When good guides go bad: empirical evaluation of all unique cas9 targets in *E. coli* reveal widespread functionality and rules for gRNA biological design

Elise K. Phillips[1,2], Riley Harrison[1,2,3], S'Khaja Charles[1,2,4], Dawn Klingemann[1], Tim Wiser[5], Carrie A. Eckert[1,2], William G. Alexander[1,2,5,*]

[1]Oak Ridge National Laboratory, Oak Ridge, TN 37830, USA
[2]Center for Bioenergy Innovation, Oak Ridge National Laboratory, Oak Ridge, TN, 37830
[3]Brown University, Providence, RI 02912, USA
[4]University of Chicago, Chicago, IL 60637, USA
[5]Truman State University, Kirksville, MO 63501, USA

* To whom correspondence should be addressed. Tel: +1 (865) 341-1730 Email: alexanderwg@ornl.gov

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

ABSTRACT

The Cas9 nuclease has become central to modern methods and technologies in synthetic biology, largely due to the ease in which it can be targeted to specific DNA loci via guide RNAs (gRNAs). Reports vary widely on the actual specificity of this targeting, with some studies observing 60% of gRNAs possessing no activity against the genome, while there is a general assumption in the *E. coli* community that inactive gRNAs are rare. To resolve these contradictions, we evaluated the activity of nearly 500,000 unique gRNAs in the *E. coli* K12 MG1655 genome. We show that the overwhelming majority of unique gRNAs are functional (at least 93%) while only 0.3% are nonfunctional. These nonfunctional gRNAs tend to exhibit strong spacer self-interaction, leading to the development of a simple set of gRNA design rules for bacteria. Finally, this work provides the greater microbial synthetic biology community a set of nearly half a million sgRNA spacers that have been empirically evaluated *in vivo* which will expedite future biological engineering projects.

INTRODUCTION

A ubiquitous class of genome editing techniques use endonucleases to achieve precise, arbitrary, and targeted genetic changes[1–3]. While nucleases of various types have been used in genome editing in the past, the current gold standard for targeted and precise genome editing uses RNA-directed endonucleases derived from bacterial CRISPR systems, with the Cas9 protein from *Streptococcus pyogenes* being the most well-known and widely used[4,5]. These CRISPR-associated endonucleases are directed to their targets by the binding of an associated RNA molecule to a complementary target DNA molecule[6,7]; the DNA molecule is then cut via the action of the two nuclease domains present in Cas9. For Cas9 in its native context, two RNA molecules are needed: a CRISPR RNA (crRNA) and a trans-activating CRISPR RNA (tracrRNA)[8]. However, during its development as a genome editor, the Cas9 crRNA and tracrRNA were fused together with a tetraloop to form a synthetic guide RNA (sgRNA or gRNA), which simplified heterologous use of the system[4,5]. Hereafter we refer to the RNA molecules that bind to Cas nucleases to direct their targeting generally as guide RNAs or gRNAs.

Cas9 gRNAs are 96 bases in length and contain two functional regions: the first 20 nucleotides are called the spacer, and the remaining 76 nucleotides is the Cas9-interaction component called the scaffold[9]. The spacer itself contains two regions: the first six bases are referred to as the PAM-Distal region (or merely “distal” for short), and the 14 nucleotides on the 3’ end of the spacer are referred to as the PAM-Proximal region (more frequently referred to as the “seed”)[10]. Mismatches in the seed are poorly tolerated by Cas9, resulting in reduced or abolished nuclease activity, while mutations in the six distal bases can convey little to no reduction in activity[11]. Additionally, the Cas9-gRNA complex will preferentially bind and act on spacer-complementary DNA that is adjacent to a short sequence specific to that nuclease. This sequence is termed the Protospacer Adjacent Motif (PAM), and for *S. pyogenes* Cas9,

this sequence is NGG and must be directly adjacent to the last base of the spacer target[12]. However, RNA-directed nucleases frequently recognize slightly altered PAM sequences, or "non-canonical" PAMs, which may serve as a mechanism to defend against selfish genetic elements evolving to eliminate PAM recognition sites from their genome. *S. pyogenes* Cas9 recognizes NAG and NGA as non-canonical PAM sequences and bind to the proceeding spacer with reduced effectiveness[13,14].

Despite the wide use of this nuclease as a tool for over a decade, contradiction and confusion exists in the literature and the community regarding the design of gRNAs[15], even in bacterial systems. For instance, up to 60% of chosen gRNAs in *E. coli* have been reported as non-functional *in vivo*[16], but within the *E. coli* community, the "lab lore" (the scientific information passed on culturally within research communities, usually verbally, and rarely published) asserts that most gRNAs simply work. Numerous computational models have been developed to predict target suitability, but these models do not apply well outside of the specific contexts in which they were developed[17].

Strangely, very few genome-scale libraries of gRNAs have been made in *E. coli*. The largest genome-scale gRNA library reported is a 120,000-member library we designed and tested[17]; despite the large relative size of this library, we still could not determine precise criteria for discriminating poorly performing gRNA targets from highly functional ones. Here, we designed and tested the remaining 405,000 unique gRNAs in the *E. coli* genome, resulting in comprehensive cutting data for 463,189 unique spacers, or 88% of the total present in the genome. We found that almost all (at least 93%) of unique spacers are highly active in *E. coli*, with only 0.3% of spacers confirmed as being inactive. We also demonstrate the role played by spacer self-interaction in gRNA inactivation and provide a method for potential gRNA reactivation. Finally, we develop a simple design rule to avoid the bulk of inactive gRNAs.

MATERIAL AND METHODS

Genome-scale gRNA Library Design, Construction, Sequencing, and Analysis

All oligonucleotide handling, plasmid library construction, host library transformation, selection, and analysis were conducted previously for the 120,000-member library[17]. The two 202,298 member libraries were produced in identical methods with the exception that the library amplicons were combined equimolarly prior to cloning, thus producing a single 404,596 library. The sequence and library name of all 524,596 gRNA spacers used in this work are found in Supplementary Data 1.

Validation gRNA Library Design and Construction

To validate the results of the larger genomic library, two hundred designs classified as functional, nonfunctional, or statistically insignificant were chosen randomly and were synthesized by Genscript in a mixed oligonucleotide pool. These oligonucleotides were then used to make plasmid libraries in the same manner as the genome-wide libraries and analysed by an identical workflow. This process was repeated, this time synthesizing and cloning the 2,365 nonfunctional guides from the genome-wide library; to further increase accuracy of functional calls, a different set of phased PCR primers was used to amplify the entire gRNA and promoter sequence from the library for sequencing (Supplementary Table 1).

Spacer self-interaction prediction

ViennaRNA[18] was used to predicted maximum free energy (MFE) and secondary structure for all functional (~444k) and nonfunctional (1,651) spacers. These data were then processed into bins and visualized.

RESULTs

Massively parallel depletion/enrichment analysis elucidates gRNA activity

Dissatisfied with the lack of reliable data on gRNA functionality *in vivo*, we decided to simply test all potential unique spacers available in the *E. coli*. A total of 524,596 unique spacers were found using a bespoke Python script; as the maximum library size available from our oligonucleotide vendor Agilent was 244,000 designs per library, meaning we'd need at least three separate oligonucleotide pools to produce the totality of gRNAs from the genome. In collaboration[17], we decided to adjust our totals to produce a 120,000-member pool and two 202,298-member pools. Results reported here are a summation of all results from all three libraries.

A total of 471,530 designs, or 90%, had at least 20 reads in both sequencing replicates. After transformation into the host strain of *E. coli*, 463,189 spacers exhibited a statistically significant change in abundance from the donor library (α=0.05). These changes in abundance were -log2-transformed to better aid in comparison and visualization, with more functional gRNAs having a higher positive score and nonfunctional gRNAs having a lower negative score. Of those gRNAs with significant changes in abundance, 438,495 (93%) of the spacers were graded as functional, as they exhibited a sizeable decrease in abundance (A-score >= 2) when transformed into their host. Additionally, 2,344 (0.5%) of the spacers demonstrated nonfunctionality (A-score <= -2) as without significant Cas9 activity these more cells with these guides survive, thus increasing their abundance. The remaining 22,350 (6.5%) of spacers were ambiguous, exhibiting statistically significant differences in their abundance despite said differences being relatively small, with A-scores between -2 and 2 (Supplementary Data 2). Surprisingly, the vast majority of unique spacers in *E. coli* are highly effective and functional.

To provide a coarse, genome-scale analysis of this activity data, we divided the *E. coli* genome into 2.5 kbp bins starting at the first base (*e.g.* 1-2500, 2501-5000, etc.). For each bin, the number of unique gRNA targets (the spacer and PAM sequence on the genome that is acted upon by the gRNA:Cas9 ribonucleoprotein complex), the number of functional gRNA targets, and the number of nonfunctional targets were totaled and plotted, along with other common genomic features, into a circos plot (Figure 1). The ratio of functional to nonfunctional targets largely remained the same throughout the genome, though the total number of gRNAs did vary as a function of GC content in a given region. The sole exception to this trend is a region between 280 and 290 kbp which is enriched for both nonfunctional and total number of targets (Supplementary Note).

Validating massively parallel data with smaller, more accurate pools

To validate results from the pooled library, we generated smaller libraries to ensure results are repeatable and accurate. We chose two hundred gRNAs each that were classified as functional or nonfunctional along with two hundred gRNAs that were statistically insignificant in their abundance changes. These lattermost designs were included as they were somewhat paradoxical: a functional gRNA is lethal and exits the population while a nonfunctional gRNA increases in abundance as it behaves like an empty vector, so a design showing no significant change of abundance is contradictory. When we transformed this library into our host strain, and almost all (99%) of the functional gRNAs replicated their functionality (Figure 2, Supplementary Data 3). Surprisingly, 170 statistically insignificant designs (85%) and 150 nonfunctional designs (75%) demonstrated functionality in this library, which suggests that the genome-wide library functionality call is conservative in nature.

Confirming poor guide activity via improved detection methods

The results from this smaller Confirmation library largely corroborate those from the genome-wide library: almost all unique gRNA targets in the *E. coli* genome result in cutting by Cas9, in direct contradiction to previous reports[16]. As the number of inactive gRNAs were quite small (2,344), we next synthesized these gRNAs for further examination, forming the

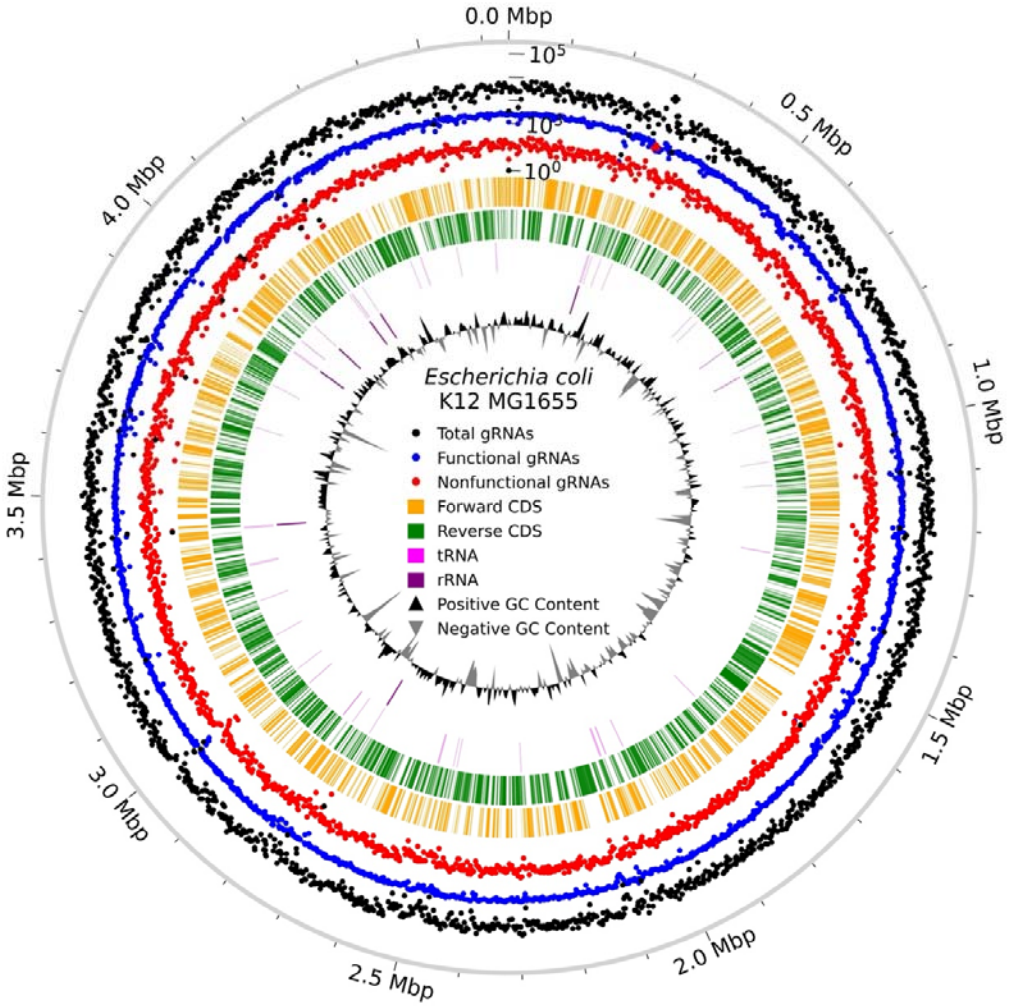

Figure 1. Functional and nonfunctional gRNA targets are uniform across the genome. A bespoke Python 3 script was written using the pycirclize[19] and matplotlib[20] libraries to generate a circos plot. The outer ring contains the genomic coordinates for the *E. coli* K12 MG1655 genome. A triple semi-log scatter plot illustrates the number of total (black dots), functional (blue dots), and nonfunctional (red dots) gRNAs present in each bin. The central dashed lines represent the location of protein-encoding (yellow and green), tRNA (pink), and rRNA (purple) genes. The internal line plot outlines the GC content of each region.

Nonfunctional library. During design of this library, an error in sorting data resulted in the inclusion of a handful of gRNAs that should have been filtered for low Donor library counts; we have decided to include these to simplify analysis, bringing the total number of Nonfunctional library designs to 2,375. We hypothesized that our sequencing method, which used amplicons primed partially off of the gRNA scaffold, could be masking mutations arising within the gRNA scaffold, which would result in a misclassification of a mutated gRNA design as "inactive" when otherwise it would be excluded during analysis. We addressed this potential inaccuracy in the Nonfunctional library analysis by 1) sequencing the entirety of the gRNA cassette and 2) sequencing both Donor and Host libraries in triplicate rather than duplicate to further decrease potential variance. Of the 2,375 gRNAs tested in the Nonfunctional library, 1,651 exhibited nonfunctionality in this setting (Supplementary Data 4). These data show that only 0.3% of the total number of unique gRNA targets in the *E. coli* genome are nonfunctional for Cas9 targeting and cutting.

Restoring gRNA activity by disrupting spacer self-interaction

Spacer secondary structure has previously been shown to inhibit gRNA activity[21–23]. During the design of the Nonfunctional library, RNAfold[18] was used to predict structure and minimum free energy (MFE) for spacers in that library. Ten library designs were manually picked that had stretches of consecutive base pairing between distal and seed regions as well as a low MFE, indicating very stable spacer structure; these gRNAs should exhibit low functionality, though hypothetically they could be restored to function by mutating one or two bases in the distal region of the spacer. Specifically, if one or two mutations could disrupt the non-additive stabilization from base stacking that occurs in the most stable spacer self-interactions, the gRNA could function again as mutations in the distal region are frequently well-tolerated[24]. To test this approach, ten "repaired" gRNAs were designed to test alongside

their predicted nonfunctional counterparts in the Nonfunctional library (Table 1, Supplementary Data 4). The repair involved manually choosing a base most likely to destabilize the spacer via hypothetical base stacking disruption, then confirming a predicted increase in MFE via RNAfold. All ten subject gRNAs matched our initial, manual prediction of non-functionality, and nine of the ten repaired versions of these gRNAs were restored to functionality as shown by a remarked depletion after transformation into the host (Table 1, Supplementary Data 4).

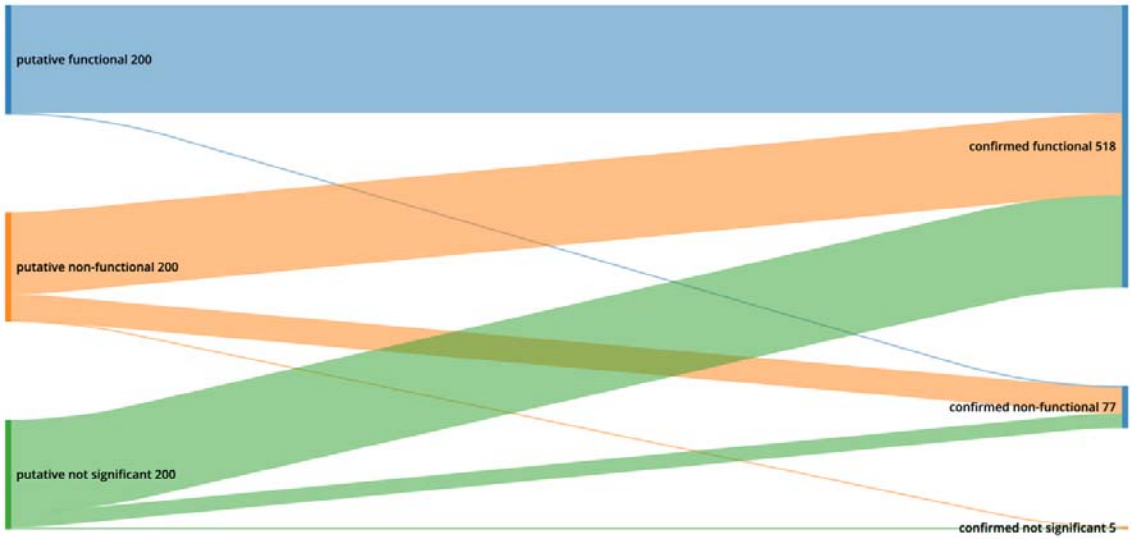

Figure 2. Functional calls from the genome-wide library are conservative. Randomly chosen gRNAs from the genome-wide library (200 each functional, nonfunctional, and statistically insignificant designs) were resynthesized by Genscript then cloned to produce the Confirmation library, which was then transformed into *E. coli* K12 MG1655 and analyzed. While almost all of the putative functional gRNAs demonstrated functionality in the Confirmation library, most of the putative nonfunctional and statistically insignificant designs demonstrated strong functionality in the Confirmation pool. This figure was produced using www.sankeydiagram.net.

Developing design rules for gRNA libraries

Our results validating that spacer self-interaction has major effects on the function of a gRNA suggest that the predicted MFE value can act as a proxy for a number of physical effects that cause strong self-interaction (number and identity of bases paired, base stacking effects, etc.). To compare the distribution of MFE values in all functional gRNAs to the 1,651 confirmed nonfunctional, we separated all gRNAs in both functional categories into bins

based on their spacer MFE, tallied the number of spacers in each bin for each category, then plotted them (Figure 3). While the vast majority of functional gRNAs had an MFE value of -5 kcal/mol or greater, the MFE of the nonfunctional gRNAs peaked at -9 kcal/mol. Graphing these bins shows that an MFE cutoff of between -6 and -7 could serve as a simple but effective cutoff during library design to eliminate a large number of poorly performing guides with little loss of high-performing gRNAs (Figure 3). -6.5 kcal/mol seems ideal, as this value would reject 61% of nonfunctional gRNAs while only losing 2.7% of functional gRNAs; however, a more or less strict cutoff could be appropriate depending on specific experimental contexts and demands on library design.

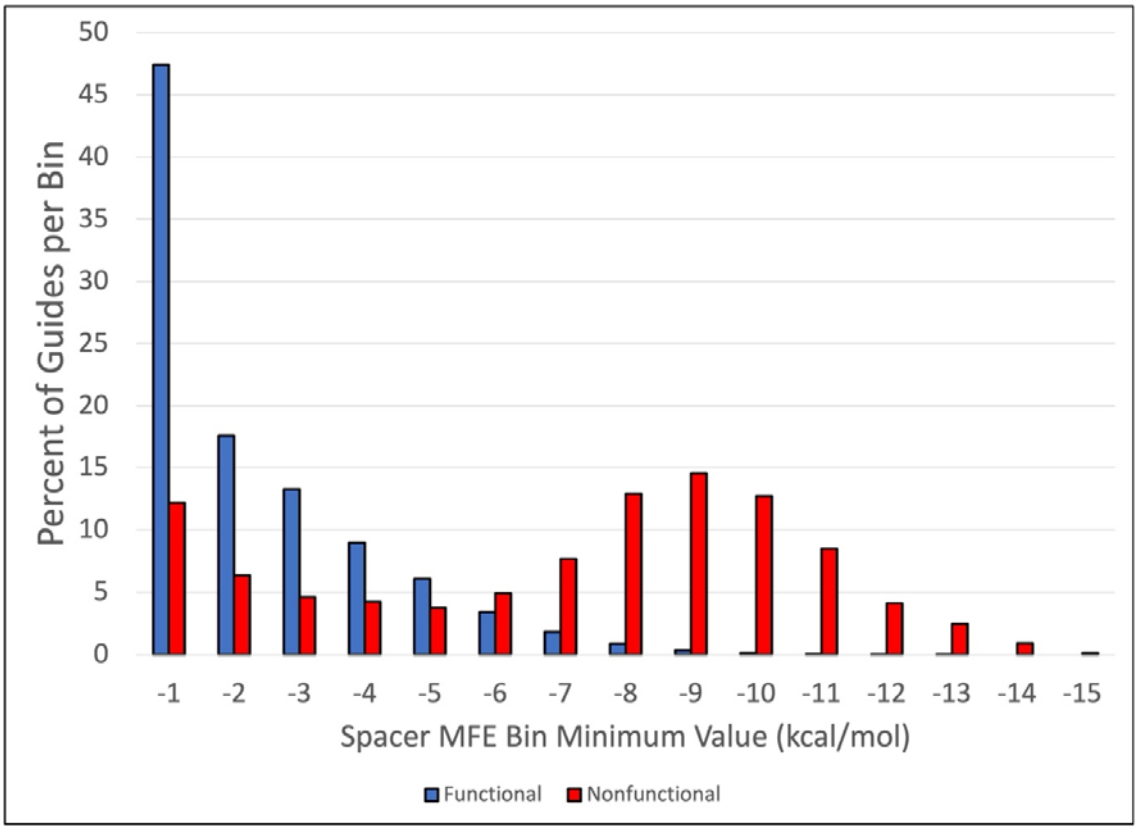

Figure 3. Functional gRNA spacers exhibit greater minimum free energy values than nonfunctional spacers. Spacer MFE values for both functional and nonfunctional gRNAs were binned, then the total for each bin was normalized to the total. The resulting percent of total gRNAs for each bin was then plotted in Excel.

DISCUSSION

Here we detail findings from the first *E. coli* genome comprehensive screen of Cas9 gRNA functionality *in vivo*. While the vast majority of Cas9 targets in *E. coli* can be acted upon effectively, we demonstrated the feasibility of restoring nonfunctional gRNAs to functionality by altering bases in the distal region and determined an empirically informed cutoff MFE value of -6.5 kcal/mol to best avoid the poorest-performing gRNA designs. Finally, we present to the synthetic biology community close to half a million gRNA sequences that have demonstrated activity *in vivo*.

While many gRNA design software packages and associated rules have been made available, they are often unreliable in practice. Numerous factors likely contribute to this disfunction, though misapplication of models derived from divergent biological systems is arguably the most common and most obvious. For example, the commonly used Doench score[25] is one of the first attempts to quantify gRNA targets suitability, but the datasets gathered were from immortalized human cell lines. Despite the very specific context in which this model was developed (eukaryotic system, heavily mutated with compromised cell cycle control and aneuploidy), it sees widespread application as a default in many popular biological CAD programs (*e.g.* Geneious, Benchling). As chromatin structure has been shown to impede Cas9 RNP access to DNA both *in vivo* and *in vitro*[23,26,27], the model employed to generate a Doench score will likely detect whatever features derive from nucleosome occupancy in the *E. coli* genome spuriously. These spurious, out of context features will then mis-score a relevant gRNA. Our results also suggest why the Doench score and other inappropriate measures of gRNA suitability were used for so long in *E. coli* design: if greater than 95% of the potential gRNAs made to act in *E. coli* are in reality functional, the criteria used to pick functional guides ceases to matter. Also, since those gRNAs flagged as

Design	Spacer	Spacer Structure	Spacer MFE	A-score	P-value
OA_J473521_091952	CAGACGGACCTTTCCGTCTG	((((((((....))))))))	-11.5	-0.324815	**0.07682572**
OA_J473521_091952_1mod	CAGAC**T**GACCTTTCCGTCTG	(((((.((....)).)))))	-5.3	8.624973	0.00204634
OA_J473521_104917	GGGCAGCTTTGTTGCCCGTA	(((((((...)))))))...	-11.5	-0.9445995	0.00302306
OA_J473521_104917_2mod	GG**T**CA**T**CTTTGTTGCCCGTA	((.(((.....)).))....	-1.8	9.76318626	0.00375396
OA_J473521_117061	GTGGAGCCTTCGGGCTTCAC	(((((((((....)))))))	-12.2	-1.2480891	0.00450849
OA_J473521_117061_2mod	GT**C**GA**T**CCTTCGGGCTTCAC	.((((....)))).....	-2.2	4.28913801	0.00287841
OA_J473522_086658	CTGGGGTAAATCCACCCCTG	..((((((.....)))))..	-8.2	-1.2709609	0.00563453
OA_J473522_086658_2mod	CTG**TT**GTAAATCCACCCCTG	.....	0	9.03206691	0.00176094
OA_J473522_125490	GCTGCGCGTTGGCGCGGCGT	((((((((....))))))))..	-14.6	-1.0181274	0.01442497
OA_J473522_125490_1mod	GCT**A**CGCGTTGGCGCGGCGT	(((.((((....)))).)))..	-8.6	7.9954639	0.00388022
OA_J473522_150359	GCCGGAGTATTACCCGGCGC	((((((.....))))))..	-10.1	-0.9454579	0.00648248
OA_J473522_150359_1mod	GC**A**GGAGTATTACCCGGCGC	((.(((.....)).))..	-3.9	7.61011715	0.00228956
OA_J473522_166700	TGGCGGACATTGCCGCCGCG	.((((((.....)))))...	-9.8	-0.5471758	**0.05694951**
OA_J473522_166700_1mod	TGG**A**GGACATTGCCGCCGCG	.((.(((.....)).))....	-4.1	9.46673057	0.00131515
OA_J473522_198868	GCGGCCATCATGCCGCTGGT	((((((.....))))))....	-9.1	-0.8006765	0.00773551
OA_J473522_198868_1mod	GC**T**GCCATCATGCCGCTGGT	((.(((.....)).))....	-2.9	8.66369937	0.00098449
OA_J473523_001496	CCGGCGGCGTTTAGTCGCCG	.(((((((.....)))))))	-12.4	-0.7106077	0.02338785
OA_J473523_001496_2mod	CCG**T**C**T**GCGTTTAGTCGCCG	.....((((.....))))..	-2.8	9.1107605	0.00180489
OA_J473523_190366	CCCGGCGCGTCGTGCCAGGG	(((((((((...)))))).)))	-10.4	-0.9764011	0.00546639
OA_J473523_190366_2mod	CC**G**GG**A**GCGTCGTGCCAGGG	((((.(.....).)).)).	-3.6	-0.2696708	**0.11001609**

Table 1. Nonfunctional gRNAs can be restored to function by disrupting spacer secondary structure. Mutations made are in red and emboldened. Functional A-scores are colored blue, while nonfunctional scores are red. Finally, statistically insignificant P-values are emboldened.

to be avoided by models, a type of confirmation bias occurs; the “inactive” gRNAs are inactive because software said so and, thus, are never tested because they should be inactive.

In addition to the possible disparities of using the wrong prediction tools for gRNA design in *E. coli*, there are other potential reasons that our results here contradict previous reports detailing poor gRNA activity within *E. coli*[16]. Cas9 expression conveys a mild fitness defect, which has also been reported previously[28] but is a familiar phenomenon for those who work frequently with Cas9 in living systems. Issues with this defect manifest when Cas9 expression is constitutive and/or too high, as cells will mutate constructs driving Cas9 expression to relieve this defect. Reducing or eliminating this growth defect in turn results in a fitness increase compared to the rest of the population, in turn causing an over-representation in the library. Finally, this Cas9- mutant, if it possesses a gRNA library plasmid, will then replicate that plasmid as it continues to grow, and the gRNA produced appears to be nonfunctional since no Cas9 is expressed. This phenomenon could explain previous reports of high levels of nonfunctional gRNAs[16], as a $Cas9^-$ mutant will behave exactly as a nonfunctional gRNA. These selective pressures must be accounted for during experimental design and, if possible, worked around to prevent these “escapees” from confounding data. Our use of arabinose-inducible Cas9 in this work allows us to control length and level of expression to avoid generating these escapees. In a parallel effort to this work, we have developed a much-improved Cas9/Lambda Red genome editing engine plasmid that makes use of layered control elements to further suppress Cas9 expression, leading to a tightly controlled editing system[29]. Second, use of appropriate, informative controls is crucial to interpret library data correctly. The data presented here were produced using three “spike-in” controls, which are three clonally derived gRNA plasmids that behave in known ways: a functional gRNA control, a nonfunctional gRNA control, and a wash control. The functional gRNA targets a site in *galM* that had been shown previously to be

highly active[30], while the nonfunctional and wash controls contain a gRNA with no target in the genome (this putative lack of activity also having been empirically tested before deployment). These controls are added to the donor library, or spiked-in, prior to transformation into the host, and thus they act as internal standards for expected behaviour within the library. The wash control, meanwhile, is spiked into the post-transformation recovery media to act as a proxy for contamination by plasmid DNA left over from transformation, which can and does interfere with PCR-based methods.

In conclusion, we speculate that the simple but validated design rules we developed here should be widely applicable for gRNA design because the sole metric used to judge the likely activity of a gRNA is derived from the spacer region. Because RNA tends to interact *in cis* and the spacer is on the 5' end of the gRNA transcript, spacer self-interactions are formed long before the scaffold has disassociated from the RNA polymerase, let alone interaction with Cas9 or the genome. Thus, spacer structure logically is a dominant feature contributing to gRNA function, as without RNA-DNA hybridization Cas9 cannot bind to the target DNA locus. This rule is also likely applicable to other nucleases, as they have similar requirements for their functionality though their differing spacer lengths mean different cutoff MFE values would likely have to be developed for each individual nuclease. Overall, this gRNA design approach represents a simple but empirical design strategy to avoid nonfunctional gRNAs in a practical, repeatable manner that could have applications across all organisms.

FUNDING ACKNOWLEDGEMENTS

This material is based upon work supported by the Center for Bioenergy Innovation (CBI), U.S. Department of Energy, Office of Science, Biological and Environmental Research Program under Award Number ERKP886. This work was also sponsored by the

Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy.,

WORKS CITED

1. Alexander, W. G. A history of genome editing in *Saccharomyces cerevisiae*. *Yeast* **35**, 355–360 (2018).
2. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911 (2018).
3. Oost, J. van der & Patinios, C. The genome editing revolution. *Trends Biotechnol.* **41**, 396–409 (2023).
4. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339**, 819–823 (2013).
5. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339**, 823–826 (2013).
6. Brouns, S. J. J. *et al.* Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* **321**, 960–964 (2008).
7. Lintner, N. G. *et al.* Structural and Functional Characterization of an Archaeal Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-associated Complex for Antiviral Defense (CASCADE) *. *J. Biol. Chem.* **286**, 21643–21656 (2011).
8. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).
9. Nishimasu, H. *et al.* Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. *Cell* **156**, 935–949 (2014).

10. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
11. Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc. Natl. Acad. Sci.* **108**, 10098–10103 (2011).
12. Sapranauskas, R. *et al.* The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res.* **39**, 9275–9282 (2011).
13. Collias, D. *et al.* A positive, growth-based PAM screen identifies noncanonical motifs recognized by the S. pyogenes Cas9. *Sci. Adv.* **6**, eabb4054 (2020).
14. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).
15. Yarrington, R. M., Verma, S., Schwartz, S., Trautman, J. K. & Carroll, D. Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proc. Natl. Acad. Sci.* **115**, 9351–9358 (2018).
16. Zerbini, F. *et al.* Large scale validation of an efficient CRISPR/Cas-based multi gene editing protocol in Escherichia coli. *Microb. Cell Factories* **16**, 68 (2017).
17. Noshay, J. M. *et al.* Quantum biological insights into CRISPR-Cas9 sgRNA efficiency from explainable-AI driven feature engineering. *Nucleic Acids Res*. **51**, 10147–10161 (2023).
18. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The Vienna RNA Websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).
19. Shimoyama, Y. pyCirclize: Circular visualization in Python. (2022).
20. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

21. Nowak, C. M., Lawson, S., Zerez, M. & Bleris, L. Guide RNA engineering for versatile Cas9 functionality. *Nucleic Acids Res.* **44**, 9555–9564 (2016).
22. Liang, G., Zhang, H., Lou, D. & Yu, D. Selection of highly efficient sgRNAs for CRISPR/Cas9-based plant genome editing. *Sci. Rep.* **6**, 21451 (2016).
23. Jensen, K. T. *et al.* Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Lett.* **591**, 1892–1901 (2017).
24. Liu, X. *et al.* Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system. *Sci. Rep.* **6**, 19675 (2016).
25. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
26. Yarrington, R. M., Verma, S., Schwartz, S., Trautman, J. K. & Carroll, D. Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9351–9358 (2018).
27. Horlbeck, M. A. *et al.* Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife* **5**, e12677.
28. Ryan, O. W. *et al.* Selection of chromosomal DNA libraries using a multiplex CRISPR system. *eLife* **3**, e03703 (2014).
29. Phillips, E. K., Klingeman, D. M., Guss, A. M., Eckert, C. A. & Alexander, W. G. Multilayered regulation by RNA thermometers enables precise control of Cas9 expression. 2025.05.12.651296 Preprint at https://doi.org/10.1101/2025.05.12.651296 (2025).
30. Garst, A. D. *et al.* Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechnol.* **35**, 48–55 (2017).