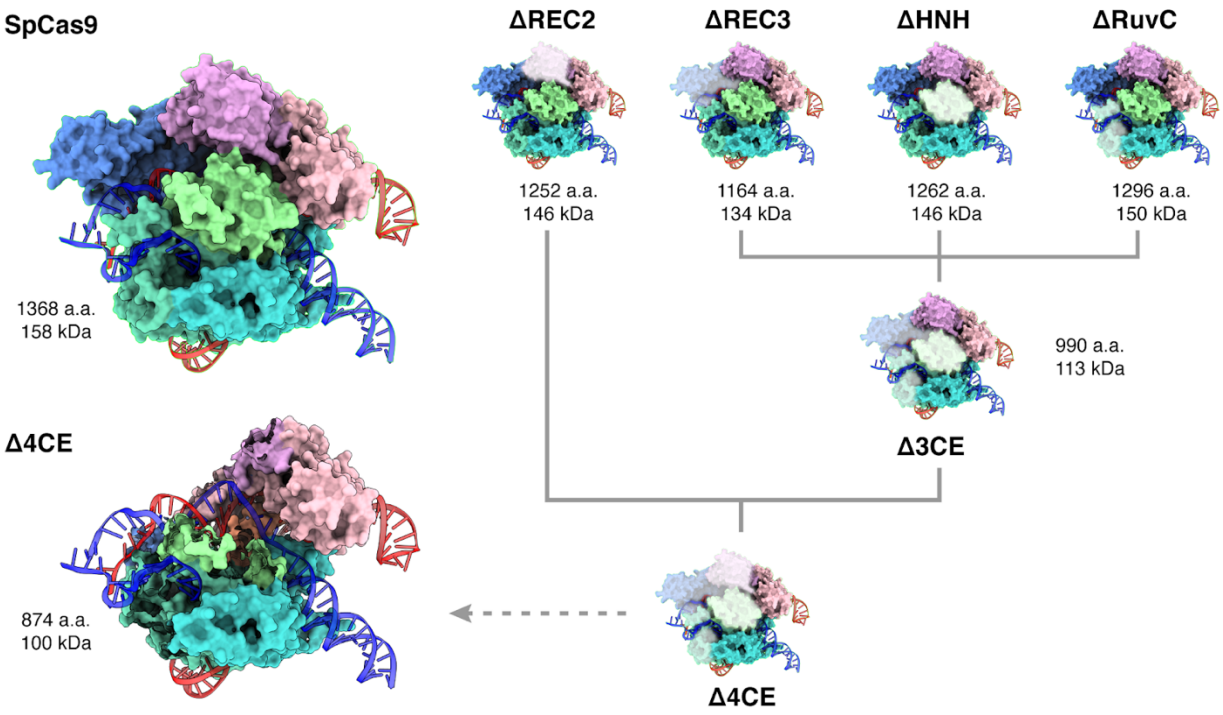
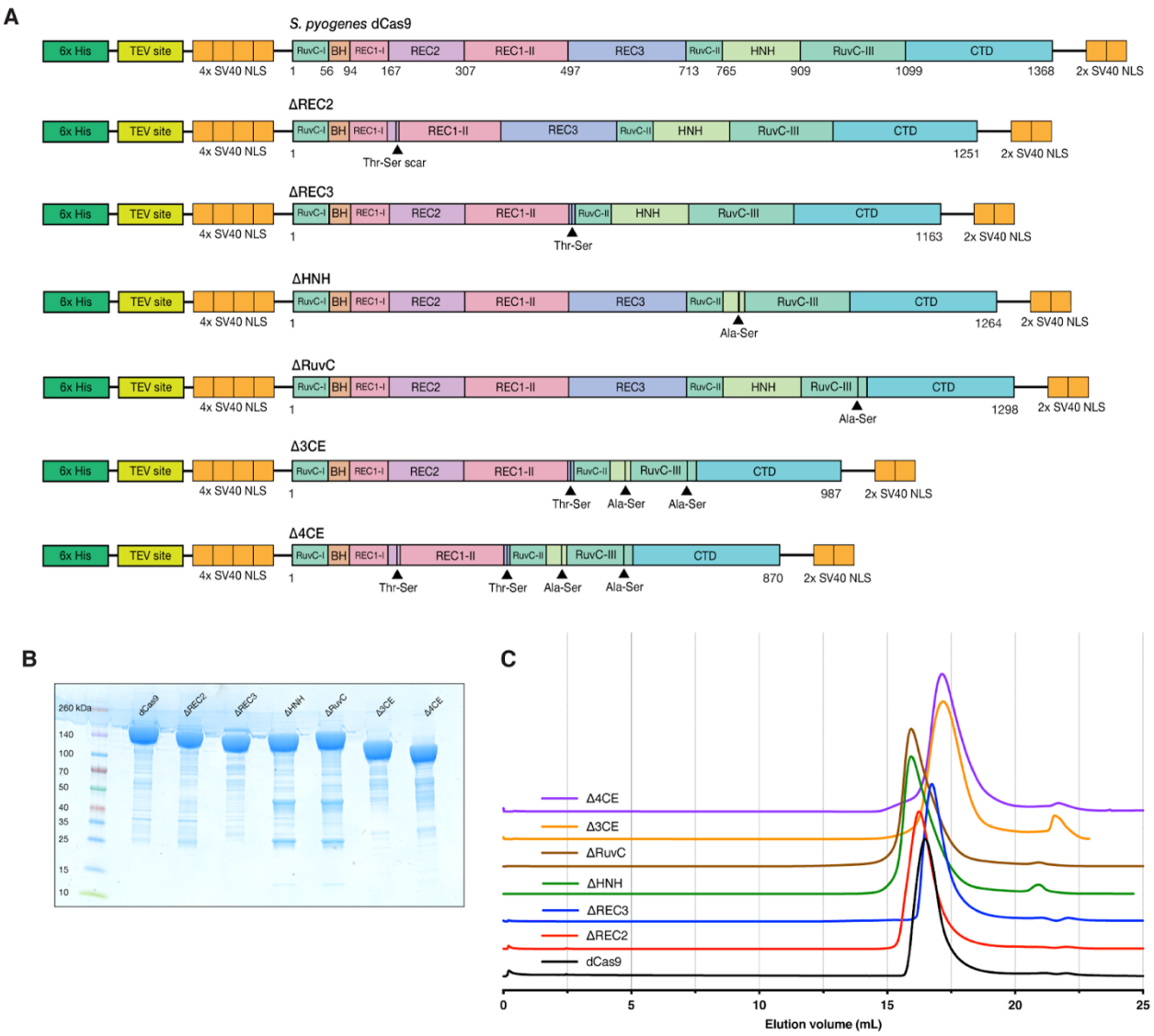


Figure 2.12: 3D comparison of complete dCas9-sgRNA-dsDNA complex and modeled MISER constructs. Model of SpCas9 complexed with sgRNA and dsDNA (PDB 5Y36), and MISER domain deletions overlaid. Δ3CE contains the REC3, HNH, and RuvC deletions, and Δ4CE contains the additional

REC2 deletion, as described in Fig. 2 and S5. The Δ4CE model is shown with the domains corresponding to MISER deletions hidden. Molecular weights are calculated by the ExPASy ProtParam tool (https://web.expasy.org/protparam/).

2.4.6 Protein expression and purification

A *Streptococcus pyogenes* Cas9 gene containing nuclease-deactivating mutations D10A/H840A (a.k.a. dCas9) was cloned into a pET14b expression vector, encoding a N-terminal 6xHis fusion tag and a C-terminal 2xNLS fusion tag. Specific MISER dCas9 variants were cloned by PCR-amplification (Q5 High-fidelity polymerase, NEB) of the dCas9 gene excluding deleted regions obtained from MISER screen (see Table S4 for primer sequences). Plasmids were verified by Sanger sequencing (UC Berkeley DNA Sequencing Facility). dCas9 and MISER constructs were overexpressed in *E. coli* BL21 (DE3) LOBSTR expression system (Kerafast). Cells were grown in Terrific Broth, modified media with 8 mM $MgCl_2$ and 0.5 glycerol and induced at ~0.6 OD with 0.5 mM IPTG. Cells were resuspended in Lysis Buffer (20 mM HEPES pH 7.5, 1 M KCl, 15 mM imidazole, 1 mM TCEP, 10% glycerol, 0.1 mM PMSF, Roche protease inhibitor tablet), lysed by sonication and clarified by centrifugation, and incubated with Ni-NTA resin to purify soluble fractions. Protein-bound Ni-NTA resin was washed with Wash Buffer (Lysis Buffer + 0.1% Triton X-114), and eluted (Elution Buffer: 20 mM HEPES pH 7.5, 150 mM KCl, 300 mM imidazole, 1 mM TCEP, 10% glycerol). Eluted fractions were subjected to a Heparin Sepharose column (GE Healthcare) for ion-exchange chromatography (300 mM KCl to 1 M gradient), concentrated, and further purified on a gel-filtration column (Superose 6 Increase, GE Healthcare). Protein Storage Buffer was as follows: 20 mM HEPES pH 7.5, 150 mM KCl, 1 mM TCEP, 10% glycerol. Purified protein aliquots were flash-frozen in liquid nitrogen and stored at -80°C. Concentrations were measured via Nanodrop A280 (ThermoFisher Scientific).



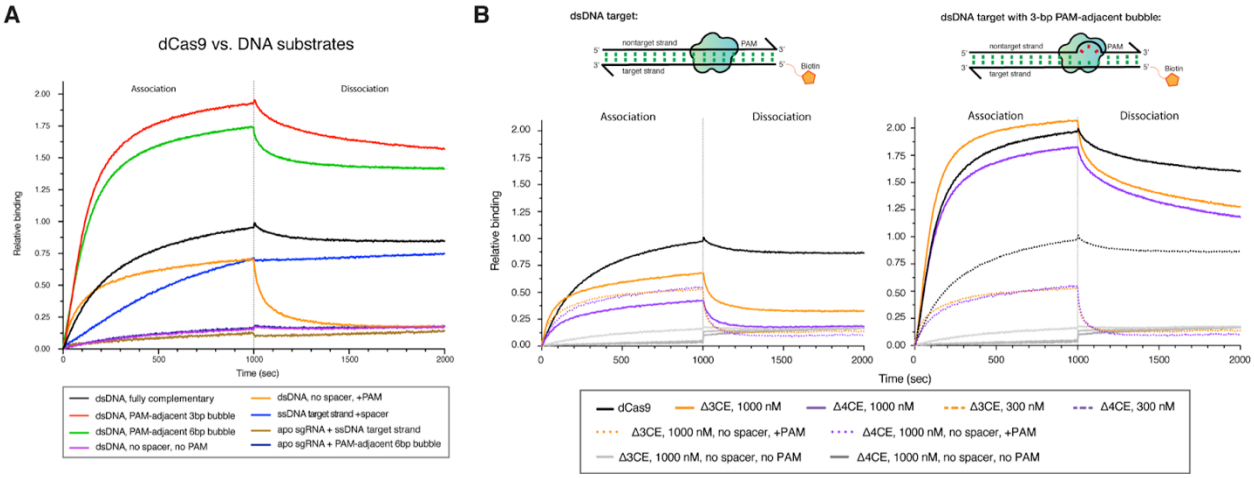
<u>Figure 2.13</u>: Expression constructs and protein purification of MISER constructs. A) Expression constructs for dCas9 containing MISER deletions and accompanying scars. All constructs were expressed using an IPTG-induced T7 promoter, and contain a N-terminal 6x His-tag, a TEV protease site, 4x SV40 NLS, and 2x SV40 NLS on the C-terminus. **B)** SDS-PAGE of purified MISER constructs. **C)** Size-exclusion chromatogram showing elution of all MISER constructs on a GE Superose 6 Increase column.

2.4.7 In vitro DNA binding assays

Purified proteins were complexed with 1.2x molar ratio sgRNA in the presence of 5 mM $MgCl_2$. 5'-biotinylated target DNA and corresponding non-target DNA was purchased from IDT as single-stranded oligos and annealed 1:1 according to standard IDT protocols. All bio-layer interferometry (BLI) measurements were performed on an Octet RED384 system (ForteBio). Biosensors coated with streptavidin (SA) were

incubated in BLI Buffer (20 mM HEPES pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 10 µg/mL Heparin, 50 µg/mL bovine serum albumin, 0.01% v/v IGEPAL CA-630, 1 mM TCEP, 10% v/v glycerol) for ~10 min prior to assay. 5'-biotinylated target DNA (ligand) and corresponding non-target DNA was purchased from IDT as single-stranded oligos and annealed 1:1 according to standard IDT protocol (See Table S4 for oligo sequences).

Biotinylated dsDNA was diluted in BLI buffer to a concentration of 10 nM. dCas9 or MISER construct RNPs were diluted in BLI Buffer at various concentrations (0.1x to 10x reported K_D). BLI step sequence was as follows: SA biosensors were incubated in BLI buffer for 60 seconds (baseline); dsDNA ligands were loaded onto SA biosensors for 300 seconds (loading); SA biosensors were incubated in BLI buffer for 60 seconds again to re-equilibrate ligand-bound tip (baseline); dsDNA-functionalized biosensors were incubated with RNP analytes for 1000 seconds (association); and biosensors were incubated in baseline wells from Step 1 for 1000 seconds (dissociation). All steps were performed at 37° C with stirring (1000 RPM). Data analysis was performed with Octet Data Analysis HT software (ForteBio).



<u>Figure 2.14:</u> Bio-layer interferometry (BLI) controls. A) BLI experiments were performed by incubating immobilized dCas9 with dsDNA containing a target spacer but no PAM (orange trace). Transient PAM interactions have a significant contribution to the k_{on} of association. The signal is lost immediately in the dissociation step, which suggests that the interaction is nonspecific. Conversely, incubation with a dsDNA containing no spacer and no PAM shows no signal (purple). **B)** BLI traces of Δ3CE and Δ4CE binding to dsDNA show that the relative binding is minimal at 300 nM, even with a 3-bp bubble in the seed region of the target (orange and purple). Subsequently a concentration of 1000 nM was used for these constructs. Dotted lines represent Δ3CE and Δ4CE RNPs interacting with a target without complementary spacers but containing NGG PAMs. Light grey and dark grey traces represent Δ3CE and Δ4CE RNPs, respectively, against dsDNA without a spacer or PAM. All data shown are normalized to the maximum signal of dCas9 vs. fully complementary dsDNA target (black).

2.4.8 Mammalian cell culture

All mammalian cell cultures were maintained in a 37°C incubator, at 5% CO_2. HEK293T human kidney cells (293FT; Thermo Fisher Scientific, #R70007) 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 µg/ml streptomycin (100-Pen-Strep; Gibco, #15140-122). U-251 human glioblastoma cells (Sigma-Aldrich, #09063001) and derivatives thereof were cultured in Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 (DMEM/F12; Gibco, #11320-033) supplemented with 10% FBS and 100-Pen-Strep. U-251 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 and U-251 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.

2.4.9 Mammalian CRISPR interference (CRISPRi) assay

For the mammalian CRISPR interference (CRISPRi) based competitive proliferation assay, human U-251 glioblastoma cells were stably transduced with lentiviral vectors (pSC066) expressing MISER or WT-dCas9 KRAB fusion proteins, followed by selection on puromycin (InvivoGen, #ant-pr-1; 1.0-2.0 µg/ml). The respective cell lines were then transduced with a secondary lentiviral vector (pCF221) expressing mCherry fluorescence protein and either CRISPRi sgRNAs targeting essential genes (sgPCNA, sgRPA1) or non-targeting controls (sgNT). After mixing with the respective parental population (at approximately an 80:20 ratio of transduced to non-transduced cells), the percentage of mCherry positive cells was monitored by flow cytometry (Attune NxT flow cytometer, ThermoFisher Scientific) over several days to assess the effect of CRISPRi with the given Cas9-variant on cell proliferation. CRISPR interference (CRISPRi) sgRNAs had been previously designed [4], as were non-targeting sgRNAs [5]. The sgRNAs were designed with a G preceding the 20-nucleotide guide for better expression from U6 promoters and cloned into the pCF221 lentiviral vector for expression [6].

2.4.10 Lentiviral vectors

A set of lentiviral vectors referred to as pSC066-GOI (gene-of-interest) – expressing an EF1a-driven polycistronic cassette containing a MISER-dCas9 or WT-dCas9 KRAB fusion protein, P2A ribosomal skipping element, and a puromycin

resistance marker – were based on the pCF525 lentiviral vector (Watters et al., 2018) derived from pCF204[6]. In brief, the original expression cassette in pCF525 was replaced by the above described EF1a-driven KRAB-MISER-dCas9-P2A-PuroR or KRAB-WT-dCas9-P2A-PuroR polycistronic constructs using custom oligonucleotides (IDT), gBlocks (IDT), standard cloning methods, and Gibson assembly techniques (NEB). Single-guide RNAs (sgRNAs) were expressed from the pCF221 vector[6]. The recipient vector (pCF221-reci) contains Esp3I (BsmBI) restriction sites for sgRNA cloning. The respective sequence (GGAGACGGAGGACGACGAACGTCTCT) is expressed as protospacer in pCF221-reci.

2.4.11 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 µg lentiviral vector, 4 µg psPAX2 and 2 µg pMD2.G were mixed in 2 ml Opti-MEM, followed by addition of 42 µg PEI. After 20-30 min incubation at room temperature, the transfection reactions were dispersed over the HEK293T cells. Media was changed 12 h post-transfection, and virus harvested at 36-48 h post-transfection. Viral supernatants were filtered using 0.45 µm cellulose acetate or polyethersulfone (PES) membrane filters, diluted in cell culture media if appropriate, and added to target cells. Polybrene (5 µg/ml; Sigma-Aldrich) was supplemented to enhance transduction efficiency, if necessary. Transduced target cell populations (U-251) were usually selected 24-48 h post-transduction using puromycin (InvivoGen, #ant-pr-1; 1.0-2.0 µg/ml).

2.4.12 Flow cytometry for mammalian CRISPRi assays

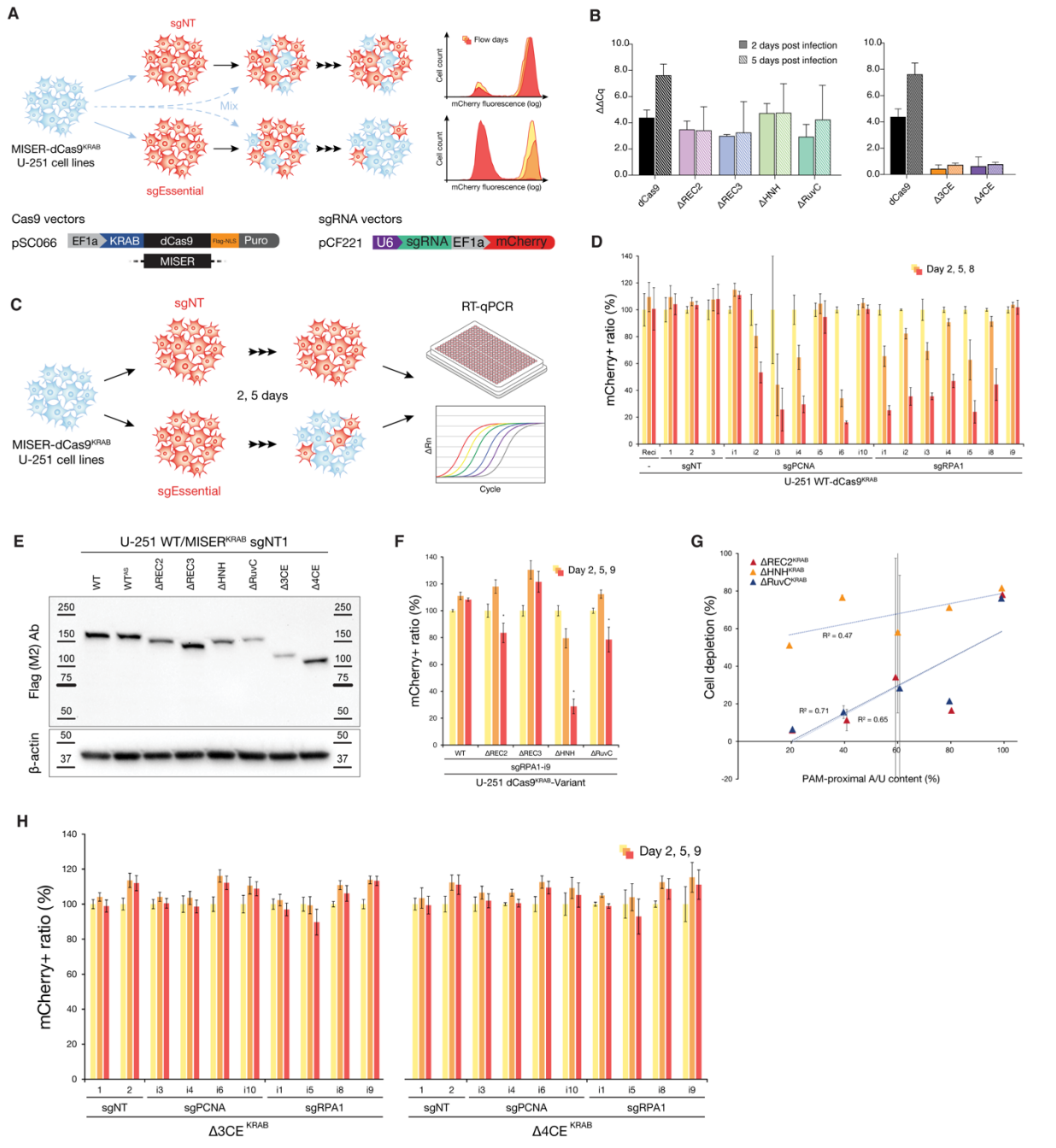
The percentage of mCherry or GFP positive cells was quantified by flow cytometry (Attune NxT flow cytometer, Thermo Fisher Scientific), regularly acquiring 10,000-30,000 events per sample.

2.4.13 Mammalian immunoblotting

Stably transduced U-251 cells expressing constructs of interest were washed with 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 + 0.1% Tween 20) for 1 h. 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. Immunoblotting was performed using the following antibodies: mouse monoclonal Anti-Flag-M2 (Sigma-Aldrich, #1804, clone M2, 1:500; HRP-conjugated mouse monoclonal Anti-Beta-Actin (Santa Cruz Biotechnology, #sc-47778 HRP, clone C4, 1:250; and HRP-conjugated sheep Anti-Mouse (GE Healthcare Amersham ECL, #NXA931; 1:5000. 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).

2.4.14 Reverse-transcription quantitative PCR (RT-qPCR)

To measure the efficacy of CRISPRi repression of essential genes by dCas9-MISER constructs in cultured mammalian cells, we performed RT-qPCR of targeted genes in human U-251 glioblastoma cells. Cells were stably transduced with lentiviral vectors encoding dCas9- or MISER-KRAB proteins, and sgRNA targeting PCNA (sgPCNA-i6) as described in the mammalian CRISPRi experiment (including non-targeting guide sgNT-1), except without any mixing with the parental population. Cells were allowed to grow and then harvested 2 and 5 days post-transduction. RNA was extracted using Trizol-chloroform and stored in -80° C [7]. RNA was reverse-transcribed to cDNA with RNA-to-cDNA EcoDry™ Premix with random hexamers (Takara Bio), using manufacturer's protocols. Quantitative PCR (qPCR) amplification of cDNA was performed using primers specific for *PCNA* (oAS089-92, Table S4) using SYBR Green PCR Master Mix (ThermoFisher Scientific) in a QuantStudio 3 Real-time PCR System (ThermoFisher Scientific). GAPDH was used as the housekeeping control (amplified with primers oAS117-118, Table S4). All results are reported relative to the expression of PCNA in cells transfected with non-target gRNA (sgNT-1, Table S4). Only amplification plots below a ΔRn threshold of 0.040 and a C_t value <35 cycles were used for analysis of expression levels. ΔC_q values were calculated by subtracting Cq values of GAPDH amplifications from PCNA, and $\Delta\Delta C_q$ values were calculated by subtracting the non-target samples from the target samples. Fold-change in expression is reported as $2^{-\Delta\Delta Cq}$.



<u>Figure 2.15:</u> Schematic of CRISPR interference (CRISPRi) based survival assay. A) Schematic of CRISPR interference (CRISPRi) based competitive proliferation assay. U-251 glioblastoma cells are stably transduced with lentiviral vectors (pSC066) expressing MISER-dCas9 or WT-dCas9 KRAB fusion proteins, followed by selection on puromycin. The various cell lines are then transduced with a secondary lentiviral vector (pCF221) expressing mCherry fluorescence protein and either sgRNAs targeting essential genes (sgEssential) or non-targeting sgRNAs (sgNT) as controls. After mixing with the respective parental population, the percentage of mCherry-positive cells is monitored by flow cytometry over several days. **B)** PCNA $\Delta\Delta C_q$ values from RT-qPCR at 2 (solid) and 5 (hatched) days post infection, calculated by subtracting target samples from sgNT samples. Values are plotted from biological duplicates as

mean±S.D. **C)** U-251 glioblastoma cells are stably transduced with lentiviral vectors (pSC066) expressing MISER-dCas9 or WT-dCas9 KRAB fusion proteins, followed by selection on puromycin. The various cell lines are then transduced with a secondary lentiviral vector (pCF221) expressing mCherry fluorescence protein and either sgRNAs targeting essential genes (sgPCNA) or non-targeting sgRNAs (sgNT) as controls. Cells are grown and harvested 2 and 5 days post-infection for RNA extraction, followed by RT-qPCR to quantitate transcription of targeted essential genes under MISER-KRAB repression. **D)** U-251 cells stably expressing a wild-type dCas9 KRAB fusion protein (WT-dCas9-KRAB) were transduced with lentiviral vectors expressing the indicated sgRNAs. At Day-2 post-transduction, cells were mixed with the parental population; mCherry fluorescence was monitored over time. sgNT, non-targeting control sgRNAs. sgPCNA and sgRPA1, sgRNAs targeting essential genes. Reci, recipient vector for sgRNA cloning. Error bars indicate SD of triplicates. **E)** Immunoblotting for Flag-tagged MISER-dCas9 or WT-dCas9 KRAB fusion proteins stably expressed in U-251 cells co-expressing a non-targeting guide (sgNT1). The indicated MISER deletions result in reduction of protein size. WT^{AS} represents an alternative out-of-frame start-codon derived from the native sequence of the KRAB domain. Beta-actin (ACTB) was used as loading control. Protein ladders indicate reference molecular weight markers in kDa. **F)** Competitive proliferation assay as in (D). Note, the indicated sgRNA (sgRPA1-i9) shows stronger depletion with some of the MISER variants when compared to the WT-dCas9 KRAB fusion. Significance in increased cell depletion was assessed by comparing samples to the wild-type control using unpaired, two-tailed t-tests (alpha = 0.01). **G)** Correlation between PAM-proximal A/U content of sgRNAs (5 most proximal bases) and cell depletion efficiency at day 9 of the competitive proliferation assay for the indicated MISER-dCas9 KRAB fusion variants. The scatter plot represents data from sgPCNA-i3/i4/i6 and sgRPA1-i1/i5/i8/i9. Dotted lines indicate linear regressions (ΔREC2 R^2 = 0.65, ΔHNH R^2 = 0.47, ΔRuvC R^2 = 0.71). **H)** Competitive proliferation assay as in Fig 1E, with stacked-deletion constructs Δ3CE and Δ4CE. Error bars indicate the SD of triplicates. Significance in cell depletion was assessed by comparing samples to their respective Day-2 controls using unpaired, two-tailed t-tests (alpha = 0.01).

2.4.15 Cryo-electron microscopy sample preparation and image acquisition

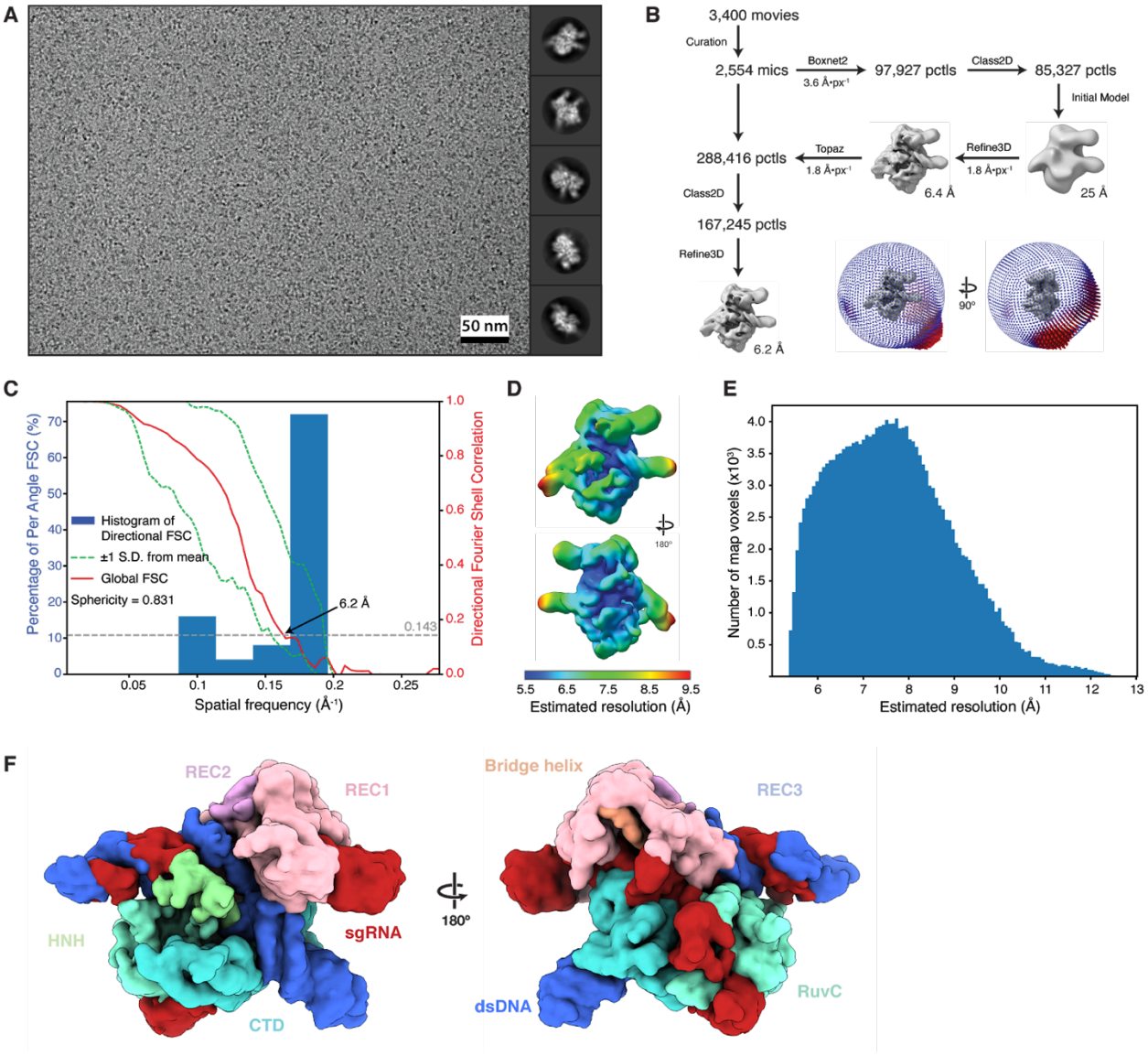
The ternary complex was prepared at 37 °C using a Δ4CE, sgRNA, and dsDNA target at a ratio of 1:1.5:2 in complexing buffer (30 mM Tris-HCl, pH 8.0, 150 KCl, 5 mM $MgCl_2$, 5 mM DTT, 2.5 % glycerol). Protein and sgRNA were incubated for 30 minutes prior to addition of dsDNA for an additional 1 hour of incubation. The sample was then desalted using a spin-column (Zeba) into Complexing Buffer containing 0.1% glycerol to be used for grid preparation. To prepare the sample for imaging, 3.2 µL of the ternary complex (around 30 nM) was applied to R1.2/1.3 Cu 200 grids (Quantifoil) coated with a thin layer of homemade continuous carbon that had been glow-discharged for 15 s immediately before use. The sample was incubated on the grid at 100% humidity and 16 °C for 10 s prior to blotting for 5 s with filter paper and plunging into liquid ethane cooled to liquid nitrogen temperatures using a Vitrobot Mark IV (TFS). The sample was imaged using a Talos Arctica transmission electron microscope (TFS) operated at 200 kV and equipped with a K3 direct electron detector (Gatan) at the Bay Area Cryo-EM facility at the University of California, Berkeley. Movies were recorded in super-resolution counting mode at an effective pixel size of 0.45 Å, with a cumulative exposure of 60 e^- $\cdot Å^{-2}$ distributed uniformly over 60 frames. Automated data acquisition was performed using image-shift and active beam tilt compensation as implemented in SerialEM-v3.7 to acquire movies from a 3x3 array of holes per stage movement [8]. In total, 3400 movies were acquired with a realized defocus range of -1.5 to -3.8 µm.

2.4.15 Cryo-EM image processing

All steps were performed using RELION-v3.1b unless otherwise indicated[9]. Movies were motion-corrected, exposure-filtered, and Fourier cropped to a pixel size of 0.9 Å using and the initial CTF parameters estimated by CTFFIND-v4.1.13 [10]. Micrographs were culled by thresholding for CTF-fit resolutions better than 8 Å and manual curation to yield a set of 2554 micrographs used in further processing. An initial set of 97,827 particles were picked using the general model of Boxnet2 [11]. These particles were extract in a 256 pixel box Fourier cropped to 64 pixels (3.6 Å·px^{-1}). Iterative rounds of reference-free 2D classification resulted in 85,327 particles, which were used to generate an ab initio 3D-reference by stochastic gradient descent. Particles were re-extracted and upsampled in a 128 pixel box (1.8 Å·px^{-1}) for further processing. Unsupervised 3D classification did not resolve distinguishable classes. Thus, all particles were subjected to 'gold-standard' 3D auto-refinement using a reference low-pass filtered to 25 Å and a soft shape-mask. This yielded a reconstruction at a nominal resolution of 6.4 Å based on the FSC0.143 criterion and using phase-randomization to correct for masking artifacts [12]. This set of particles was then used to train a picking model with Topaz-v0.2.3 [13]. This approach resulted in a set of 288,416 particle coordinates. The new set of particles was extracted in a 128 pixel box (1.8 Å·px^{-1}) and subjected to reference-free 2D classification, which resulted in a set 167,245 particles. Additional attempts at 3D classification did not resolve distinguishable classes. This final set of particles was used for 3D auto-refinement as described above and resulted in a 6.2 Å reconstruction. Further processing using reference-based fitting of particle motion and CTF parameters did not yield improvements. Resolution anisotropy of the final reconstruction was assessed using the 3DFSC web server [14].

2.4.16 Modelling of the cryo-EM map

The previously published coordinate model for the 5.2Å cryo-EM structure of SpCas9 ternary complex (PDB ID 5Y36) was used as an initial model [15]. To this end, the protein domains were deleted from 5Y36 to match those of Δ4CE. The edited coordinate model was then docked as a rigid-body into the RELION post-processed map using ChimeraX-v1.0, which resulted in a cross-correlation value of 0.73 against a 6.2 Å map simulated from the coordinate model [16]. For display purposes, a denoised version of the Δ4CE map was generated with LAFTER as part of the CCPEM-v1.4.1 suite [17].



<u>Figure 2.16:</u> Single-particle cryo-EM of the Δ4Cas9 ternary complex. A) Exemplar micrograph at approximately 3 microns defocus with scale indicated and representative reference-free 2D class averages from the Topaz-picked particle set. Diameter of 2D mask is 150 Å in all averages. **B)** Single-particle reconstruction work-flow as described in methods and orientation distribution of the final reconstruction inset. (C) Directional FSC for final reconstruction. **D) and E)** Local resolution estimates calculated in RELION shown by coloration on the map and as a histogram, respectively. **F)** Density map of Δ4CE with putative domains segmented and colored according to their relative position within a 20 Å radius when overlaid on WT SpCas9 (PDB 5Y36).

EMDB-22518	
Data Collection	
Microscope	Talos Arctica
Magnification	45,000
Voltage (kV)	200
Detector	K3
Electron exposure (e-/Å^2)	60
Defocus range (μm)	1.5 to 3.8
Pixel size (Å)	0.45[a]
Reconstruction	
Symmetry imposed	C1
Box size (pixels/Å)	128/230
Initial particle images (no.)	288,416[b]
Final particle images (no.)	167,245
Map resolution (Å)	6.2
FSC threshold	0.143
Sharpening factor (Å^2)	-395
Map resolution range (Å)	5.5-9.5
Sphericity	0.831
Modeling	
Method	Rigid-body
Initial Model	5Y36
CC	0.73

[a]Super-resolution
[b]from picking with Topaz

Table 2.5: Cryo-EM data collection & reconstruction statistics.

Oligo ID	Purpose	Sequence (5'-3')
SAH_284	Recombineering amplification: universal forward	AACACGTCCGTCCTAGAACT
SAH_285	Recombineering amplification: universal reverse	ACTTGGTTACGCTCAACACT
SAH_286	Recombineering amplification: SpeI-specific reverse	GATCTGAGTGTACCGCTTGC
SAH_287	Recombineering amplification: NheI-specific reverse	GATCGCCTAGACAACTCCTG
sgRNA-B9	sgRNA for Cas9 RNP, used in BLI and cryo-EM	AGUCGGUGUCGACCCGGACCCAAAAUCUCGAUC UUUAUCGUUCAAUUUUAUUCCGAUCAGGCAAUAG UUGAACUUUUUCACCGUGGCUCAGCCACGAAAA
oAS081	5'-biotinylated ssDNA target for BLI, sgRNA-B9	GCTCAATTTTGACAGCCCACCAGGCCCAGCTGTG GCTGATGGCATCCTTCCACTC
oAS003a	non-target ssDNA for BLI (complementary to oAS081)	GAGTGGAAGGATGCCATCAGCCACAGCTGGGCCT GGTGGGCTGTCAAAATTGAGC
oAS114	5'-biotinylated ssDNA non-target for BLI (no spacer, no PAM)	GTGTGCACACATGCAATAACATTGTGCACATGATA CATTGCAATGACAATTAACC
oAS036	non-target ssDNA for BLI (complementary to oAS081, 3-bp PAM-proximal bubble)	GAGTGGAAGGATGCCATCAGCCACAGCTGGGCC GATTGGGCTGTCAAAATTGAGC
oAS116	unlabeled ssDNA target for BLI, sgRNA-B9. Used for cryo-EM RNP complex	GCTCAATTTTGACAGCCCACCAGGCCCAGCTGTG GCTGATGGCATCCTTCCACTC
sgNT-1	Non-targeting gRNA for mammalian CRISPRi	GGCCAAACGTGCCCTGACGG
sgNT-2	Non-targeting gRNA for mammalian CRISPRi	GCGATGGGGGGGTGGGTAGC
sgPCNA-i1	*PCNA* targeting gRNA for mammalian CRISPRi	GGGGCGAACGTCGCGACGAC
sgPCNA-i2	*PCNA* targeting gRNA for mammalian CRISPRi	GGCGTGGTGACGTCGCAACG
sgPCNA-i3	*PCNA* targeting gRNA for mammalian CRISPRi	GCGCTCCCGCCAAGCACCGG
sgPCNA-i4	*PCNA* targeting gRNA for mammalian CRISPRi	GAAGCGCTCCCGCCAAGCAC
sgPCNA-i5	*PCNA* targeting gRNA for mammalian CRISPRi	GCCCGGCCCGCCTGCACCTC
sgPCNA-i6	*PCNA* targeting gRNA for mammalian CRISPRi	GCGGACGCGGCGGCATTAAA
sgPCNA-i10	*PCNA* targeting gRNA for mammalian CRISPRi	GGCCATCCGCGCCTTCTCAT
sgRPA1-i1	*RPA* targeting gRNA for mammalian CRISPRi	GGGAAGCTGGAGCTGTTGCG
sgRPA1-i2	*RPA* targeting gRNA for mammalian CRISPRi	GGCGACGGGGGATGAACGCG
sgRPA1-i3	*RPA* targeting gRNA for mammalian CRISPRi	GTGCGCAGCGCGCGGGACCC

sgRPA1-i4	*RPA* targeting gRNA for mammalian CRISPRi	GTGAGCCGCGCGCACGTCGG
sgRPA1-i5	*RPA* targeting gRNA for mammalian CRISPRi	GGCGGTGCGCGCAACTTCTC
sgRPA1-i8	*RPA* targeting gRNA for mammalian CRISPRi	GCGAGCCTCGCGGAGTAGAG
sgRPA1-i9	*RPA* targeting gRNA for mammalian CRISPRi	GCCGCGCGCTGCGCAGTTAT
oAS085	Forward primer for *RPA1* cDNA reverse transcription, set 1	GCAGTTGGAGTGAAGATTGG
oAS086	Reverse primer for RPA1 cDNA RT, set 1	CACTTGGACTGGTAAGGAGT
oAS087	Forward primer for *RPA1* cDNA RT, set 2	CCGAGCTACAGCTTTCAATG
oAS088	Reverse primer for RPA1 cDNA RT, set 2	GCAGATCCCGATGATGTCTA
oAS089	Forward primer for *PCNA* cDNA RT, set 1	ACTCAAGGACCTCATCAACG
oAS091	Reverse primer for PCNA cDNA RT, set 1	TGAACCTCACCAGTATGTCC
oAS090	Forward primer for *PCNA* cDNA RT, set 2	CGTTATCTTCGGCCCTTAGT
oAS092	Reverse primer for PCNA cDNA RT, set 2	CGTGCAAATTCACCAGAAGG
oAS117	Forward primer for *GAPDH* RT	TCAAGGCTGAGAACGGGAAG
oAS118	Reverse primer for *GAPDH* cDNA RT	TGGACTCCACGACGTACTCA
oAS034	Forward primer for cloning dCas9 and MISER constructs into expression vector	GGTATCAACTTTTCGTTTCTT
oAS035	Reverse primer for cloning dCas9 and MISER constructs into expression vector	CAAAGCCCGAAAGGAAG

Table 2.6: Oligonucleotides used in this study.

2.5 Acknowledgments

This work was supported by NIH grants 1R01GM127463 (D.F.S.) and RM1HG009490 (J.A.D., D.F.S). Additional support and reagents were provided by Agilent Technologies. A.S. was supported by the NSF GRFP (grant no. 1752814), S.A.H. was supported by NIH Training Grant 5T32GM066698-10, and M.A. was supported by the ARCS Foundation. C.F. was supported by a U.S. NIH K99/R00 Pathway to Independence Award (K99GM118909, R00GM118909) from the NIGMS. J.A.D. is an Investigator of the Howard Hughes Medical Institute (HHMI), and this study was supported in part by HHMI. This work used the Vincent J. Coates Genomics Sequencing Laboratory at University of California, Berkeley, supported by NIH S10 instrumentation grants (S10RR029668 and S10RR027303). We thank Mary West and the CIRM/QB3

Shared Stem Cell Facility/High-Throughput Screening Facility for technical support, as well as Timothy Brown (Thermo Fisher Scientific) for flow cytometry support. We thank Daniel Toso and Paul Tobias for assistance with cryo-EM data collection at Bay Area Cryo-EM facility at the University of California, Berkeley. We also thank Emeric Charles, Shin Kim, Rob Nichols, Luke Oltrogge, Avi Flamholz, and Joshua Cofsky for technical support and productive discussions.

2.6 Declaration of Interests

UC Regents have filed a patent related to this work. S.A.H. is an employee of Scribe Therapeutics. B.L.O. and B.T.S. are co-founders and employees of Scribe Therapeutics. C.F. is a co-founder of Mirimus, Inc. J.A.D. is a co-founder of Caribou Biosciences, Editas Medicine, Intellia Therapeutics, Scribe Therapeutics, and Mammoth Biosciences. J.A.D. is a scientific advisory board member of Caribous Biosciences, Intellia Therapeutics, eFFECTOR Therapeutics, Scribe Therapeutics, Synthego, Metagenomi, Mammoth Biosciences, and Inari. J.A.D is a member of the board of directors at Driver and Johnson & Johnson. D.F.S. is a co-founder of Scribe Therapeutics and a scientific advisory board member of Scribe Therapeutics and Mammoth Biosciences. All other authors declare no competing interests.

2.7 Data Availability Statement

All custom scripts are available at https://github.com/savagelab. All sequencing data that support the findings of this study are available from the authors upon reasonable request. Cryo-EM data are available at the Electron Microscopy DataBank (EMDB); accession code 22518. All other relevant data are available from the corresponding author on request.

2.8 Author Affiliations

Arik Shams[1,*], Sean A. Higgins[1,2,3,*], Christof Fellmann[1,4,5], Thomas G. Laughlin[1,6], Benjamin L. Oakes[1,2,3], Rachel Lew[4], Shin Kim[1,2], Maria Lukarska[1,2], Madeline Arnold[1], Brett T. Staahl[1,2,3], Jennifer A. Doudna[1,2,4,7,8,9,10,11], David F. Savage[1,^]

Affiliations:

[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[2]Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA 94720, USA
[3]Scribe Therapeutics, Alameda, CA 94501, USA
[4]Gladstone Institutes, San Francisco, CA 94158, USA
[5]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA
[6]Division of Biological Sciences, University of San Diego, San Diego, CA 92093, USA
[7]Graduate Group in Biophysics, University of California, Berkeley, Berkeley, CA 94720, USA
[8]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720, USA
[9]Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA 94720, USA
[10]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[11]Department of Chemistry, University of California, Berkeley, Berkeley, CA 94720, USA

* These authors contributed equally to this work.
^Correspondence: savage@berkeley.edu.

2.9 References

1. Teichmann, S. A., Park, J. & Chothia, C. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. USA* **95,** 14658–14663 (1998).
2. Apic, G., Gough, J. & Teichmann, S. A. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310,** 311–325 (2001).
3. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300,** 1701–1703 (2003).
4. Koonin, E. V., Aravind, L. & Kondrashov, A. S. The impact of comparative genomics on our understanding of evolution. *Cell* **101,** 573–576 (2000).
5. Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. & Teichmann, S. A. Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14,** 208–216 (2004).
6. Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A. & Clarke, J. The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **8,** 319–330 (2007).
7. Weiner 3rd, J., Beaussart, F. & Bornberg-Bauer, E. Domain deletions and substitutions in the modular protein evolution. *FEBS J.* **273,** 2037–2047 (2006).
8. Basu, M. K., Carmel, L., Rogozin, I. B. & Koonin, E. V. Evolution of protein domain promiscuity in eukaryotes. *Genome Res.* **18,** 449–461 (2008).
9. Koonin, E. V., Makarova, K. S. & Zhang, F. Diversity, classification and evolution of CRISPR-Cas systems. *Curr. Opin. Microbiol.* **37,** 67–78 (2017).
10. Kriventseva, E. V. *et al.* Increase of functional diversity by alternative splicing. *Trends Genet.* **19,** 124–128 (2003).
11. Dueber, J. E., Yeh, B. J., Bhattacharyya, R. P. & Lim, W. A. Rewiring cell signaling: the logic and plasticity of eukaryotic protein circuitry. *Curr. Opin. Struct. Biol.* **14,** 690–699 (2004).
12. Guntas, G. & Ostermeier, M. Creation of an allosteric enzyme by domain insertion. *J. Mol. Biol.* **336,** 263–273 (2004).
13. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147,** 1564–1575 (2011).
14. Oakes, B. L. *et al.* Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nat. Biotechnol.* **34,** 646–651 (2016).
15. Atkinson, J. T., Jones, A. M., Zhou, Q. & Silberg, J. J. Circular permutation profiling by deep sequencing libraries created using transposon mutagenesis. *Nucleic Acids Res.* **46,** e76 (2018).
16. Oakes, B. L. *et al.* CRISPR-Cas9 Circular Permutants as Programmable Scaffolds for Genome Modification. *Cell* **176,** 254–267.e16 (2019).
17. Simm, A. M., Baldwin, A. J., Busse, K. & Jones, D. D. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 beta-lactamase. *FEBS Lett.* **581,** 3904–3908 (2007).
18. Hecky, J. & Müller, K. M. Structural perturbation and compensation by directed evolution at physiological temperature leads to thermostabilization of beta-lactamase. *Biochemistry* **44,** 12640–12654 (2005).

19. Ma, D., Peng, S., Huang, W., Cai, Z. & Xie, Z. Rational Design of Mini-Cas9 for Transcriptional Activation. *ACS Synth. Biol.* **7,** 978–985 (2018).
20. Jones, D. D. Triplet nucleotide removal at random positions in a target gene: the tolerance of TEM-1 beta-lactamase to an amino acid deletion. *Nucleic Acids Res.* **33,** e80 (2005).
21. Arpino, J. A. J., Reddington, S. C., Halliwell, L. M., Rizkallah, P. J. & Jones, D. D. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure* **22,** 889–898 (2014).
22. Pisarchik, A., Petri, R. & Schmidt-Dannert, C. Probing the structural plasticity of an archaeal primordial cobaltochelatase CbiX(S). *Protein Eng Des Sel* **20,** 257–265 (2007).
23. Ostermeier, M., Shim, J. H. & Benkovic, S. J. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.* **17,** 1205–1209 (1999).
24. Morelli, A., Cabezas, Y., Mills, L. J. & Seelig, B. Extensive libraries of gene truncation variants generated by in vitro transposition. *Nucleic Acids Res.* **45,** e78 (2017).
25. Araya, C. L. & Fowler, D. M. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.* **29,** 435–442 (2011).
26. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
27. Oakes, B. L., Nadler, D. C. & Savage, D. F. Protein engineering of Cas9 for enhanced function. *Meth. Enzymol.* **546,** 491–511 (2014).
28. Jiang, F. & Doudna, J. A. CRISPR-Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.* **46,** 505–529 (2017).
29. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507,** 62–67 (2014).
30. O'Connell, M. R. *et al.* Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature* **516,** 263–266 (2014).
31. Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **374,** 20180087 (2019).
32. Jiang, F., Zhou, K., Ma, L., Gressel, S. & Doudna, J. A. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348,** 1477–1481 (2015).
33. Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351,** 867–871 (2016).
34. Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Sci. Adv.* **3,** eaao0027 (2017).
35. Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156,** 935–949 (2014).
36. Nishimasu, H. *et al.* Crystal Structure of Staphylococcus aureus Cas9. *Cell* **162,** 1113–1126 (2015).
37. Thomason, L. C., Costantino, N., Shaw, D. V. & Court, D. L. Multicopy plasmid modification with phage lambda Red recombineering. *Plasmid* **58,** 148–158 (2007).

38. Higgins, S. A., Ouonkap, S. V. Y. & Savage, D. F. Rapid and programmable protein mutagenesis using plasmid recombineering. *ACS Synth. Biol.* **6,** 1825–1833 (2017).
39. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152,** 1173–1183 (2013).
40. Larson, M. H. *et al.* CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat. Protoc.* **8,** 2180–2196 (2013).
41. Babu, K. *et al.* Bridge helix of cas9 modulates target DNA cleavage and mismatch tolerance. *Biochemistry* **58,** 1905–1917 (2019).
42. Bratovič, M. *et al.* Bridge helix arginines play a critical role in Cas9 sensitivity to mismatches. *Nat. Chem. Biol.* **16,** 587–595 (2020).
43. Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513,** 569–573 (2014).
44. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550,** 407–410 (2017).
45. Sternberg, S. H., LaFrance, B., Kaplan, M. & Doudna, J. A. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527,** 110–113 (2015).
46. Huai, C. *et al.* Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nat. Commun.* **8,** 1375 (2017).
47. Huang, T. P. *et al.* Circularly permuted and PAM-modified Cas9 variants broaden the targeting scope of base editors. *Nat. Biotechnol.* **37,** 626–631 (2019).
48. Xu, H. *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25,** 1147–1157 (2015).
49. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci. USA* **108,** 10098–10103 (2011).
50. Grishin, N. V. Fold change in evolution of protein structures. *J. Struct. Biol.* **134,** 167–185 (2001).
51. Wright, A. V. *et al.* Rational design of a split-Cas9 enzyme complex. *Proc. Natl. Acad. Sci. USA* **112,** 2984–2989 (2015).
52. Klein, M., Eslami-Mossallam, B., Arroyo, D. G. & Depken, M. Hybridization Kinetics Explains CRISPR-Cas Off-Targeting Rules. *Cell Rep.* **22,** 1413–1423 (2018).
53. Fong, J. H., Geer, L. Y., Panchenko, A. R. & Bryant, S. H. Modeling the evolution of protein domain architectures using maximum parsimony. *J. Mol. Biol.* **366,** 307–315 (2007).
54. Prakash, A. & Bateman, A. Domain atrophy creates rare cases of functional partial protein domains. *Genome Biol.* **16,** 88 (2015).
55. Grieger, J. C. & Samulski, R. J. Packaging capacity of adeno-associated virus serotypes: impact of larger genomes on infectivity and postentry steps. *J. Virol.* **79,** 9933–9944 (2005).
56. Wu, Z., Yang, H. & Colosi, P. Effect of genome size on AAV vector packaging. *Mol. Ther.* **18,** 80–86 (2010).

57. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520,** 186–191 (2015).
58. van Haasteren, J., Li, J., Scheideler, O. J., Murthy, N. & Schaffer, D. V. The delivery challenge: fulfilling the promise of therapeutic genome editing. *Nat. Biotechnol.* **38,** 845–855 (2020).
59. Crudele, J. M. & Chamberlain, J. S. Cas9 immunity creates challenges for CRISPR gene editing therapies. *Nat. Commun.* **9,** 3497 (2018).
60. Mehta, A. & Merkel, O. M. Immunogenicity of cas9 protein. *J. Pharm. Sci.* **109,** 62–67 (2020).
61. Wang, A. S. *et al.* The Histone Chaperone FACT Induces Cas9 Multi-turnover Behavior and Modifies Genome Manipulation in Human Cells. *Mol. Cell* **79,** 221–233.e5 (2020).
62. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. BioNumbers--the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* **38,** D750-3 (2010); BNID 108263.
63. Chylinski, K., Makarova, K. S., Charpentier, E. & Koonin, E. V. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Res.* **42,** 6091–6105 (2014).
64. Crudele, J. M. & Chamberlain, J. S. AAV-based gene therapies for the muscular dystrophies. *Hum. Mol. Genet.* **28,** R102–R107 (2019).
65. Onda, M. Reducing the immunogenicity of protein therapeutics. *Curr Drug Targets* **10,** 131–139 (2009).
66. Baker, M. P., Reynolds, H. M., Lumicisi, B. & Bryson, C. J. Immunogenicity of protein therapeutics: The key causes, consequences and challenges. *Self Nonself* **1,** 314–322 (2010).

Chapter 3

Perspectives on Genetically Engineered Microorganisms and their Regulation[+]

[+] The work presented in this chapter is adapted from a manuscript in preparation with permission from the authors:

Shams A, Kliegman M, Fischer A, Bodnar A.

3.1 Introduction

Genetic engineering has ushered in a new technological age characterized by harnessing biological systems for human utility. While genetic modification has existed since ancient history (in the form of selective breeding in plant and animal husbandry), the direct manipulation of genomes via recombinant DNA and molecular biology has catapulted both our understanding and control of complex biological systems into a new "biotechnology era." Recent advances in technologies like genome editing, oligonucleotide synthesis, sequencing, and bioinformatic processing have pushed us into the realm of synthetic biology, or the de novo synthesis of life. These alterations started first in the smallest and most manipulatable forms of life, microorganisms, where due to their small size and rapid generation times, altering the genetic code of organisms is relatively simple. These technological advances raise obvious questions about ethical development and deployment, as well as the safety of releasing bioengineered products; any future where genetically engineered microorganisms play a major role must take into account not just safety for human health but also ecosystem balance and resource consumption. To that end, the regulation of bioengineered products is critical and needs to adapt alongside technological development while being nimble to accommodate products in the pipeline.

3.1.1 The Future of Microbial Biotechnology: From Research to Regulation

In February 2022, the Innovative Genomics Institute, the United States Department of Agriculture (USDA), and Phytobiomes Alliance jointly hosted a virtual workshop to foster open discussion between regulators, scientists, and developers on the future of GEMs (genetically engineered microorganisms), titled "The Future of Microbial Biotechnology: From Research to Regulation" [32]. The purpose of the workshop, which was open to the public, was to convene stakeholders that research, develop and regulate GEMs and GEM-based products to interact directly in an effort to demystify and develop a better understanding of each group's approach. All participants highlighted that their approaches use engineering in ways that leverage existing capabilities in natural microbes and most participants have worked with both natural and engineered microbes to achieve the same goals. In some cases, participants have remarked that there are no large-scale, promising products coming out of exploiting existing, wild-type species and therefore synthetic biology or genetic engineering have made valuable inroads in tackling specific challenges.

The workshop and the conversations that it fostered came in the midst of critical circumstances that are actively and drastically changing the field of biotechnology. Climate change necessitates that we quickly employ a broad and diverse range of solutions for emergent global crises: feeding a growing population, loss of biodiversity, greenhouse gas emissions, and fossil fuel dependence [33–35]. Biotechnology, and in particular GEMs, hold potential solutions or adaptations to many of these issues, and

have attracted huge boosts in interest and investment. This, coupled with major scientific advances, commercial opportunity, and changes in regulatory attitudes present a unique moment for developers and regulators alike.

In this chapter we will cover existing GEM products on the market, those in the pipeline and some of the major scientific advances that we expect will intersect with the regulatory agencies in the near future. This chapter will provide a summary of the meeting proceedings, specifically focused on the scientific research and the future of the field, while a companion article will provide an overview of how these products are regulated. We will, however, also highlight instances where there are regulatory gaps and make recommendations for cohesive, yet adaptive regulatory coverage of innovative new technologies in the future.

GEMs hold a tremendous amount of promise for keeping pace with a changing world and growing population. As a society we must have a clear understanding of products being developed, benefits conferred, risks posed and details of safety evaluations. This clear understanding is crucial for consumer confidence but similarly crucial for investors and developers as they navigate the uncertain territory of innovation.

3.1.2 GEMs: What are they?

Microorganisms have been controlled and cultivated in the production of foods, fuels, and materials for millennia. Even before humans knew of microbes' existence, they were used to make foods such as cheese, wine, and yogurt. In the 1800s, microorganisms' role in causing disease and food spoilage was discovered, and immediately spurred the race to manipulate them. Today, genetically engineering microorganisms to maximize efficiency and make new products is an increasingly common practice in industry.

The term microbe or microorganism itself is not clearly defined scientifically, as it contains species from all three domains of life (Archaea, Bacteria and Eukarya), but is rather defined in a more functional sense as "whole organisms which are in the microscopic scale." Entities around the world define the term differently, to suit various criteria, especially to apply regulatory oversight onto.

The USDA defines genetic engineering as "manipulation of an organism's genes by introducing, eliminating or rearranging specific genes using the methods of modern molecular biology, particularly those techniques referred to as recombinant DNA techniques [36]." However, when it comes to regulations, the USDA, Environmental Protection Agency (EPA), and Food and Drug Administration (FDA) regulate "biotechnology products" instead of the tools used to produce them under the guidance laid out in the Coordinated Framework updated in 2017 [28]. In other words, as long as a product itself does not contain recombinant DNA, such as a modified yeast or probiotic additive, it is not under the purview of the Coordinated Framework.

The emergence of genome-editing has added to the fluidity of these definitions, as it is now possible to make precise nucleotide changes within a gene to render it nonfunctional, or even introduce sequences from other naturally-occurring organisms.

Arguably, these changes are not “unnatural,” and may very well exist in nature as a result of random mutagenesis, splicing, or horizontal gene transfer [38]. In this light, the approach of the US government in regulating biotechnology products is functional and risk-based on a per-product level. This is a significantly more liberal view than the one taken by the European Food Safety Authority (EFSA) of the European Union, which does consider any product created with genetic engineering at any step of its manufacturing as a “genetically-modified organism,” subject to the requisite regulations and labeling standards [39]. In this review we make a deliberate distinction between “genetically modified” versus “genetically engineered.” While the former can include any manipulation made to an organism’s genome – by selective breeding, random mutagenesis, etc. – the latter term refers to specific, known changes made to the genome by engineering technologies to result in a desired phenotype. In many modern cases of genetic modifications where a genomic target is selected and then altered by insertion, deletion, or substitution, “engineered” is the more appropriate term. We chose to avoid the term “genetically modified organism” for two reasons: firstly, it is vague and outdated in the context of the spectrum of modifications possible; and secondly, the term has been a flashpoint of debate within the scientific as well as public and political spheres, and has also changed significantly in the public consciousness since the original definition of “any organism that has been genetically altered in a way not found in nature” [39]. The obvious issues with this definition include our incomplete knowledge of the full gamut of genetic variations found in nature and its narrow scope in considering only the product and not the process.

Throughout this review we use the term “genetically engineered microorganism” or “GEMs” to describe a microscopic organism with a genome directly altered using biotechnological methods. We will focus primarily on the technological development, current and potential uses, and intersection of policy and scientific advances in the field of engineered microbes. We note that there are critical distinctions between a product and a process in terms of their regulation: commercial microbes or microbial products are not categorically regulated by the USDA, FDA, or EPA, but *engineered* microbes are, as they fall under the scope of bioengineered organisms.

3.2 GEMs in Food and Agriculture

Microorganisms present a unique opportunity for the food and agriculture sector and many new organizations are using microbial engineering as a means to develop new food products, reduce greenhouse gas emissions and improve agricultural sustainability. GEMs in the food and agriculture industry generally fall into two categories; A) engineering microbes to alter plant-microbe interactions providing added benefits for plant cultivation or B) as food and feed additives that alter the qualities of the final product. Although there are many applications being developed in the early discovery phases, below we focus on GEM products that are either on or nearing the commercialization stage.

3.2.1 Fertilizer

Microbes can be found living in close association with plants and are required for many key plant metabolic functions [40,41]. Nitrogen (N), phosphorus (P) and potassium (K) are all essential macronutrients for plant growth and function that are typically applied to agricultural fields as synthetically-derived fertilizers. In all three cases, bioavailability of these nutrients to the plant is mediated, at least in part, by microorganisms. Replacing or reducing the need for synthetic fertilizer has been one area of intense focus for genetically engineered microorganisms.

Nitrogen fixation in legume plants occurs through a symbiotic relationship with micro-organisms who fix atmospheric nitrogen and provide nitrogen in the form of ammonia to plants. While there have been improvements to this process in legumes [42] which already form symbiosis with nitrogen-fixing bacteria, engineering approaches have had limited tractability in cereal crops which provide a majority of the world's calories and where most of the world's synthetic nitrogen fertilizer is directed [43].

To date, limitations in nitrogen availability have been circumvented through the industrialized Haber-Bosch process which uses high temperatures and pressures to combine atmospheric nitrogen with hydrogen gas to form ammonia [44]. In addition to the use of energy and release of greenhouse gasses (GHG) during the manufacturing process, globally ~50% of N applied to agricultural fields is lost to the environment [45]; industrially synthesized N is more prone to leaching and volatilization as nitrous oxide (N_2O) after application than biologically fixed nitrogen. Thus, improving N fixation by microorganisms presents a unique opportunity to improve environmental sustainability and GHG emissions.

During our workshop we heard from both Joyn Bio and Andes Ag, two biotechnology companies who are working to replace synthetic nitrogen fertilizer using GEMs. Both companies aim to alter the genes of naturally occurring microbes involved in nitrogen fixation with a focus on engineering symbiosis with cereal crops. In the case of Andes Ag, microbes are delivered to farmers on seeds to improve efficacy by ensuring interaction with the plant and reduced competition with soil microbes already present.

One product already on the market is PROVEN by Pivot Bio which reduces the use of nitrogen fertilizers by bioengineering microbes. In the case of Pivot Bio, naturally occurring bacteria were screened for nitrogen fixation qualities and then engineered to possess constitutive activation of genes involved in nitrogen production. These bacteria are applied with corn seeds upon planting to facilitate root assocation [46,47]. While Pivot Bio uses genetic engineering, all of the genes involved are endogenous to the species of bacteria and, as such, the bacteria are not considered transgenic. Other organizations and initiatives that aim to engineer microbes to reduce the need for nitrogen fertilizers include the Engineering Nitrogen Symbiosis for Africa initiative.

Second to nitrogen, both phosphorus and potassium are key to plant metabolism and physiology. Both P and K are present in soil but limited in bioavailable forms, they can be converted to forms more bioavailable to plants via microbial processes [48,49]. Although there are references in the scientific literature to engineering microorganisms to provide more P and K, there are no commercial attempts being made at this time that we are aware of.

3.2.2 Pesticides

Biological pesticides are an important part of modern agricultural practices. Prevalent in the organic sector, biopesticides are attractive due to their specificity, lack of toxicity to non-target and biodegradability when compared to chemically synthesized broad spectrum pesticides [50,51]. They have been effectively applied in the GEM space. The strain *Bacillus thuringiensis* (Bt), which produces proteins that impair pest digestive function, is one of the most widely used biopesticides and has been applied as a commercial product against lepidoptera[52]. A related approach has been to integrate specific genes from *B. thuringiensis* into the DNA of plants to avoid frequent applications due to environmental degradation.

One example of a microbial biopesticide approved for field trials in Europe is a fungicidal amoeba developed by French company Amoéba. They have developed an engineered amoeba, *Willaertia magna* C2c Maky, that feeds on the fungi that cause wheat rust, and has similar efficacy to chemical fungicides, according to the company's press release [53]. Another commercial product, Velifer, is a strain of the fungus *Beauveria bassiana* that is in use as a biopesticide against many insects and phytopathogenic bacteria. Other such fungi are also being researched as a biocontrol agent that is less toxic than conventional chemical pesticides [54].

3.2.3 Food enzymes and additives

One of the areas where there has been significant use of microbial biotechnology is in food enzymes and additions. Biotechnology has long been used to make a key enzyme in rennet (chymosin) which is used to make most hard cheeses on the market. Another dairy product, yogurt, was the source of the discovery of CRISPR systems, as

scientists sought to engineer bacteria at Danisco [55]. Milk is cultured with bacteria to make buttermilk, sour cream, cottage cheese and yogurt, typically with the bacteria *Lactobacillus acidophilus* and *Streptococcus thermophilus*. Use of non-engineered bacterial strains in cultured milk products are designated by the FDA as "Generally Regarded As Safe," or GRAS.

A full listing of notices and response letters together with a search engine is available on the GRAS notice website, which allows searches through the inventory that is up-dated monthly. Supplementary Table S1 gives the lactic acid bacteria (LABs) for which GRAS notices have been submitted to the FDA up to December 2019, and which are intended as live food additives. It only includes LABs for which the FDA has not questioned the GRAS conclusion of the notifier. For the use of harmless LABs, as optional ingredients in specified standardized foods, such as *Lactobacillus acidophilus*, prior approvals have been recognized (i.e., before 1958). Use of these bacteria is permitted in cultured milk (which includes buttermilk), sour cream, cottage cheese, and yogurt, provided that the mandatory cultures of *Lactobacillus bulgaricus* and *Streptococcus thermophilus* are also used in the yogurt [56].

There has been a recent boom in the use of genetic engineering to make beer, with some companies such as Omega Yeast [57] and Berkeley Yeast [58] providing new strains. Brewer's yeast is designated as safe by the FDA and thus does not need to undergo special regulatory approvals.

Another new area is that of replacement meat using microbial derived products as a means of improving sustainability related to the production of GHGs during meat production. Bruce Friedrich from the Good Foods Institute, an advocacy organization focused on alternative proteins explained that three-fourths of the land used for agriculture is dedicated towards growing feed crops or grazing ruminants [59]. Precision fermentation, using genetically engineered microbes, can produce similar flavor profiles, fats, proteins, and textures for consumers interested in meat alternatives. Some of the companies developing these products include Motif (bovine heme, proteins found in eggs and milk), Perfect Day and Nobel (dairy alternatives), Clara Foods, and Impossible Foods (legume hemoglobin).

3.3 GEMs in Biomanufacturing and the Environment

Synthetic biology and engineering of microorganisms can offer many benefits in the fields of biomanufacturing and industry that cannot be achieved by conventional petrochemical or organic synthesis-based methods. Microorganisms are already used in many industries to make fine, high-value chemicals, such as additives, pharmaceuticals, fragrances and flavors [60]. Common antibiotics (penicillin, erythromycin, vancomycin, etc.) [61–63], food additives (vitamins, monosodium glutamate) [64], and other pharmaceuticals have been successfully produced by microbial biosynthesis at scale. Insulin production in the U.S. is also primarily via microbial fermentation in *E. coli* and yeast [65]. Artemisinin, an antimalarial drug, is produced by engineered yeast by Amyris, Inc [66].

3.3.1 Fuels and commodity chemicals

Mass production of chemically simpler commodities like ethanol, fertilizers, and materials have traditionally relied on organic synthesis, often from petrochemicals as input. In the past, microbial fermentation has struggled to reach carbon parity with organic synthesis for many commodity chemicals, as the products from commercial heterotrophs like yeast are often reduced, lower-carbon products (e.g. ethanol) derived from higher-carbon raw materials (e.g. starch) [67]. Additionally, the production of CO_2 as a byproduct of fermentation makes it difficult to lower the carbon footprint. However, these issues are solvable in large part by genetic engineering of industrial microbes. Researchers have exploited the natural metabolic pathways of autotrophs like *Clostridium autoethanogenum*, which uses carbon monoxide (CO) and CO_2 to make acetate, a precursor to acetone and isopropanol [68].

As the biotechnology and green industries boom, scaling up microbial production of commodity chemicals are becoming more and more economical and carbon-neutral. Existing industrial biotechnologies, even today, promise to make significant reductions in GHG emissions, non-renewable energy consumption, and losses in efficiency [69,70]. Many companies have developed, or are developing, new genetically engineered microbes to improve fermentation efficiency, use novel raw materials, reduce carbon footprints, and redirect biological byproducts to more downstream biosynthetic production streams.

LanzaTech, for example, employs proprietary GEMs to use concentrated waste gasses at industrial facilities in order to make several products, including plastics and biofuels. By redirecting efflux from steel mills, Lanza can use concentrated CO, CO_2 and hydrogen gasses as input to improve efficiency and scale. Producing acetone and isopropanol is traditionally done via cracking or reforming propene, which is very energy intensive. Even conventional fermentation using sugar feedstocks is inefficient, as sugars are relatively complex molecules which release as a byproduct of lysis. Gas fermentation is a more efficient alternative since it recaptures waste CO_2 and other hydrocarbons to use as feedstock. Reengineered *C. autoethanogenum*, which is already capable of producing ethanol via fermentation, can redirect ethanol production to that of acetone and isopropanol [71].

3.3.2 Materials

Microorganisms are increasingly being used to make a variety of raw materials, such as plastics, fabric, building materials, coatings, etc. Plastics are one of the more sought-after materials for bioproduction, owing to the enormous ecological costs associated with conventional petroleum-based, non-biodegradable plastics ubiquitous in human society. Bioplastics have several advantages over conventional plastics: their production is more eco-friendly, they can be degraded by the environment over time, they can be biocompatible (non-toxic), and they can often use existing organic or industrial wastes as feedstocks [72]. One of the most well-known forms of bioplastics is

polyhydroxyalkanoate (PHA), which is made by more than 300 species of prokaryotes [73] in the form of granules within their cells. the range of sizes (chain length) of PHAs and the granules' macromolecular structures allow a diverse set of PHA-derived oligomers to be created. PHAs have similar chemical properties to conventional plastics and are easily biodegraded into simpler carbon compounds by certain bacteria. However, there are still significant costs associated with making PHAs in bulk, due to their extraction and purification costs.

More recently, polylactates have emerged as another promising bioplastic, similar in properties to PHAs, but with a simpler synthetic pathway. Polylactates are derived from – and broken down into – lactic acid, which is a common metabolite in most chemoautotrophs. Several engineered microbes have been developed that can make PLAs, including *E. coli* [74] and cyanobacteria [75]. Additionally, these GEMs can use plant biomass and industrial waste gasses as feedstock, as well as be grown in large quantities, which can help offset many of the costs associated with making bioplastics. One company, Neste, uses biomass from waste feedstocks and oil byproducts to make a variety of bioplastics [76].

Another biomaterial that has been well-studied and is now being tested in the field is biocement, which is cement derived from microorganisms by precipitation of calcium compounds [77,78]. These materials are aimed at replacing conventional Portland cement manufacturing and usage. The manufacturing process for Portland cement is quite carbon-heavy, as it requires enormous amounts of fossil fuel combustion to power kilns used to bake the cement precursors, and is responsible for 8% of global CO_2 emissions [79]. Furthermore, building construction and operation in general contributes 50% of emissions worldwide [78], so there is plenty of space for improvement. By contrast, biocement can potentially be carbon-negative, since it sequesters CO_2 from the air and groundwater to make cementitious materials. Additionally, cement sequesters CO_2 throughout its lifetime [80], reducing CO_2 emissions during production can result in a large reduction in overall carbon footprint. The process of biocementation can also be done *in situ*, which involves inoculation and culturing microbial colonies on soil that needs to be strengthened. In addition to being released into the environment, biocement can itself be composed of dried microbial biomass, so there are regulatory concerns by the EPA. bioMason, a bioconstruction company that utilizes soil microorganisms to produce calcium carbonate as a cementitious material which can then be used to build biologically-inert products like tiles, walls, foundations, etc. In our workshop, Kent Smith, Director of Research and Development at bioMason, stated that government regulations are a key factor in deployment of their product, especially as they develop higher-performing GEMs in the future.

Fuel or material manufactured by engineered microorganisms are not themselves regulated as bioengineered. And, since they are not agricultural or food products, they are mostly exempt from USDA and FDA regulations. The proprietary GEMs, however, make the companies subject to regulations under the EPA Office of Pollution Prevention and Toxics (EPA-OPPT). They must undergo a risk assessment by the OPPT to ensure that the GEMs are not producing toxins, are not pathogenic to plants and animals, and

are safeguarded against accidental release. For obvious environmental safety reasons, the EPA evaluates GEM products designed to be released to the outside world [81].

3.3.3 Bioremediation

GEMs also have a role to play in bioremediation from degrading hydrocarbons, pesticides, plastics and heavy metals. Besides degrading the usual organic matter and returning nutrients to the soil, GEMs can break down xenobiotic compounds as well [82]. Several microorganisms have shown broad efficacy against toxic pesticides like profenofos [83], pyrethroids [84], and endosulfan [85]. They have also been successful in treating soils contaminated with toxic hydrocarbons [86,87] and heavy metals [88]. In the case of plastics, diverse microbial communities have been shown to be most effective for degrading the wide range of polymers present in environmental plastic waste [89,90].

While the success of wild-type microbial isolates is promising, it also opens new possibilities for bioremediation using GEMs. The range of chemical contaminants that are biodegradable could be vastly improved if existing microorganisms could be purpose-built for certain compounds like toluene, xylene, and salicylate, that would normally be toxic to endogenous microorganisms [33,91]. Modern technologies like directed evolution can accelerate the search for engineered bacteria that are capable of novel chemistries for specific compounds, as was performed for the biodegradation of the pesticide atrazine [92].

As is the case with bioconstruction, bioremediation also involves environmental release of GEMs and GEM communities, and as such is regulated by the EPA. Arguably, however, bioremediation is more closely in line with the EPA's core mission of protecting and reducing harm to the environment.

3.4 Regulation of GEMs: Present and Future

3.4.1 Current regulations

Regulatory oversight of genetically altered organisms in the United States involves multiple government agencies; specifically, the US Department of Agriculture (USDA), Food and Drug Administration (FDA), and Environmental Protection Agency (EPA). Although the Coordinated Framework for Regulation of Biotechnology, instituted in 1986, sought to simplify and update oversight of biotechnology in food and agriculture, different governing bodies still maintain distinct but overlapping roles in regulating biotechnology and products thereof. The main tenets of the framework are as follows: 1) regulations should focus on the products of biotechnology rather than the process used to make them, 2) regulations should be grounded in scientific fact and reasoning, and 3) existing federal regulations are sufficient to review biotechnology products. Since its initiation, these tenets have held firm and lay the foundation for the US response to regulating biotechnology. The framework has undergone several updates since its

initiation, notably in 1992 and again more recently in 2017 [28]. The 2017 update was intended to clarify the roles and specific areas of focus for each agency. Fig. 3.1 outlines a simplified process for determining which agency will regulate a particular GEM intended for the market.

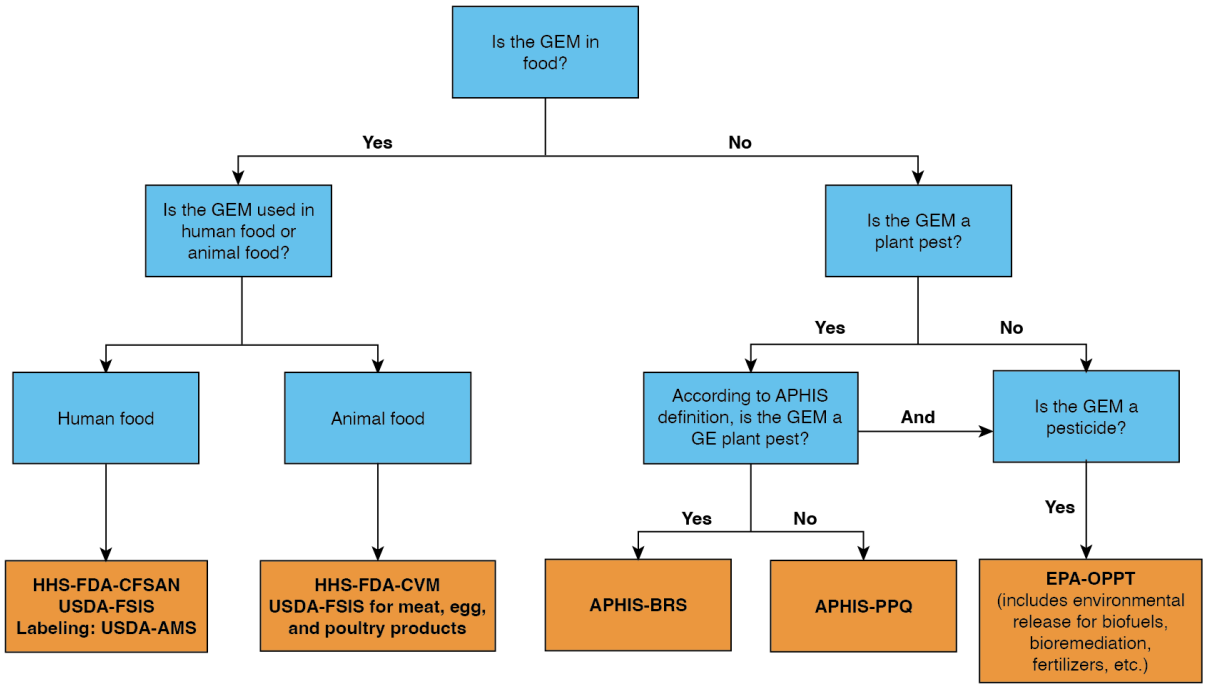


Figure 3.1*:* Flowchart outlining the steps required to determine which U.S. agency offices are responsible for regulating a GEM.

In addition to reaffirming the governing tenets described previously, one of the aims of the 2017 update was to revamp the framework such that emerging technologies and products could be more easily reviewed and regulated under existing guidelines, in an effort to modernize and "future-proof" the US government response to biotechnology. A unified website for the Coordinated Framework was established after the 2017 update [27].

The third tenet, which states that new products derived from biotechnology exist on a spectrum with all other products and can therefore fall under the purview of pre-existing federal guidelines, circumvents the need for specific rules for products made with recombinant DNA or genetic modifications.

In this section, we will attempt to broadly describe the responsibilities and purviews of each of the federal agencies charged with regulating biotechnology in the U.S.The USDA Animal and Plant Health Inspection Service (APHIS) is the branch responsible for regulating plants and plant pests, including DNA inserted from any plant pest. Under APHIS, a GEM can be regulated if it is classified taxonomically as a plant pest, contains DNA from a plant pest, or if it can be used to biologically control plant

pests. In these cases, APHIS regulates the interstate movement and importation, as well as potential environmental release of the GEM, and any company that is planning to do so must obtain a permit from APHIS.

There are multiple avenues employed by the FDA for regulation of GEMs, depending on the product, and as such use multiple definitions. For instance, the Food Safety and Inspection Service (FSIS) takes a final product-based outlook on any food or food additive that may use GEMs. FSIS reviews new technologies used in food production irrespective of whether or not a GEM is used. The FDA Center for Food Safety and Applied Nutrition (CFSAN) ensures safety of foods and food additives and undertakes a review process to present the “Generally Regarded As Safe” (GRAS) label to products. If a GEM is present in the final product, the permit application must provide details about the genetic lineage, modifications, and pathogenic safety of the organism. FDA-CVM process described here.

The EPA also has multiple offices responsible for the regulation of GEMs in scenarios where environmental release may be a concern. The Biopesticides and Pollution Prevention Division (BPPD), for example, oversees biologically based pesticides, which includes chemicals derived from GEMs or GEMs themselves (e.g. bacteriophages and fungicides). The Office of Pollution Prevention and Toxics (OPPT) is responsible for evaluating new chemicals or pathogens that may be released into the environment, and usually oversees GEMs used in biomanufacturing, fermentation, and biofuels industries.

Companies that market food products made by or containing GEMs are also required to abide by the National Bioengineered Food Disclosure Standard (the Standard). According to the Standard, a bioengineered food is one that contains genetic material that has been modified through recombinant DNA techniques, and that cannot be obtained through conventional breeding or found in nature.

3.4.2 Future-proofing regulations

In April 2012, the White House under the Obama administration released the National Bioeconomy Blueprint, which laid out several long-term goals for U.S. investment into biotechnology for the 21st century[93]. Even then, the U.S. government identified several trends regarding biotechnology in the scientific, commercial, and public opinion spheres. In the areas of energy, agriculture, and environment, climate change was identified as a potential crisis as they all relied on limited and outdated resources for a growing national and global population.

In 2022, ten years later, the Biden administration announced the National Biotechnology and Biomanufacturing Initiative (NBBI), an executive order to adopt a "whole-of-government approach to advance biotechnology and biomanufacturing towards innovative solutions in health, climate change, energy, food security, agriculture, supply chain resilience, and national and economic security" [94]. This initiative promises $2 billion in investments into research and development, market expansion, training, data sharing, and modernization of the U.S. bioindustry. Overall, this initiative represents a significant commitment by the U.S. government to bring its agencies and offices up-to-speed and in step with the private biotechnology sector, which has been moving at a rapid pace with emerging technologies.

Many of the issues the NBBI targets are familiar, having been specified in the 2012 Blueprint, and are unfortunately yet to be realized. Importantly, and relevant to this review, modernizing and streamlining the U.S. regulatory framework for biotechnology products has been a longstanding goal for the government, and a missed opportunity for developers.

What is evident from the Future of GEMs Workshop and examining the US regulatory system is just how complicated the regulatory landscape is. There are three agencies, six offices within those agencies and up to 12 different laws, regulations, and guidelines that a developer of a GEM may need to consult. This complexity may lead to a slow response when new technologies arrive – for instance one agency may be able to quickly produce guidance documents in relation to a new technology while others may take additional time.

One notable example is the American Chestnut Restoration Project, which aims to produce a blight-tolerant chestnut tree for restoration of American forests [95]. Since 2020, the USDA has been evaluating the Darling 58 chestnut tree for large-scale planting and propagation of the blight-resistant oxalate oxidase gene (OxO; isolated from wheat). If approved, this would the first time a bioengineered product was approved for ecological restoration. Supporters of the project argue that ecological restoration should be considered under the same regulatory scope as human health (as undertaken by the FDA), and that such projects are actually critical to broader ecosystem health that is universally beneficial [96]. However, the current U.S. system does not allow such a consideration. It is to be expected that coordinating a response among multiple agencies will take longer than just one office acting. One approach to streamlining this system could be the creation of a cross-organization panel or review body specifically designed to periodically review new innovations in the pipeline and coordinate agency responses.

In other countries, having just one biotechnology agency has also proven to be a workable approach; e.g. Kenya, and Argentina [97,98]. In Argentina, for example, the regulatory process for a new GE product involves three sequential steps: i) an environmental safety evaluation, ii) a food and feed safety evaluation, and iii) an evaluation of the product's impact on the market. Each step requires a favorable assessment to proceed to the next one, and the process is the same for GE plants, microorganisms, or animals [99]. While the myriad of laws in the U.S. are likely to remain in the future, having a centralized agency to interpret those laws and/or only one application to submit for developers could go a long way in streamlining the process and leading to faster, more decisive decisions on how new technologies will be treated in the future.

In addition to the complexities of the approvals process, another major issue is obsolescence of current regulation. Despite incremental updates, the Coordinated Framework as well intra-agency rules are 20+ years old, and not made for the pace of the current bioindustry. Many new products don't fall into the application spaces which the agencies have historically established; as an example, some new applications for GEMs may focus solely on sustainability endpoints or may have the goal of being deployed in natural environments for climate applications. As molecular biotechnology continues to advance, our ability to make genetic alterations become both more precise and more ambitious, i.e., bigger changes are now easier to make. Current definitions of genetically modified organisms are not on the same continuum: they are inherently categorical and do not accurately represent the nature of the modifications made, and how the end-product is "different" from the starting wild-type organism.

It is important to note that regulations are often slow to adapt by design, to allow innovation within the broad space provided and avoid different standards for different products. However, the landscape has changed such that innovation is no longer merely a business edge: climate change, overpopulation, food scarcity, ecosystem collapse, pandemic threats, etc. necessitate innovation for survival. Regulations need to be faster, more dynamic, more adaptable, and leaner (less bureaucratic). Additionally, regulations should also be a driver for innovation in certain areas, meaning the regulations can carve out a particular space for innovation for the public good, and incentivize large-scale problem-solving.

Conversely, product development should also happen with an eye towards regulations. Product development should be incentivized for problems such as climate, environment, food, etc. such that it makes economic sense to make products for public benefit, as opposed to potentially more profitable but niche products such as tastier foods, precision medicine, cosmetics, etc.

3.5 Conclusions

In 1975, a group of scientists (including Nobel Laureates David Baltimore and Paul Berg), lawyers and journalists convened in Asilomar, California to discuss the at-the-time nascent technology of recombinant DNA. A few years prior, Paul Berg had conducted an experiment to isolate gene fragments from the SV40 virus and insert it into an *E. coli* bacterium. This was the first recorded incidence of creating a genetically altered organism. The ethical and biosecurity implications of this experiment led to the Asilomar retreat, where the scientists present called for a moratorium on certain recombinant DNA experiments in the laboratory, which included (but was not limited to) cloning genes from pathogenic organisms, and recommended biocontainment procedures [100]. The discussion and decisions were made public and presented to the government as an assurance that the technology was in safe hands. The meeting allayed much public anxiety about recombinant DNA; however, it set a precedent for regulation of biotechnology for years to come. The Asilomar meeting essentially placed the responsibility – and privilege – of safeguarding the public from risky biological experiments on scientists in the laboratory. Top-down government regulation of research was rejected as a feasible option, and researchers and developers were given the burden to ensure their work was safe and for the greater good – a system that remains in effect to this day [101].

One may argue that this is an undemocratic process by which the public eye is removed from what is essentially a service performed by scientists. In some cases, it may foment distrust of science, backlash against new technologies, and potentially dangerous consequences. However, the fact of the matter is that this hands-off, "limited" regulation of laboratory research is what has built the current biotechnology industry. Innovation and creativity have fostered incredible new developments in the fields of medicine, agriculture, energy, and overall quality of life. New technologies have not just been created by for-profit companies; public universities and laboratories have also thrived in the freedom to pursue breakthroughs without the heavy hand of regulation. This view is also the official position of the U.S. regulatory agencies, as enshrined in the Coordinated Framework, as they place the burden of proving efficacy of any product going to market on the stakeholders themselves.

Conversely, however, the inherently capitalistic nature of this position has also allowed private stakeholders free reign to invest money and resources on the most lucrative of biotechnology products in favor of most beneficial. Enormous amounts of money are dedicated to industries like personalized medicine, anti-aging treatments, pet health, novelty foods, etc. "Boutique" biotech companies cater to a niche market of wealthy individuals. Furthermore, companies retain exclusive intellectual property and manufacturing rights, which perpetuate lack of access to beneficial technologies. The COVID-19 mRNA vaccine is one notable example [102,103]. Critics of deregulation have argued that regulatory strategies that are overly favorable for private, for-profit companies are problematic, as they can either compromise on public safety, or limit access [104]. Instead, non-profits and advocacy groups that regularly interact with

government policymakers, the public, and academia should be included in the dialogue for what effective governance should be.

Times have changed, and taking a back seat is no longer an option for anyone, especially those bodies charged with protecting the public good. Society is on the brink of several simultaneous crises that will require an all-hands-on-deck approach to mitigate. Both stakeholders and regulators need to work together to come up with large-scale solutions to climate change. Regulators, for their part, can work to expand the innovation space for products that directly benefit the public (e.g., high-yield crops, biofuels, biomaterials), incentivizing development in those directions by easing and simplifying regulatory burdens and setting useful boundary-conditions for development. The answer is not to limit innovation, even in niche markets – boutique biotech companies also advance scientific knowledge and serve the bioeconomy. Rather, regulations should be more proactive in how they mold the development space. Simultaneously, companies should redirect resources to public goods that have laxer regulatory burden and can therefore move to market easier and faster.

In 1974, when Paul Berg was poised to make a laboratory *E. coli* strain with fragments of SV40 inserted, he was dissuaded by colleagues due to the potential implications of opening such a door [100]. He and his colleagues realized the potential of the technology and decided that any progress must be made deliberately, carefully, and collaboratively. Nearly fifty years later, the landscape of recombinant DNA technology is astounding. We live in a world built on genetically engineered organisms and products and continue to advance the frontier in exponentially new ways. But for better or worse, a paradigm of biotechnology is upon us, with new problems and solutions. Time is running out, and public and private partnerships are more critical than ever to collaborate on scalable solutions.

3.6 Author Affiliations

Arik Shams[1], Melinda Kliegman[2], Alexandria Fischer[3], Anastasia Bodnar[3]

Affiliations:
[1]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[2]Innovative Genomics Institute, University of California, Berkeley, Berkeley, CA 94720, USA
[3]United States Department of Agriculture, Washington, DC, USA

Disclaimer:

The content of this chapter reflects the opinions of the authors, and this chapter is not intended to constitute a statement of the official policy or actions of the U.S. Department of Agriculture.

3.7 References

1. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
2. Jiang, F. & Doudna, J. A. CRISPR-Cas9 Structures and Mechanisms. *Annu. Rev. Biophys.* **46**, 505–529 (2017).
3. Zhang, H. *et al.* Application of the CRISPR/Cas9-based gene editing technique in basic research, diagnosis, and therapy of cancer. *Mol. Cancer* **20**, 126 (2021).
4. Doudna, J. A. The promise and challenge of therapeutic genome editing. *Nature* **578**, 229–236 (2020).
5. Singh, V. Chapter 1 - An introduction to genome editing CRISPR-Cas systems. in *Genome Engineering via CRISPR-Cas9 System* (eds. Singh, V. & Dhar, P. K.) 1–13 (Academic Press, 2020). doi:10.1016/B978-0-12-818140-9.00001-5.
6. Nakamura, M., Gao, Y., Dominguez, A. A. & Qi, L. S. CRISPR technologies for precise epigenome editing. *Nat. Cell Biol.* **23**, 11–22 (2021).
7. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
8. Scholefield, J. & Harrison, P. T. Prime editing - an update on the field. *Gene Ther.* **28**, 396–401 (2021).
9. Nishimasu, H. *et al.* Crystal Structure of Staphylococcus aureus Cas9. *Cell* **162**, 1113–1126 (2015).
10. Acharya, S. *et al. Francisella novicida* Cas9 interrogates genomic DNA with very high specificity and can be used for mammalian genome editing. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 20959–20968 (2019).
11. Hou, Z. *et al.* Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15644–15649 (2013).
12. Harrington, L. B. *et al.* A thermostable Cas9 with increased lifetime in human plasma. *Nat. Commun.* **8**, 1424 (2017).
13. Wei, T., Cheng, Q., Min, Y.-L., Olson, E. N. & Siegwart, D. J. Systemic nanoparticle delivery of CRISPR-Cas9 ribonucleoproteins for effective tissue specific genome editing. *Nat. Commun.* **11**, 3232 (2020).
14. Wang, D., Zhang, F. & Gao, G. CRISPR-Based Therapeutic Genome Editing: Strategies and In Vivo Delivery by AAV Vectors. *Cell* **181**, 136–150 (2020).
15. Yourik, P., Fuchs, R. T., Mabuchi, M., Curcuru, J. L. & Robb, G. B. Staphylococcus aureus Cas9 is a multiple-turnover enzyme. *RNA* **25**, 35–44 (2019).
16. Ma, D., Peng, S., Huang, W., Cai, Z. & Xie, Z. Rational Design of Mini-Cas9 for Transcriptional Activation. *ACS Synth. Biol.* **7**, 978–985 (2018).
17. Simm, A. M., Baldwin, A. J., Busse, K. & Jones, D. D. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 beta-lactamase. *FEBS Lett.* **581**, 3904–3908 (2007).
18. Morelli, A., Cabezas, Y., Mills, L. J. & Seelig, B. Extensive libraries of gene truncation variants generated by in vitro transposition. *Nucleic Acids Res.* **45**, e78 (2017).

19. Arpino, J. A. J., Reddington, S. C., Halliwell, L. M., Rizkallah, P. J. & Jones, D. D. Random single amino acid deletion sampling unveils structural tolerance and the benefits of helical registry shift on GFP folding and structure. *Structure* **22**, 889–898 (2014).
20. Teichmann, S. A., Park, J. & Chothia, C. Structural assignments to the Mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14658–14663 (1998).
21. Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. Evolution of the protein repertoire. *Science* **300**, 1701–1703 (2003).
22. Shapiro, R. S., Chavez, A. & Collins, J. J. CRISPR-based genomic tools for the manipulation of genetically intractable microorganisms. *Nat. Rev. Microbiol.* **16**, 333–339 (2018).
23. Harbola, A., Negi, D., Manchanda, M. & Kesharwani, R. K. Chapter 27 - Bioinformatics and biological data mining. in *Bioinformatics* (eds. Singh, D. B. & Pathak, R. K.) 457–471 (Academic Press, 2022). doi:10.1016/B978-0-323-89775-4.00019-5.
24. Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J. Clin. Invest.* **132**, (2022).
25. Song, L.-F., Deng, Z.-H., Gong, Z.-Y., Li, L.-L. & Li, B.-Z. Large-Scale de novo Oligonucleotide Synthesis for Whole-Genome Synthesis and Data Storage: Challenges and Opportunities. *Front Bioeng Biotechnol* **9**, 689797 (2021).
26. Adrio, J. L. & Demain, A. L. Genetic improvement of processes yielding microbial products. *FEMS Microbiol. Rev.* **30**, 187–214 (2006).
27. The Unified Website for Biotechnology Regulation. https://usbiotechnologyregulation.mrp.usda.gov/biotechnologygov/resources.
28. Modernizing the Regulatory System for Biotechnology Products: Final Version of the 2017 Update to the Coordinated Framework for the Regulation of Biotechnology. Preprint at https://usbiotechnologyregulation.mrp.usda.gov/2017_coordinated_framework_update.pdf (2017).
29. Karavolias, N. G., Horner, W., Abugu, M. N. & Evanega, S. N. Application of Gene Editing for Climate Change in Agriculture. *Frontiers in Sustainable Food Systems* **5**, (2021).
30. Wen, A. *et al.* Enabling Biological Nitrogen Fixation for Cereal Crops in Fertilized Fields. *ACS Synth. Biol.* **10**, 3264–3277 (2021).
31. Hatfield, J. L. & Dold, C. Water-Use Efficiency: Advances and Challenges in a Changing Climate. *Front. Plant Sci.* **10**, 103 (2019).
32. “The Future of Microbial Biotechnology” Workshop Summary - Innovative Genomics Institute (IGI). *Innovative Genomics Institute (IGI)* https://innovativegenomics.org/the-future-of-microbial-biotechnology-workshop-summary/ (2022).
33. Pant, G. *et al.* Biological approaches practised using genetically engineered microbes for a sustainable environment: A review. *J. Hazard. Mater.* **405**, 124631 (2021).

34. Roell, M.-S. & Zurbriggen, M. D. The impact of synthetic biology for future agriculture and nutrition. *Curr. Opin. Biotechnol.* **61**, 102–109 (2020).
35. Liao, J. C., Mi, L., Pontrelli, S. & Luo, S. Fuelling the future: microbial engineering for the production of sustainable biofuels. *Nat. Rev. Microbiol.* **14**, 288–304 (2016).
36. United States Department of Agriculture. Agricultural Biotechnology Glossary. https://www.usda.gov/topics/biotechnology/biotechnology-glossary.
37. Hall, R. J., Whelan, F. J., McInerney, J. O., Ou, Y. & Domingo-Sananes, M. R. Horizontal Gene Transfer as a Source of Conflict and Cooperation in Prokaryotes. *Front. Microbiol.* **11**, 1569 (2020).
38. Hilbeck, A., Meyer, H., Wynne, B. & Millstone, E. GMO regulations and their interpretation: how EFSA's guidance on risk assessments of GMOs is bound to fail. *Environmental Sciences Europe* **32**, 1–15 (2020).
39. Meena, S. S. & Mohanty, A. Ethical, Patent, and Regulatory Issues in Microbial Engineering. in *Engineering of Microbial Biosynthetic Pathways* (eds. Singh, V., Singh, A. K., Bhargava, P., Joshi, M. & Joshi, C. G.) 133–142 (Springer Singapore, 2020). doi:10.1007/978-981-15-2604-6_8.
40. Jacoby, R., Peukert, M., Succurro, A., Koprivova, A. & Kopriva, S. The Role of Soil Microorganisms in Plant Mineral Nutrition-Current Knowledge and Future Directions. *Front. Plant Sci.* **8**, 1617 (2017).
41. Abedini, D., Jaupitre, S., Bouwmeester, H. & Dong, L. Metabolic interactions in beneficial microbe recruitment by plants. *Curr. Opin. Biotechnol.* **70**, 241–247 (2021).
42. Wang, Q., Liu, J. & Zhu, H. Genetic and Molecular Mechanisms Underlying Symbiotic Specificity in Legume-Rhizobium Interactions. *Front. Plant Sci.* **9**, 313 (2018).
43. Rosenblueth, M. *et al.* Nitrogen Fixation in Cereals. *Front. Microbiol.* **9**, 1794 (2018).
44. Frink, C. R., Waggoner, P. E. & Ausubel, J. H. Nitrogen fertilizer: retrospect and prospect. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 1175–1180 (1999).
45. Coskun, D., Britto, D. T., Shi, W. & Kronzucker, H. J. Nitrogen transformations in modern agriculture and the role of biological nitrification inhibition. *Nat Plants* **3**, 17074 (2017).
46. Voigt, C. A. Synthetic biology 2020-2030: six commercially-available products that are changing our world. *Nat. Commun.* **11**, 6379 (2020).
47. Temme, K., Tamsir, A., Bloch, S., Clark, R. & Tung, E. Methods and compositions for improving plant traits. *US Patent* (2021).
48. Haro, R. & Benito, B. The Role of Soil Fungi in K+ Plant Nutrition. *Int. J. Mol. Sci.* **20**, (2019).
49. Jain, D. *et al.* Chapter 11 - Potassium solubilizing microorganisms as soil health engineers: An insight into molecular mechanism. in *Rhizosphere Engineering* (eds. Dubey, R. C. & Kumar, P.) 199–214 (Academic Press, 2022). doi:10.1016/B978-0-323-89973-4.00007-7.
50. Syed Ab Rahman, S. F., Singh, E., Pieterse, C. M. J. & Schenk, P. M. Emerging microbial biocontrol strategies for plant pathogens. *Plant Sci.* **267**, 102–111 (2018).
51. Razaq, M. & Shah, F. M. Chapter 2 - Biopesticides for management of arthropod pests and weeds. in *Biopesticides* (eds. Rakshit, A. et al.) 7–18 (Woodhead Publishing, 2022). doi:10.1016/B978-0-12-823355-9.00005-5.

52. Liu, X. *et al.* Overview of mechanisms and uses of biopesticides. *Int. J. Pest Manage.* **67**, 65–72 (2021).
53. Results of the 2nd cereal trial campaign. *Amoeba* https://amoeba-nature.com/en/results-of-the-2nd-cereal-trial-campaign/ (2021).
54. Gupta, R. *et al.* The Entomopathogenic Fungi Metarhizium brunneum and Beauveria bassiana Promote Systemic Immunity and Confer Resistance to a Broad Range of Pests and Pathogens in Tomato. *Phytopathology* **112**, 784–793 (2022).
55. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
56. Plavec, T. V. & Berlec, A. Safety Aspects of Genetically Modified Lactic Acid Bacteria. *Microorganisms* **8**, (2020).
57. Ingredient Series: Gene-Edited Beer Yeast. *Omega Yeast* https://omegayeast.com/news/ingredient-series-gene-edited-beer-yeast.
58. How we Bioengineer Yeast to have Improved Traits. *Berkeley Yeast* https://berkeleyyeast.com/science-lesson/how-we-bioengineer-yeast-to-have-improved-traits/ (2022).
59. Land use in agriculture by the numbers. *Food and Agriculture Organization of the United Nations* https://www.fao.org/sustainability/news/detail/en/c/1274219/.
60. Gleizer, S., Bar-On, Y. M., Ben-Nissan, R. & Milo, R. Engineering Microbes to Produce Fuel, Commodities, and Food from CO2. *Cell Reports Physical Science* **1**, 100223 (2020).
61. Microbiología, M., Campos Muñiz, C., Cuadra Zelaya, T. E., Rodríguez Esquivel, G. & Fernández, F. J. Penicillin and cephalosporin production: A historical perspective. *Microbiología* **49**, 3–4 (2007).
62. Wu, J. *et al.* Toward improvement of erythromycin A production in an industrial Saccharopolyspora erythraea strain via facilitation of genetic manipulation with an artificial attB site for specific recombination. *Appl. Environ. Microbiol.* **77**, 7508–7516 (2011).
63. Baltz, R. H. Molecular engineering approaches to peptide, polyketide and other antibiotics. *Nat. Biotechnol.* **24**, 1533–1540 (2006).
64. Smith, J. L. & Palumbo, S. A. Microorganisms as Food Additives. *J. Food Prot.* **44**, 936–955 (1981).
65. Baeshen, N. A. *et al.* Cell factories for insulin production. *Microb. Cell Fact.* **13**, 141 (2014).
66. Kayani, W. K., Kiani, B. H., Dilshad, E. & Mirza, B. Biotechnological approaches for artemisinin production in Artemisia. *World J. Microbiol. Biotechnol.* **34**, 54 (2018).
67. Scown, C. D. & Keasling, J. D. Sustainable manufacturing with synthetic biology. *Nat. Biotechnol.* **40**, 304–307 (2022).
68. Zhu, P. & Chen, X. Carbon-negative biomanufacturing of chemicals from waste gases. *Chem* **8**, 1178–1180 (2022).
69. Hermann, B. G., Blok, K. & Patel, M. K. Producing bio-based bulk chemicals using industrial biotechnology saves energy and combats climate change. *Environ. Sci. Technol.* **41**, 7915–7921 (2007).

70. Jullesson, D., David, F., Pfleger, B. & Nielsen, J. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnol. Adv.* **33**, 1395–1402 (2015).
71. Liew, F. E. *et al.* Carbon-negative production of acetone and isopropanol by gas fermentation at industrial pilot scale. *Nat. Biotechnol.* **40**, 335–344 (2022).
72. Luengo, J. M., García, B., Sandoval, A., Naharro, G. & Olivera, E. R. Bioplastics from microorganisms. *Curr. Opin. Microbiol.* **6**, 251–260 (2003).
73. Anjum, A. *et al.* Microbial production of polyhydroxyalkanoates (PHAs) and its copolymers: A review of recent advancements. *Int. J. Biol. Macromol.* **89**, 161–174 (2016).
74. Taguchi, S. *et al.* A microbial factory for lactate-based polyesters using a lactate-polymerizing enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 17323–17327 (2008).
75. Katayama, N., Iijima, H. & Osanai, T. Production of Bioplastic Compounds by Genetically Manipulated and Metabolic Engineered Cyanobacteria. *Adv. Exp. Med. Biol.* **1080**, 155–169 (2018).
76. Neste successfully concludes first series of processing trial runs at industrial scale with liquefied waste plastic in Porvoo, Finland. *Neste worldwide* https://www.neste.com/releases-and-news/circular-economy/neste-successfully-concludes-first-series-processing-trial-runs-industrial-scale-liquefied-waste (2022).
77. Stabnikov, V. & Ivanov, V. 3 - Biotechnological production of biopolymers and admixtures for eco-efficient construction materials. in *Biopolymers and Biotech Admixtures for Eco-Efficient Construction Materials* (eds. Pacheco-Torgal, F., Ivanov, V., Karak, N. & Jonkers, H.) 37–56 (Woodhead Publishing, 2016). doi:10.1016/B978-0-08-100214-8.00003-8.
78. Achal, V. & Mukherjee, A. A review of microbial precipitation for sustainable construction. *Construction and Building Materials* **93**, 1224–1235 (2015).
79. Olivier JGJ, Janssens-Maenhout G, Muntean M, Peters JAHW. *Trends in global CO2 emissions; 2016 Report, The Hague: PBL Netherlands Environmental Assessment Agency*. http://edgar.jrc.ec.europa.eu or http://www.pbl.nl/en/trends-in-global-co2-emissions (2016).
80. Xi, F. *et al.* Substantial global carbon uptake by cement carbonation. *Nat. Geosci.* **9**, 880–883 (2016).
81. TSCA biotechnology regulations, US EPA OCSPP. https://www.epa.gov/regulation-biotechnology-under-tsca-and-fifra/tsca-biotechnology-regulations (2015).
82. Rebello, S. *et al.* Bioengineered microbes for soil health restoration: present status and future. *Bioengineered* **12**, 12839–12853 (2021).
83. Malghani, S., Chatterjee, N., Yu, H. X. & Luo, Z. Isolation and identification of Profenofos degrading bacteria. *Braz. J. Microbiol.* **40**, 893–900 (2009).
84. Zhang, X. *et al.* Isolation and identification of the Raoultella ornithinolytica-ZK4 degrading pyrethroid pesticides within soil sediment from an abandoned pesticide plant. *Arch. Microbiol.* **201**, 1207–1217 (2019).
85. Bhalerao, T. S. & Puranik, P. R. Biodegradation of organochlorine pesticide, endosulfan, by a fungal soil isolate, Aspergillus niger. *Int. Biodeterior. Biodegradation* **59**, 315–321 (2007).

86. Kumari, S., Mangwani, N. & Das, S. Synergistic effect of quorum sensing genes in biofilm development and PAHs degradation by a marine bacterium. *Bioengineered* vol. 7 205–211 (2016).
87. Furukawa, K. “Super bugs” for bioremediation. *Trends Biotechnol.* **21**, 187–190 (2003).
88. Rebello, S. *et al.* Hazardous minerals mining: Challenges and solutions. *J. Hazard. Mater.* **402**, 123474 (2021).
89. Cf, S. F. *et al.* Bioprospecting of gut microflora for plastic biodegradation. *Bioengineered* **12**, 1040–1053 (2021).
90. Anani, O. A. & Adetunji, C. O. Bioremediation of Polythene and Plastics Using Beneficial Microorganisms. in *Microbial Rejuvenation of Polluted Environment: Volume 3* (eds. Adetunji, C. O., Panpatte, D. G. & Jhala, Y. K.) 281–302 (Springer Singapore, 2021). doi:10.1007/978-981-15-7459-7_13.
91. Kumar, N. M., Muthukumaran, C., Sharmila, G. & Gurunathan, B. Genetically Modified Organisms and Its Impact on the Enhancement of Bioremediation. in *Bioremediation: Applications for Environmental Protection and Management* (eds. Varjani, S. J., Agarwal, A. K., Gnansounou, E. & Gurunathan, B.) 53–76 (Springer Singapore, 2018). doi:10.1007/978-981-10-7485-1_4.
92. Renata, H., Wang, Z. J. & Arnold, F. H. Expanding the enzyme universe: accessing non-natural reactions by mechanism-guided directed evolution. *Angew. Chem. Int. Ed Engl.* **54**, 3351–3367 (2015).
93. The White House Office of Science and Technology Policy. *National Bioeconomy Blueprint*. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/national_bioeconomy_blueprint_april_2012.pdf (2012).
94. Executive Office of the President. Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy. *Federal Register* vol. 87 56849–56860 Preprint at https://www.federalregister.gov/d/2022-20167 (2022).
95. Westbrook, J. W., Holliday, J. A., Newhouse, A. E. & Powell, W. A. A plan to diversify a transgenic blight-tolerant American chestnut population using citizen science. *Plants People Planet* **2**, 84–95 (2020).
96. Aucott, M. & Parker, R. A. Medical biotechnology as a paradigm for forest restoration and introduction of the transgenic American chestnut. *Conserv. Biol.* **35**, 190–196 (2021).
97. Kenya National Biosafety Authority. BIosafety Clearing-House Kenya. *National Node of the BCH - Kenya* http://ke.biosafetyclearinghouse.net/ (2012).
98. Convocatoria CONABIA. *Argentina.gob.ar* https://www.argentina.gob.ar/agricultura/bioeconomia/biotecnologia/convocatoria-conabia (2018).
99. Lewi, D. M. & Vicién, C. Argentina’s Local Crop Biotechnology Developments: Why Have They Not Reached the Market Yet? *Front Bioeng Biotechnol* **8**, 301 (2020).
100. Berg, P. & Singer, M. F. The recombinant DNA controversy: twenty years later. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 9011–9013 (1995).

101. Hurlbut, J. B. Limits of Responsibility: Genome Editing, Asilomar, and the Politics of Deliberation. *Hastings Cent. Rep.* **45**, 11–14 (2015).
102. Time is running out for COVID vaccine patent waivers. *Nature Publishing Group UK* http://dx.doi.org/10.1038/d41586-022-00878-x (2022) doi:10.1038/d41586-022-00878-x.
103. Loft, P. *Waiving intellectual property rights for Covid-19 vaccines*. https://commonslibrary.parliament.uk/research-briefings/cbp-9417/ (2022).
104. Gordon, D. R. *et al.* Responsible governance of gene editing in agriculture and the environment. *Nat. Biotechnol.* **39**, 1055–1057 (2021).

Chapter 4

Conclusions

4.1 The When, How, and Who of CRISPR Technology

In the last decade, genome-editing has progressed from one protein, SpCas9, to many new orthologs, new CRISPR types and molecules, engineered variants, and chimeric effectors. The field has evolved from "molecular scissors," to an ever-expanding molecular toolbox for DNA and RNA. However, one could argue that there have been some caveats to the speed at which the field has progressed. Today, hundreds of new CRISPR variants exist, each slightly different from the next, often with single nucleotide variations and minute differences in performance [1]. The rate of discovery and engineering of new CRISPR-Cas variants has outpaced our ability to fully characterize each new variant that is being created and touted as the next game-changing molecule. Additionally, there is a symbiotic tension between engineering new variants from existing Cas proteins, versus mining for new variants in nature that may already have a desired function. Newly discovered molecules are engineered into new subvariants before the original is fully understood.

In many ways, this push for "new" molecules is driven by commercialization. CRISPR has also shifted paradigms in how technology is patented and profited from, in that many patents arising from CRISPR happen from within academic laboratories as opposed to existing research or biomedical companies [2]. Licenses are granted for individual molecules for broad applications, so making a slightly different variant and commercializing its use is relatively easy and incentivized. Arguably, the potential innovation space enabled by this system can be beneficial overall, since this leads to a vast number of options for various applications. However, this innovation space is rapidly being carved out by exclusive rights for any new CRISPR protein, without much consideration for how useful it actually is in practice. Even licensing complications aside, currently there is no literature or unified database for every Cas variant out there, and subsequently, no standard for best practices exists about which molecule is best for which function.

Conversely, there are certain benefits to patenting technology. Patents fundamentally restrict use of an invention by anyone other than the licensee. However, this also allows the licensor to restrict the use of the technology for certain applications, i.e., obligate ethical use of the technology by the licensee [3]. Such is the case for the Broad Institute's CRISPR license to Bayer (formerly Monsanto) [4]. Bayer can use the technology to edit plants, but not to develop gene drives, terminator seeds, or tobacco products. While this practice is ethically promising in theory, it does not guarantee such an outcome, nor answer the question of whether the application of extremely powerful technologies like CRISPR should be decided by the patent owner.

The many debates surrounding CRISPR can be overwhelming, especially when trying to decide on a practical and productive way forward. The use of CRISPR in modern society is no longer a matter of *if* the technology is going to be used, but rather *when* and *how*. The development of genome-editing technology, and subsequently the biotechnological revolution it has birthed converges on a critical time for our society and the planet. As climate crises loom over every aspect of human life, scientists and policymakers are scrambling for large-scale solutions. Genome-editing has tremendous

potential in this regard and is evermore necessary for mitigating the worst effects of climate change. In this context, the rapid implementation of CRISPR technology is paramount, and every effort must be made to facilitate access to the technology. Now more than ever, public-private partnerships are critical to create solutions at scale. At the same time, there must be a reckoning for how the technology must be used. Less discussed than the *when* and *how* of CRISPR implementation is the *who*, even though the answer seems simple: everybody. Without advocating for a total dissolution of CRISPR as intellectual property, we must proactively pursue the approach that benefits the most people – something that will require a whole-of-society effort.

Even as the scientific advances continue unabated, its implementation is far less rapid. Translating CRISPR from the laboratory to crop fields, refineries, and hospitals is a more complex problem with many variables such as local politics, diverse standards of ethics, resources, and accessibility. Exporting scientific knowledge and capacity often abuts against the invisible wall of policy. During the course of this PhD and this dissertation, I have attempted to understand the interactions between science research and science policy. My main research project has been a purely scientific project, to try to engineer – and consequently better understand – Cas9. In the process, I came to appreciate the scale of the implementation problem of CRISPR. In the second part, I participated in a workshop that brought together groups of GEM developers and regulators and reviewed the current state of the field from a policy perspective. My hope is that taken together, this document describes a complex and important relationship between science and policy, and one that emphasizes the need for better synergy between the two.

4.2 References

1. Kondrateva, E., Demchenko, A., Lavrov, A. & Smirnikhina, S. An overview of currently available molecular Cas-tools for precise genome modification. *Gene* **769**, 145225 (2021).
2. Brinegar, K. *et al.* The commercialization of genome-editing technologies. *Crit. Rev. Biotechnol.* **37**, 924–932 (2017).
3. Sherkow, J. S. Patent protection for CRISPR: an ELSI review. *J Law Biosci* **4**, 565–576 (2017).
4. Begley, S. Monsanto licenses CRISPR technology to modify crops — with key restrictions. *Stat News* (2016).