Using the CRISPR-Cas9 System Analogous to Restriction Enzymes Whose Recognition Site is 23 Nucleotides to Alter Pooled CRISPR Libraries

Using the CRISPR-Cas9 System Analogous to Restriction Enzymes Whose Recognition Site is 23 Nucleotides to Alter Pooled CRISPR Libraries

Briana L. Fritchman

A Thesis in the Field of Biology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

March 2022

Abstract

Pooled CRISPR libraries in recent years have been widely used to perform genetic screens. Such a library contains tens of thousands of single guide RNA (sgRNA) constructs that target a set of genes of interest. A perfect library is a collection of all intended constructs, and all library members are uniformly distributed. However, in reality, that is never the case. During library production, unwanted sgRNA constructs might be introduced and some constructs are more abundant than other. These issues can cause issues with screen setup and data analysis. For example, a library that has high proportion of unwanted constructs, or a library that have lopsided abundance distribution may force us to scale up the screen setup, in order to have large portion of the library members screened. Currently there is no good way to remove sgRNA constructs from plasmid pooled CRISPR libraries, other than to remake the library. Knowing that the CRISPR-Cas9 system can cleave a 20-nucleotide region, this study aimed to show that sgRNA constructs present within a CRISPR Library could be removed by using sgRNA-Cas9 complexes in similar manner to restriction enzymes by using a 23-nucleotide recognition sequence.

First, cleaving a plasmid DNA containing a single sgRNA construct was demonstrated, by designing a sgRNA oligo targeting the 20-nucleotide sgRNA construct present within a plasmid DNA sample. After the initial successful test, an additional seven sgRNA oligos were designed to target the complimentary sgRNA constructs present within a genome-wide CRISPR library, a library of 70,000 members. The results

from the first attempt came out inconclusive, due to shallow sequencing that resulted in non-targeted sgRNA constructs appearing to be depleted after analyzing the sequencing data. The second attempt however, with some optimizations was able to prove that sgRNA constructs could be removed up to 8-fold from a pooled CRISPR library. An additional experiment showed that after digestion of plasmid DNA with a sgRNA-Cas9 complex, the cleaved 20-nucleotide sequence has the ability to be ligated back into the opened plasmid with blunt-end ligation. Just like restriction enzymes, in our sgRNA-Cas9 system, the cleavage was not 100% in either the single plasmid DNA or the pooled CRISPR library, further optimizations past this study will need to be completed to help obtain higher cleavage efficiency.

Dedication

I'd like to dedicate this work to my parents, who have continued to believe in me and support me along the way. Love you both.

Acknowledgments

I would like to first thank my thesis director and manager Dr. Xiaoping Yang. Thank you for all of your time, help, and guidance throughout this project.

I would like to also thank by husband Ryan Steward. Thank you for always being there to support me and make me extra pots of coffee while working late on this project.

An additional thank you to Edith Sawyer, Berta Escude Velasco, Luke Sprenkle, and Monica Robinson for cheering me on throughout this journey in the lab.

Also, a final thank you to Dr. David Root and Dr. Xiaoping Yang for giving me the opportunity to complete this project in the Genetic Perturbation Platform at the Broad Institute.

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

Restriction enzymes are commonly used as a tool to cleave specific regions of DNA. These enzymes are proteins produced by bacteria that recognize, in most common cases, a specific four to eight nucleotide sequence and cut around or within the sequence (Vincze et al., 2003). Different types of bacterial species recognize different recognition sequences. Currently four different types of restriction enzymes have been recognized (Britannica 2020). Out of the four types, Type II enzymes are the most frequently used, because they cleave within a recognition sequence, compared to the other three types that either cut outside of the recognition sequence or cut at random (Pray 2008). As of the early 2000s over 3000 type II enzymes have been identified (Pingoud et al., 2001). Hundreds of these type II restriction enzymes are commercially available through many sites such as New England Biolabs and Thermofisher Scientific.

There are many applications within biology where restrictions enzymes are used. One common application is using restriction enzymes to cut and using DNA ligase to stitch two separate DNA fragments together. This is performed by cleaving both fragments with the same restriction enzyme. Cleavage of fragments at the recognition sequence will leave the ending strands with a 5' phosphate and a 3' OH-end, which are required for ligation. Additionally, these cleaved DNA strands will either have 5' or 3' overhangs, referred to as sticky ends, or both strands will have blunt ends (Pingoud et al. 2001). Sticky ends are generally favored over blunt ends, because they are more efficient during ligation and only sticky ends with complimentary strands will ligate together

(Brown 2002). With the resulting two DNA fragments containing compatible sticky ends or blunt ends, an additional enzyme called DNA ligase will seal the two fragments together and form a new circular plasmid (Pray 2008).

Another common application for restriction enzymes, is to fingerprint DNA molecules, for example, plasmid DNA, genomic DNA or forensic DNA, by looking at the specific banding patterns generated by cleaving the DNA at specific recognition sites. Generally, one or two enzymes will be digested with plasmid DNA that will result in a unique banding pattern based on the cut sites. Tools such as SnapGene, can simulate the banding pattern that would be seen on an agarose gel of a given plasmid based on what restriction enzymes were chosen to cleave the DNA. When choosing restriction enzymes for comparing two plasmids, it's important that the chosen enzymes produce distinct banding patterns between the two plasmids.

Type II restriction enzymes generally have good specificity in recognizing their recognition sequence. Depending on the set-up conditions, and the enzyme itself however, off-target cleaving can occur. Cleavage of off target recognition sequences in restriction enzymes is referred to as star activity (Lundin et al., 2015). To decrease the chances of star activity occurring, many restriction enzyme distribution sites, such as NEB encourage the user to follow their specific protocol for each enzyme. For example, different Type II enzymes can require different reaction buffers, and depending on the buffer that is being used NEB will provide the percent cutting efficiency of the enzyme when used with different buffers. Additionally, NEB has come out with high-fidelity versions of many of their restriction enzymes, which they have noted helps with reducing star activity.

Function of the CRISPR Cas9 System

The CRISPR-Cas9 system has become in recent years a common practice for perturbing genes by either in-vitro or in-vivo methods (Hartenian & Doench 2015). The system consists of a single-guide RNA (sgRNA) paired with *Streptococcus pyogenes (S.pyogenes)* Cas9 (Sun et al., 2016). By changing the target sequence of a sgRNA, the Cas9 can target a different location within the genome (Doench et al., 2016). The sgRNA is made up of two components, CRISPR RNA (crRNA) and Trans-activating crRNA (tracrRNA) (Kelley et al., 2016). The second component of the sgRNA is the tracrRNA. TracrRNA is a conserved sequence that with the addition of a linker sequence the crRNA and tracrRNA can form the sgRNA. The tracrRNA region is important for the binding of Cas9 to the sgRNA, and the crRNA portion is important to guide the Cas9 to its correct target sequence (Hartenian and Doench 2015). The crRNA is made up of a unique 20-nucleotide sequence that is the reverse compliment of the target sequence within a given gene (Zhang et al., 2017). One important factor when designing the crRNA sequence, is that a protospacer adjacent motif (PAM) site must be present on the 3'end following the guide sequence, on the target DNA strand, the strand that carries the same sequence as the sgRNA (Sun et al., 2016). The PAM site for *S.pyogenes* Cas9 is NGG, with N being any of the four nucleotides. This site is necessary in order for Cas9 to recognize and bind to the target sequence region.

Once the sgRNA-cas9 complex is bound to the desired target sequence, the Cas9 will cleave the DNA three nucleotides upstream of the PAM site, resulting in a double strand break (DSB). The DSB in cells will generally be repaired by nonhomologous end

joining (NHEJ) (Wu et al., 2014). Due to NHEJ being very error prone the resulting repaired strand will have deletions or insertions, making the target gene inactive (Scott, et al. 2019). If the sgRNA-Cas9 complex is added to DNA in-vitro with no cells, there is no repair of the double stranded break by NHEJ, resulting in two DNA fragments with blunt ends (Liu et al., 2015).

When designing sgRNAs to target a given region within the genome, factors such as on and off target activity need to be taken into consideration, as they can vary greatly from one sgRNA to another. Many studies have shown that off-target effects are mainly due to either the design of the sgRNA itself or the experimental conditions (Doench et al., 2016). To help alleviate this problem, there are now a few online tools available that use specific rule sets to predict on and off target activity for a given sgRNA. Additionally, these tools can be used to show you how well your sgRNA design scored amongst another sgRNA's targeting a given region (Wu et al., 2014). It is important to note that these rule sets continue to be improved by gathering more information and data after previous in-vivo CRISPR screens (Doench et al., 2016).

CRISPR Pooled Libraries

Pooled screening has become a favorite way to screen a large number of genetic perturbations within a given population of cells. These screens are generally represented as negative or positive selection screens, where the ending result is to look at genes that have either been enriched or depleted during the screen (Piccioni et al, 2018). To assure

that only a single perturbation is incorporated in a single cell a multiplicity of infection (MOI) is calculated for viral transduction (Sanson et al., 2018).

In order to perform a CRISPR pooled screen, a pooled CRISPR library needs to be designed and made. On average, CRISPR libraries contain four to six sgRNAs per gene (Wu et al., 2014). To design each sgRNA, a tool such as STARS is used. This tool uses rule sets that can predict the on and off-target activity of each sgRNA for a given gene and ranks them based on their predicted performance (Doench et al., 2016). Pooled libraries are very customizable, because they can be designed to only contain sgRNAs targeting specific genes of interest (Sanson et al., 2018). CRISPR libraries range from 200 sgRNAs up to genome wide scale, ~ 90K. Once the sgRNAs have been picked for a given library, they can be cloned into their desired viral vector (Doench 2018).

Generating a pooled library requires multiple steps. First, oligos for all the sgRNAs are made by synthetic oligo synthesis. The ordered oligos are typically referred to a chip, because they are synthesized on microchips (Hughes and Ellington 2017). Companies such as Twist and Genscript generate these microchips. Many times, multiple libraries are put onto the same chip, and are amplified out separately by having specific primer sites added onto each sgRNA sequence beforehand (Sanson et al., 2018). Once the chip has been received from the vendor, PCR Amplification is performed, to amplify and separate out each pooled CRISPR library. Once the oligos have been purified post PCR, they are cloned into a vector backbone using Golden Gate Technology (Hu et al., 2019). Once the library has been transformed and the plasmid DNA has been extracted, the library can be sequenced by Illumina sequencing.

Illumina sequencing data for libraries are deconvoluted and analyzed to make sure that there are not too many sgRNA sequences missing, and that the correct sgRNA constructs are present. Additionally, it is important that the sgRNA constructs within the library are well distributed, meaning that each sgRNA sequence within the library has a similar number of read counts in comparison to the total number of read of the library.

Due to possible errors during sgRNA design, oligo-synthesis, PCR, or contamination, un-intended sgRNA constructs could be present within the library. Synthetic oligo synthesis error rates average around 0.5%, that is, 1 error per 200 nucleotides (Hughes and Ellington 2017). If these error rates are applied to a genome wide library of 20-nucleotide sgRNA, then roughly 900 (10%) out of 90,000 sgRNAs will be affected, by oligo synthesis error alone. During PCR, possible single nucleotide switches can occur due to DNA polymerase errors (Kebschull and Zador 2015). This is important, because a study had shown that if a mismatch that is 10 to 12- nucleotides within the PAM proximal region occurs, there is a decrease in on-target activity (Wu et al., 2014). Another study had shown that single mismatches at the 5' end of the target region are more tolerated, but if they are closer to the PAM region, on-target activity will be decreased (Kelley et al., 2016). These un-intended sequences are not ideal for pooled CRISPR screening, as a study has shown that even a single base mutation within the sgRNA sequence can cause decreased sgRNA activity (Doench et al., 2016). Additionally, if certain sgRNA sequences account for a higher number of read counts, the downstream effects could be misrepresentation of gene hits from the pooled screen (Jun et al., 2020). Therefore, once the CRISPR library has been sequenced it is important to

make sure that the library looks correct before moving into screening. Currently if there are issues with the library there are no good ways to fix the problems, other than addressing the problems once the screening data has come back or remaking the library, which is costly and time consuming.

Research Problem

There are some limitations when using restriction enzymes in cleaving DNA. One main restriction, is there are limitations to where you can cut within a plasmid, based on where restriction recognitions sequences are located. This can be avoided by pre-determining your cut sites while designing the plasmid, but this is not always the case. Additionally, sometimes an ideal recognition site cannot be used, because there is more than one recognition sequence within the plasmid, which will result in more than one cleavage site. Most importantly if you are trying to remove sgRNA sequences within a CRISPR library you will be unable to, as every sgRNA sequence within the library is cloned into the same vector backbone.

The aim of the study was to determine if a 20-nucleotide sgRNA sequence within a plasmid could be removed in a similar manner to a restriction enzyme, by using the CRISPR-Cas9 system with a 23-nucleotide recognition sequence. The 23-nucleotide recognition sequence includes both the 20-nucleotide sgRNA sequence and the additional 3-nucleotide PAM sequence. An additional aim of this study was to not only try to remove a single sgRNA sequence from DNA, but to remove multiple sgRNAs from a

pooled CRISPR library. The results to these experiments, may highlight the opportunity for the CRISPR-Cas9 system to help edit pooled CRISPR libraries, and possible further applications outside of only targeting sgRNA sequences.

Chapter II.

Materials and Methods

A single sgRNA was synthesized and paired with Cas9 to form a complex that targeted a plasmid DNA containing only a single 20-nucleotide sgRNA construct. Results were analyzed by a gel. An additional seven sgRNAs were then synthesized to cleave sgRNA constructs present in a pooled CRISPR library. Data from treated and non-treated libraries was deconvoluted and analyzed to look for sgRNA construct read count depletion. Additionally, ligation to pair blunt-ended DNA together after cleave was performed. Techniques and protocols used throughout the experiments are described in this chapter.

Restriction Enzyme Digestion of DNA

Restriction enzyme NotI-high fidelity (HF) (NEB) was digested with plasmid DNA to generate a single cut, resulting in linear plasmid DNA. The Digest was set up by combining 1.5uL of plasmid DNA at 400ng, 3uL of 10X Cutsmart Buffer (NEB), 1uL of NotI-HF, and 24.5 uL of nuclease-free water into one well of a 8-well PCR Strip Tube. Reaction was mixed by pipetting total volume up and down. The strip tube was then incubated for 18 hours at 37 °C in a thermocycler. After incubation, enzymes were inactivated by incubating tube for 5 minutes at 65°C in a thermocycler.

Pick and Design of sgRNAs

The sgRNA oligo sequences were designed by taking the 20-nucleotide reverse compliment sequence of each targeted sgRNA construct. Each 20-nucleotide sequence was immediately followed by a PAM site on the 3' end. The PAM site was not included in the oligo sequence. The sgRNA synthesis kit protocol (NEB) required a T7 promoter sequence (TTCTAATACGACTCACTATAG) to be added before each 20-nucleotide sequence on the 5' end. In between the T7 promoter sequence and target sequence, nucleotide G was added as the transcription start site. On the 3' end of each 20-nucleotide sequence, a 14-nucleotide overlap sequence (CTGTAGTGCAGGCATTGGG) was added to help form the tracrRNA during the synthesis reaction. The oligo sequences were ordered as lyophilized single stranded DNA (IDT) and then resuspended in nuclease-free water once received.

Synthesis of sgRNA and Purification

The ordered oligos were synthesized into sgRNA using a sgRNA synthesis kit (NEB). To set up the reaction 2uL of nuclease-free water, 10uL of EnGen2X sgRNA reaction Mix, S. pyogenes, 5uL of (1uM) ordered oligo, 1uL of (0.1M) DTT, and 2uL of EnGen sgRNA Enzyme Mix, were combined into a 8-well PCR strip tube. Total reaction volume was 20uL. After the reaction was mixed thoroughly by pipetting up and down, the strip tube was placed into a thermocycler for 30 minutes at 37 °C. To stop the

reaction, 30uL of nuclease-free water and 2uL of DNase I were added to the reaction. The tube was then placed back onto the thermocycler for another 15 minutes at 37 °C.

Excess salt, nucleotides, and proteins were purified from newly synthesized RNA using a RNA Cleanup Kit (NEB). To begin purification, 100uL of RNA Cleanup Binding Buffer was added to 50uL of RNA in a 1.5mL microcentrifuge tube. Then 150uL of 95% Ethanol was added to the tube and mixed by pipetting up and down. Total volume was then transferred into a spin column that was set within a collection tube. Tube was spun for 1 minute at 16,000 x g. After spin was finished, flow through was discarded. Once column was placed back within collection tube, 500uL of RNA Cleanup wash buffer containing 95% Ethanol was added to column. Tube was re-spun for an additional 1 minute at 16,000 x g and flow through was discarded. Wash step was repeated a second time. After second wash, the column was placed inside a new 1.5ml microfuge tube. Column was eluted with 50uL of nuclease-free water and then spun for 1 minute at 16,000 x g to collect purified sample inside the microfuge tube. Purified RNA was then quantified using a single channel nanodrop, Nanodrop One (Thermofisher).

In vitro digestion of DNA with sgRNA-Cas9 complex

Formation of a sgRNA-Cas9 complex followed by in-vitro digestion with plasmid DNA containing target sequence, was taken from protocol "In vitro digestion of DNA with Cas9 Nuclease, S. pyogenes" (NEB). Cas9 Nuclease. *S.pyogenes* was purchased at a concentration of 20uM (NEB). Cas9 was diluted down to 1uM by doing a 1:20 dilution with NEBuffer r3.1. The Cas9 dilution was re-made for every experiment, because the

NEB protocol instructed that once Cas9 was diluted in buffer, the enzyme needed to be used immediately or needed to be diluted in additional buffers for long time storage. Forming the sgRNA-Cas9 complex was done by combining 20uL of nuclease-free water, 3uL of NEBuffer r3.1, 3uL of sgRNA at 300nM, and 1uL of Cas9 Nuclease. *S pyogenes* at 1uM into a 8-well PCR strip tube. Nuclease-free water was used to replace sgRNA or Cas9 when omitted from the reaction. Once all reagents were mixed thoroughly, the strip tube was incubated in a thermocycler for 10 minutes at 25 °C. After the sgRNA-Cas9 complex was formed, either 1uL or 3uL of DNA at 3nM was added into the reaction mixture and mixed thoroughly. Strip tube was then placed back in a thermocycler for 15 minutes at 37°C. The reaction was then stopped by either directly incubating the sample for 5 minutes at 65°C in a thermocycler, or by adding 1uL of Proteinase K (Omega) and then incubating for 10 minutes at room temperature.

DNA Cleanup via SPRI Purification

Digested DNA was purified using a technique called Solid Phase Reverse Immobilization (SPRI). AMPure XP (Beckman) solution containing the SPRI beads were added at a 1:1 ratio with the sample in one well of a polystyrene 96-well plate. Once thoroughly mixed by pipetting, the sample was incubated for 5 minutes at room temperature. The 96-well plate was then placed on top of a 96-well magnetic plate and sat for 3 minutes to allow beads to form a circular ring around the edge of the well. The supernatant inside the well was aspirated by a pipette and discarded. With the 96-well plate remaining on the magnet, the well was washed twice with 100uL of 70% ethanol.

The 96-well plate was removed off the magnet and sat for 1 minute at room temperature to remove residual 70% ethanol. After the minute, the beads were resuspended 40uL in nuclease-free water by mixing thoroughly with a pipette. The 96-well plate was then placed back on the magnet for 2 minutes, until beads formed a ring along the edges of the well. The supernatant, containing the purified DNA was aspirated from the well and placed in a new 1.5ml microfuge tube.

Removal of double and single stranded linearized DNA

A combination of Exonuclease I (NEB) and III (NEB) were used to remove nucleotides from both double and single stranded linearized DNA. Reaction was set up by combining 25uL of DNA cleaned up via SPRI, 1uL of Exonuclease I, 1uL of Exonuclease III, 10uL of NEB1 Buffer, and 15uL of nuclease-free water into a 8-well PCR strip tube. Once all reagents were mixed, the strip tube was placed into a thermocycler and incubated for 30 minutes at 37°C.

5’ & 3’ Phosphate Removal

Calf intestinal alkaline phosphatase (CIP) removes 5’ and 3’ phosphates from linearized DNA. To prevent ligation from happening, DNA cleaved by a sgRNA-Cas9 complex had 1uL of Quick CIP (NEB) added into tube of a 8-well strip tube. The strip

tube was then placed in a thermocycler for 10 minutes at 37°C. Reaction was stopped by incubating tube for 2 minutes at 80°C.

Purified DNA using PCR Spin Column Cleanup

Samples used during the ligation experiment were purified by the QIAquick PCR Purification Kit (Qiagen). To set up reaction, 80uL of Buffer PB was added to the CIP and non-CIP treated sample that each contained volume of 16uL, making a 1:5 ratio of DNA to Buffer PB. The reaction mixture was mixed thoroughly by vortexing. The total volume was then transferred into a QIAquick spin column that was inserted in a 2mL collection tube. To bind the DNA to the column, the tube was centrifuged for 1 minute at 16,000 x g. Flow through in collection tube was discarded. To wash the column, 750uL of Buffer PE was added into the QIAquick spin column and then centrifuged for 1 minute at 16,000 x g. Flow through was discarded. To remove any residual buffer, the column was placed back in the centrifuge and spun for an additional minute at 16,000 x g. QIAquick spin column was then placed inside a new 1.5mL microcentrifuge tube. DNA was eluted by adding 17uL of nuclease-free water and spinning the tube for 1 minute at 16,0000 x g.

Ligation

T4 Ligase enzyme (NEB) was used to ligate together blunt end DNA strands. The ligation protocol provided with the T4 ligase was used. A 20uL reaction was set up in a 8-well PCR strip tube containing 2uL 10X T4 DNA ligase buffer, 1uL T4 DNA ligase enzyme, and 17uL of purified plasmid DNA containing both the opened plasmid and the 20-nucleotide sgRNA region that was cleaved by the sgRNA-Cas9 complex. Reaction was mixed by pipetting up and down and incubated for 2 hours at room temperature on a lab bench. T4 ligase was inactivated by incubating tube for 10 minutes at 65°C in thermocycler.

Analysis via Gel

DNA was visualized by a 1% ethidium bromide agarose gel. The agarose gel was set up by combining 150mL of 1X TAE buffer and 1.5g of agarose power in a 500mL Erlenmeyer flask. After swirling to mix the powder and buffer, the flask was placed in a microwave for ~2 minutes until powder was dissolved. Once the flask had slightly cooled, 15uL of Ethidium Bromide (Invitrogen) was added to liquid agarose and swirled gently to be incorporated. Mixture was then poured into a gel cast containing a 12-well comb and sat undisturbed until solidified. Once the gel had solidified, the comb was removed and 1X TAE buffer was added into gel box until gel fully submerged.

Before loading samples onto gel, a 1:5 ratio of sample to purple loading dye (Invitrogen) was used to calculate how much dye to add to the samples. The dye was mixed until thoroughly combined by pipetting up and down. Samples containing dye were loaded into each well of the gel. Ranges of volume added into the wells varied from each experiment based on initial sample volume. The first well on the left-hand side contained 15uL of 1 kb+ ladder (Invitrogen). Gels were run for 2 and a half hours at 90 volts until the 1 kb+ ladder bands had distinct separation. The gels were then placed on a Gel Doc Imager (Bio-Rad). Using ethidium bromide protocol on the imager software, the gel images were taken.

Electroporation, Plasmid Prep, PCR, and Sequencing Libraries

Libraries were transformed using Stbl4 cells (Invitrogen). To set up electroporation, 1uL of library was added to 1 tube containing 100uL of Stbl4 cells. For the first experiment only one tube of Stbl4 cells were used to electroporate the treated library. For the second experiment, three Stbl4 tubes were used per treated and non-treated library. After DNA was mixed gently with the cells, 30uL was aspirated from the cell tube and dispensed into a MicroPulser Electroporation cuvette (BioRad) with a 0.1cm gap. The cuvette was then placed into a Micropulser (BioRad) set to the E. coli setting (EC1). Once pulse was complete, the cells were quenched in 500uL of Super Optimal Broth (SOC). SOC media containing cells was placed into a 15mL cell culture tube that contained an additional 2.5mL of SOC. This process was completed two

additional times for each tube of cells, resulting in a total of three culture tubes per 1 tube of Stbl4. Tubes were shaken in a 37 °C warm room for one hour. During the incubation hour, agarose bioassays containing (100ug/mL) Carb were placed into the 37 °C warm room to be pre-warmed. After one hour, the bioassays were taken out of the warm room, approximately 30 glass beads were added per bioassays to help spread around bacteria culture. One 3mL tube of culture, represents one bioassay, and therefore one tube of Stbl4 cells generates three tubes of culture and three bioassays. Before plating, all culture tubes for a given library are pooled and mixed together before dispensing 3mL of culture onto each bioassay. Once culture has been dispensed, the bioassays are gently shaken by hand to help distribute the glass beads throughout the plate. Once glass beads have spread liquid culture around, beads are discarded from bioassay into waste bin. The bioassays are then put into the 37 °C warm room overnight for a total of 16 hours.

An additional bioassay was pre-warmed to perform a serial dilution that would be used the following day to calculate roughly the number of colonies per construct, based on the total number of estimated bacteria colonies on each bioassay. To set up colony count plate, 90uL of SOC was added into wells B01, C01, D01 and E01 of a 96-well plate. Additional columns can be used if more libraries are being set up. Once the culture tubes were combined and mixed, before dispensing onto the bioassays, 90uL was aspirated and dispensed into well A01 of the 96-well plate. To begin the serial dilution 10uL from well A01 is dispensed into well B01 and mixed by pipetting up and down. This is repeated by then taking 10uL from B01 to C01, C01 to D01 and D01 to E01. The following dilutions generated were a 1:10, $1:10^2$, $1:10^3$, $1: 10^4$, and $1:10^5$. Once the serial

dilution was complete, 10uL from each well was then dispensed onto the extra bioassay and then tipped to allow the liquid to run down in a straight line across the plate.

After 16 hours, the bioassays were taken out of the warm room. Before scraping, the plate containing the serial dilutions was counted. To estimate how many colonies were on the bioassays, a calculation was done. Before doing the calculation, a dilution on the plate was chosen that had individual colonies for easy counting. The dilution chosen varied for each library transformation depending on how well the colonies grew. Once the dilution set was chosen, the individual colonies were counted. The raw count of colonies was multiplied by the total mL that were plated. For example, if only one tube of cells was used for transformation then the total milliliters would be 9, because there was 3mL in 3 separate tubes. The raw count multiplied by the total number of milliliters was then multiplied by 100 and then by the dilution factor that was used to count the colonies. The final calculated number was the estimated total number colonies across all the bioassays. This number was then divided by the total number of constructs in the library, to provide the estimated colonies for each construct. In general, a good number is 100 colonies per construct. If the number is lower than 100, generally there will not be a good distribution of the sgRNA constructs within the library.

Once the colony count had been taken, the bioassays were scrapped. The bioassays contained lawns of bacteria, meaning the bacteria colonies were not easily distinguishable. Repeating twice, 15mL of LB only media was dispensed onto each bioassay. A plastic plate spreader was used to scrape and collect the bacteria colonies off the agar. The media containing the collected bacteria was then poured into 250mL conical. Bioassays containing the same library and treatment type were all collected into

the same 250mL conical. The 250mL conical was then pelleted using a table top centrifuge for 5 minutes at 2250 rpm. Once a bacterial pellet formed at the bottom of the conical, the supernatant was then discarded. DNA was extracted from bacteria pellets using a HiSpeed Maxi Kit (Qiagen).

Prepped libraries were set up for PCR to attach the necessary Illumina primers onto the sgRNA sequences. Forward primer, P5 stagger mix, contains both a stagger region and the 5' Illumina standard primer. Reverse primer, P7 contains the 3' Illumina standard primer and a unique 8 mer-index that can be used to identify multiple libraries, when submitted in one lane of sequencing. The P7 primer was set up in a 96-well PCR plate, with each well containing a unique 8-mer index. The vector backbone present in the library contains a specific P5/P7 primer pair.

A 100uL PCR reaction was set up in a 8-well PCR Strip Tube containing 10uL of 10X Titanium Taq Buffer (Takara), 8uL of dNTP (Takara), 1.5uL of Titanium Taq Enzyme (Takara), 5uL DMSO, 0.5uL(100uM) P5 primer, 1uL of plasmid library, and 10uL (5uM) from one of the P7 primer wells. Record of what P7 primer well was used for each library was noted in a spreadsheet. The reaction mix was then placed in a thermocycler, using these cycling conditions: 5 minutes at 95°C, repeating these steps 10 times, 30 seconds at 94°C, 30 seconds for 53°C, 20 seconds at 72°C, and then finished by 10 minutes at 72°C and held at 4°C for infinite time.

To purify the PCR Product, SPRI purification was done, however with the AMPure beads at 1.8x the volume of the PCR Product. Therefore, to set up 100uL of the PCR product and 180uL of AMPure beads were added to a well of a polystyrene 96-well plate. Once beads and PCR Product were mixed, the 96-well plate sat for 5 minutes at

room temperature before being placed onto 96-well magnetic plate. Once the beads had formed a ring around the edges, the supernatant was aspirated from the well and discarded. The well was washed twice with 180uL of 70% Ethanol. After washes were complete, the 96-well plate was removed from the magnet and sat for 1 minute to allow for residual ethanol evaporation. Beads in the well were then mixed with 50uL of nuclease-free water. After mixing, the plate was put back on magnet. Once beads formed a ring around the edges, the supernatant was aspirated and put into a clean 1.5mL microcentrifuge tube. DNA was then quantified using a nanodrop. Libraries were pooled together based off their concentrations from the nanodrop and their library size, to make sure every pool had equal weight per construct. The pooled libraries were submitted to the Broad Institute's Genomics Platform for MiSeq sequencing with 50 cycles.

Sequencing Analysis

Raw sequencing data was deconvoluted using a software tool called PoolQ (PoolQ, 2019), where it is able to pair raw sequencing reads with individual sgRNA construct barcodes present within the submitted library. The construct barcode sequences for pooled CRISPR libraries are the 20-nucleotide sgRNA sequences. The tool uses a condition file that lists what library corresponds to which P7 unique 8-mer index used during NGS PCR set up. This tool will use this 8-mer index to separate out the multiple libraries submitted in one lane of sequencing. Next, the tool takes the search premix CACCG, which is present right before the 20-nucleotide sequence when cloned into vector backbone, to identify the sgRNA construct barcodes. The tool then filters out the identified from the raw sequencing data. The identified construct barcode sequences are

then matched to the reference file that contains all sgRNA construct barcode sequences present in a given pooled CRISPR library.

Once the raw sequencing data was deconvoluted a count file was generated. One column contains all the sgRNA construct barcode sequences in the library and a second column contains a number representing the number of times that each sequence was counted in the sequencing data. This file was then used for further analysis of the reads.

The cumulative distribution graphs were plotted to show the area under the curve (AUC). The graph has the sgRNAs ranked by abundance on the x-axis. This numbers on the x-axis were calculated by first ranking the sgRNA constructs in descending order of their total read counts. The rank of each construct was then divided by the total number of constructs within the library to generate a number between 0 and 1. The y-axis has the fraction of each sgRNA represented within the total number reads from the library. The graphs for both experiments were generated by an in-house tool called Apron.

The number of read counts that were read out for each sgRNA construct were divided by the total number of reads for each library to generate a fraction. The fraction was taken for both the library treated with the sgRNA-Cas9 complex and the non-treated library. Fraction values were log2-transformed for both the treated and non-treated sgRNA constructs. The log2-fold change (LFC) was calculated by taking the log2 norm of the treated library minus the log2 norm of the non-treated library.

Using the graphing tool PRISM, a scatterplot was generated by plotting the log2 norm treated values on the y axis, and the log2 norm non-treated values on the x-

axis. The sequences that were targeted using the sgRNA-Cas9 complex were labeled in red.

The log2-fold change value for every sgRNA construct within the library was ranked. A graph generated by python plotted the ranked value of each construct on the x-axis and log2-fold change on the y axis. The targeted sgRNA constructs were labeled in red.

Chapter III.

Results

Various experiments were set up to demonstrate that 20-nucleotide sgRNA constructs could be removed from both single construct plasmid DNA and pooled CRISPR libraries. Additionally, an experiment showed that the cleaved 20-nucleotide sequence could be ligated back into the opened plasmid. The results collected from these experiments, are described in this chapter.

Proof of Principle

To first show that a sgRNA-Cas9 complex can cleave a specific 20 nucleotide sequence, plasmid DNA containing a single construct was used. The plasmid DNA contained a single guide sequence, GCCCAATGCCTGCACTACAG, that was cloned into vector pXPR003, Addgene #52963. To design the oligo sequence of the sgRNA, a plasmid map of the construct cloned into the vector backbone was generated via SnapGene (Figure 1). The 20-nucleotide constructs sequence was then identified within the plasmid map, along with the PAM site, CGG, on the reverse compliment strand. The oligo sequence was then designed by taking the 20-nucleotide sequence on the complimentary strand of the constructs sequence working from the 5’ to 3’ direction, making sure the sequence was directly followed by the PAM site on the 3’ end.

NEBs sgRNA synthesis kit instructed to add a T7 promoter sequence, nucleotide G, and 14 nucleotides overlap sequence onto the designed oligo sequence. The T7 promoter sequence is necessary for transcription, along with the G nucleotide, that is used as the transcription start site. The T7 promoter sequence, followed by the G were added

onto the 5' end before the oligo sequence. To the 3' end after the oligo sequence, a 14 nucleotide overlap sequence was added, which would then turn into the tracrRNA during sgRNA synthesis. Once the additional sequences were added, the oligo was ordered.

The target specific 20-nucleotide oligo was synthesized into sgRNA by following the sgRNA synthesis kit instructions (Figure 2). The reaction began with the *S.pyogenes* Cas9 scaffold oligo provided in the reaction mix, annealing with the 14-nucleotide oligo sequence that had been added onto the generated oligo. The *S.pyogenes* Cas9 scaffold oligo became the tracrRNA. After annealing was complete, the DNA polymerase extended the DNA on both the 5' and 3' ends generating double stranded DNA. RNA polymerase then recognized the T7 promoter region of the double stranded DNA and began transcription at the start site. Once transcription was complete, the final generated product was RNA with both the crRNA and the tracrRNA. While the sgRNA was being actively used for reactions the tube was placed on ice. For long term storage the sgRNA was kept in a freezer set to -80°C.

To ensure that both sgRNA and Cas9 were necessary for proper cleavage of our 20-nucleotide sequence, two additional reactions were set up. One reaction was missing sgRNA and the other Cas9. Nuclease-free water was used in replacement to make sure reaction volume stayed the same. The protocol provided by NEB was used for the formation of the sgRNA-Cas9 complex. Immediately after the complex was formed, the complex was digested with the plasmid DNA containing the single sgRNA construct (Figure 3). All three reactions set up had DNA added. After digestion, all samples were heated to 65°C to in-activate the Cas9, if present.

After digestion, all three conditions were purified using AMPure beads and resuspended in nuclease-free water. This was done to help remove any buffers from previous reactions. To show that linear DNA was produced by cleavage with the sgRNA-Cas9 complex, half of the purified material for each condition was treated with Exonuclease I and III, and the other half was not treated. The original DNA was linearized by making a single cut in the plasmid with restriction enzyme NotI-HF. Once digestion was complete, sample was also purified using AMPure beads, and half of the material was treated with Exonuclease I and III. All samples treated with and without Exonuclease were run on a 1% casted agarose gel. Gel was then imaged and analyzed (Figure 4).

Lane 1 was run with original uncut DNA and lane 2 was run with original DNA digested with NotI-HF. The gel shows that when cut the DNA becomes linear and is around the expected band size of 8.3kb. Since the original uncut DNA is supercoiled it travels faster on the gel and does not show its proper size of 8.3kb. Lane 3 was run with DNA digested with NotI-HF and treated with exonuclease I and III. No band was identified in this lane, indicating that the exonuclease treatment worked on the linear double stranded DNA. Lane 4 and 5 contained the original DNA digested with the sgRNA-Cas9 complex targeting the 20-nucleotide constructs sequence present in the DNA. Comparing lane 4 to lane 2, both showed bright bands around 8.3kb indicating that the DNA was cleaved by the Cas9, resulting in linear DNA. To further show that the DNA became linear, lane 5 was treated with exonuclease I and III. No band was present in lane 5, similar to lane 3 concluding that the DNA became linear after being cleaved by Cas9. The faint band present in lanes 4 & 5 lines up with the original uncut DNA

indicating that not all of the DNA was properly cleaved by the sgRNA-Cas9 complex. The DNA ran in lanes 6 and 7 were digested with only the Cas9, and lanes 8 and 9 with only sgRNA. Lanes 7 and 9 were treated with exonuclease I and III. The bands in lanes 6 through 9 lined up with the uncut DNA, both sgRNA and Cas9 are required to cleave the DNA. Lanes 7 and 9 showed slightly higher banding compared to the non-exonuclease treated counterparts. There is currently no explanation for this. The gel results demonstrated that a specific 20-nucleotide sequence can be cleaved from DNA, when targeted by a sgRNA-Cas9 complex.

Experiment 1

Since the proof of principle was able to show successful cleavage of a 20-nucleotide sgRNA construct sequence by gel results, the next step was to see if multiple 20-nucleotide construct sequences could be removed in a pooled CRISPR library. The genome-wide Brunello library, also referred to as CP0043, contains 77,441 sgRNA constructs. This library was chosen, because the library contained the same sgRNA construct sequence used in the proof of principal. The one difference is the library's constructs are cloned into a different vector, known as pXPR023, Addgene #98293. The goal of this experiment was to target a total of eight sgRNA construct within the entire library and see if they were all able to be successfully cleaved by Cas9.

The seven additional sgRNA construct sequences were picked within the Brunello Library. Oligo sequences for the seven additional target-sgRNAs were designed and

ordered by IDT (Table 1). RNA for each targeted construct was synthesized individually using the same sgRNA synthesis kit as the proof of principal. All RNA samples were then purified using the RNA cleanup kit. All eight purified sgRNAs were pooled together by combining 3uL of each sgRNA into a microfuge tube. The tube containing the eight mixed sgRNAs had Cas9 added at a 1:1 ratio to form sgRNA-Cas9 complexes. The Brunello library was then added into the tube at a 1:10 ratio with the complex. To in-activate the Cas9, Proteinase K was added. The treated Brunello library was then transformed by electroporation using Stbl4 cells and plated onto 3 bioassays. The following day the bioassays contained bacteria lawns. However, the lawns were less dense than expected and the colony count plate indicated roughly 3 colonies per construct, which was significantly lower than expected. The bacteria on the bioassays were still collected however and the DNA was prepped by using a single maxi column and then quantified. Both the original non-treated, and sgRNA-Cas9 complex treated Brunello library were then set up for NGS PCR in replicates. A total of four wells were set up during NGS PCR and then pooled together to be submitted for one lane of Illumina Sequencing.

Once the raw sequencing data was received, the data was deconvoluted through a tool called PoolQ. Using an in-house tool called Apron, the distributions of all sgRNA constructs for each library were graphed (Figure 5). The x-axis represented the sgRNA's constructs ranked by abundance and the y-axis represented the fraction of total reads. The area under the curve (AUC) was calculated for each replicate. Both replicates had the same exact AUC value calculated, concluding that there was no huge sequencing difference of each sgRNA construct between the two replicates. An ideal perfect

distribution of sgRNAs would result in an AUC value of 0.5. Due to the large number of sgRNAs the original library had an AUC value of 0.64. The Brunello library that was made after digesting with the eight sgRNAs had an AUC value of 0.68. This indicated that the distribution of sgRNAs of the newly made library was not as good.

Using the counts file generated from PoolQ, each sgRNA construct was normalized by first summing together the read counts for each replicate together and then dividing by the total summed read counts of the library. The output number was then multiplied by 1x10e6 to show reads per million (RPM) and then the number one was added onto each RPM. This number was then log2- transformed. To calculate the log2-fold change (LFC) the log2-RPM value of the non-treated library was subtracted from the treated library (Table 2).

To show how well the eight targeted sgRNAs were cleaved, a scatterplot plotting the log2 norm of the treated vs non-treated was generated with the eight targeted sgRNAs labeled in red (Figure 6). The scatterplot did not show tight distribution of the non-targeted sgRNA constructs, concluding that the estimated 3 colonies per constructs in the treated library after transformation resulted in many of the non-targeted sgRNA constructs showing depleted read counts compared to the original library.

The LFC of each sgRNA construct within the library ranked from highest LFC to lowest. A graph plotting the LFC and ranking for each sgRNA construct was generated (Figure 7). The eight sgRNA constructs that were targeted were marked in red. The results showed that each sgRNA showed varying LFCs, indicating that cleavage of each sgRNA varied. Additionally, many sgRNA constructs that were not targeted fell within the rankings of our targeted sgRNAs. The average LFC of the eight targeted sgRNA

constructs was -2.47 and the average LFC of the entire library was -0.69 indicating highlighting that non-targeted sgRNA constructs showed read count depletion (Table 3). From this experiment the results were unable to confidently show that the targeted sgRNA constructs read counts were reduced due to proper cleavage of the 20-nucleotide sequences by Cas9. This experiment was then repeated using some modifications.

Experiment 2

A second attempt to target the same eight sgRNA constructs in the Brunello library was done with some modifications. First, the eight individual sgRNAs were re-pooled together. Next, two different conditions were set up. One condition contained the pooled sgRNAs and Cas9 to form the complexes and the second condition only contained the sgRNAs without Cas9. Both conditions then had 1uL of the Brunello library at 498ng/uL added. Each condition was then transformed by electroporation using Stbl4 cells. Instead of only transforming with one tube of cells and plating onto 3 bioassays, as in experiment 1; the reactions were scaled up. Each condition was transformed using a total of four tubes of Stbl4 and plated onto twelve bioassays to help increase the number of colonies per construct. The number of colonies in the bacteria lawn had significantly increased from experiment 1. The colony count plates this time showed that the sgRNA only treated library had 106 colonies per construct, and the sgRNA-Cas9 complex library had 101colonies per construct. Bacteria was then scrapped off each bioassay and pooled together per each condition. The bacteria culture was then thoroughly mixed and split into four conicals and pelleted. Only one pellet per condition was maxi prepped. One well

per condition was set up for NGS PCR. Both wells were then pooled together and submitted for Illumina Sequencing.

Once the raw sequencing data was received, the data was deconvoluted again using PoolQ. Apron was used again to generate the distribution graphs and to calculate the AUC's for both conditions (Figure 8). The AUC values for both the library treated with sgRNA only and with sgRNA-Cas9 complex were 0.64. Not only did both treatment conditions have the same AUC value, but the value was also identical to the non-treated Brunello library from Experiment 1, indicating better sgRNA construct distribution across the library.

Taking the counts file for both conditions the sgRNA construct data was normalized by taking the read counts for each individual construct and dividing by the total number of reads per condition. The number resulting from the fraction was then multiplied again by 1x10e6 reads per million. With the addition of the number 1, the log-2 was transformed. Because there were no replicates of each condition submitted for sequencing, there was no summing of the reads before calculating the fraction. The LFC was calculated by subtracting the sgRNA only treated library from the sgRNA-Cas9 treated library. The average LFC of the entire library was 0.0014 and the average LFC of the eight targeted constructs was -2.106. LFC (Table 3).

A scatterplot was used to again to show the log2 norm of both treatment conditions, and the eight targeted sgRNAs were labeled in red (Figure 9). The results showed a tighter distribution of the non-targeted constructs, indicating that both libraries had similar sgRNA representation. The graph nicely showed that seven out of the eight targeted sgRNAs constructs dropped out compared to the other constructs in the library.

All sgRNA constructs within the library were ranked from highest, most negative log2 value, to lowest, most positive log2 value. Six out the eight targeted sgRNA constructs were the highest ranked in LFC. The remaining two were ranked 87, and 67237 out of 77441. The construct that was ranked 67237 had also shown a very low LFC in experiment 1, indicating that the on-target efficiency of this sgRNA sequence is low. The LFC and ranking of each constructs LFC was plotted again on a graph, with the eight targeted constructs labeled in red (Figure 10). Compared to experiment 1, the eight targeted sgRNA constructs showed better read count depletion. This experiment showed that targeted sgRNA constructs could be removed up to 8-fold from the library.

Experiment 3

To determine if the 20-nucleotide sequence was able to be re-ligated back into the now opened plasmid, a T4 ligase reaction was set up using the plasmid DNA from the proof of principal. A new sgRNA-Cas9 complex was made using the same method as the proof of principal, followed by the addition of DNA at 30nM. The digested sample was then split into two reaction conditions. One condition got treated with Quick CIP to remove 5' and 3' phosphates, and the other condition was left untreated. Both tubes then had T4 buffer and enzyme added and were incubated according to NEB instructions. Both samples were then run on a 1% casted agarose gel, alongside the DNA treated with the complex, or with either sgRNA or Cas9 only to help identify the banding patterns. Gel was then imaged and analyzed (Figure 11).

Lane 1 ran with CIP treated sample showed a band around 8.3kb, indicating that the DNA after ligation remained linear and no ligation occur. This result was expected, because T4 ligation requires 5' and 3' phosphates to be present. Lane 2 was run with the non-CIP treated sample, and results showed a bright band still at 8.3kb, but also showed fainter band that was higher than 8.3 kb. The higher band in the non-CIP treated lane was compared to the three other lanes containing DNA treated with and without the complex that also showed the higher ban. Based off of knowing how DNA runs through gels depending on its current state, the higher band is believed to be relaxed circular DNA. This assumption cannot be fully proven, as these bands were not sequenced for full confirmation. Overall, the results showed that it is possible to ligate the 20-nucleotide sequence into the DNA after it had been cleaved using T4 ligase.

Chapter IV.

Discussion

Type II restrictions enzymes can recognize specific 4 to 8-nucleotide sequences within DNA and result in cleavage of that region or regions depending on how many times that sequence is present within the plasmid. Even though restriction enzymes are commonly used, they are not always the most suitable option for removal of a given sequence within a plasmid. For instance, to remove undesired sgRNA constructs present in a pooled CRISPR library, a restriction enzyme could not be used because all the sgRNA constructs are cloned into the same vector backbone. This aim of this study was to see if the CRISPR-Cas9 System could behave in a similar manner to a restriction enzyme, by using a 23-nucleotide recognition sequence to remove a 20-nucleotide sequence.

Significance of Results

To first show that the CRISPR-Cas9 system can work to remove a single 20-nucleotide sequence, a plasmid DNA containing a single sgRNA sequence was used. The goal was to remove the 20-nucleotide sgRNA sequence from the plasmid. Gel results, showed that the plasmid DNA treated with the sgRNA-Cas9 complex resulted in a linear band. The linear band showed the same size as the linearized original plasmid DNA by a restriction enzyme. These results demonstrated the sgRNA-Cas9 complex is able to cleave the DNA, resulting in linear DNA. Proof that the entire 20-nucleotide sequence was removed, was unable to be demonstrated by using gel analysis only. When DNA was treated with only Cas9 or sgRNA, the bands displayed identical sizing to the uncut original plasmid DNA, concluding that the DNA remained supercoiled. This supported the notion that both sgRNA and Cas9 are required in order to cleave the DNA.

A faint band around the same size as the supercoiled DNA was present in the lane where the DNA was digested with the sgRNA-Cas9 complex. This faint band suggests that not all the plasmid DNA in the reaction was cleaved by the Cas9. Tweaking the ratio of the input DNA to the sgRNA-Cas9 complex would likely help increase the amount of cleavage within the reaction.

Once the proof of principle showed that a sgRNA construct sequence could be removed from the plasmid DNA, eight sgRNA sequences within the Brunello Library were targeted to be removed. The eight sgRNAs were picked based on the read count abundance in the original library. Because most of the DNA in the initial library was cleaved, the same ratio of sgRNA to Cas9 to DNA was applied for this experiment.

The following day after electroporation, the colony count for the treated library was taken. The colony count calculation was way of estimating how many colonies amongst the bioassays are specific to one sgRNA construct. Colony count calculations revealed that only three colonies per sgRNA construct were present. Typically, a good colony count is around one-hundred colonies per construct or greater, which increases the odds of good distribution amongst the sgRNAs. Most likely, because this treated large library was transformed with only one tube of cells, the amount of unique sgRNAs to get into the cells was limited. Experiment still proceeded however with DNA extraction, and sequencing to look if the sgRNAs sequences were still able to be cleaved within the library.

Deconvoluted Illumina Sequencing data was run through an in-house tool called Apron to look at the AUCs for both the treated and non-treated library. The higher AUC in the treated library, 0.68 compared to 0.64 of the original library suggested that the distribution of sgRNA constructs within the treated library was lower. A scatterplot, plotting the log2 norms of both libraries did not show a tight correlation of the non-targeted sgRNA construct, indicating that read counts for all sgRNA constructs in the treated library varied from the non-treated library. These results we reflected when calculating the LFCs, which showed that the LFC values for the eight targeted sgRNAs varied greatly, with only one of the targeted sgRNAs ranking in the top 10 highest LFCs. Many of the non-targeted sgRNA sequences ranked higher for most negative LFC. Even though we were able to see that the read counts for seven out of eight of our targeted sgRNAs depleted; too many other non-targeted sequences also had depletion in read counts. The low number of colonies per sgRNA construct is reflected in the low number

of read counts for each sgRNA construct. Attempting to compare the non-treated original library that was transformed with four tubes of cells and a total of 12 bioassays to the treated library, that was transformed with only one tube of cells and 3 bioassays, appeared to be the wrong comparison.

Based off the collected results from the previous experiment, modifications were made to the following experiment. First, the library treated with the eight sgRNA-Cas9 complexes would be compared to the same library treated with only the eight sgRNAs. This was done to highlight that read count depletion in the targeted sgRNAs was due to the sgRNA-Cas9 complex and not any other outside factors. To further keep these libraries similar and to increase the amount of sgRNAs per construct, the libraries were transformed with four tubes of cells, and twelve bioassays each. This was to mimic the way the original library was made. The colony count calculations, this time showed that both transformed libraries had over 100 colonies per construct compared to our previous experiment of 3 colonies per constructs.

The deconvoluted Illumina Sequencing data was again run through Apron, and the AUCs for both libraries were 0.64. This AUC value was the same value as the original non-treated library in the previous experiment, suggesting that the distribution of sgRNA constructs within all three prepped libraries are similar. Another scatterplot was plotted with the log2 norm counts of each sgRNA construct for both the sgRNA-Cas9 complex treated library and sgRNA only treated library. The results this time compared to the previous experiment show a much tighter correlation in log2norms between the non-targeted sgRNA constructs in both libraries. The plot nicely showed that seven out of the eight targeted sgRNAs dropped out of the non-targeted sgRNAs on the graph. This

helped to prove that the sgRNA-Cas9 complex cleaved the targeted sgRNA constructs. The calculated LFC changes were again ranked, but this time six of the targeted sgRNAs were ranked in the top ten. The graph plotting the sgRNA rank compared to LFC showed that seven out of the 8 targeted sgRNA constructs were ranked high up. Additionally, this graph showed that most of the non-targeted constructs did not have a LFC greater than negative one.

The sgRNA targeted construct, AGGAAGATAGAGGAGGAACA, was unable to be cleaved in both experiments. To figure out why no cleavage occurred, the on- target score of the designed sgRNA for this construct was calculated using the Rule Set2 On-Target Scoring (Doench et al, 2016). A 0.53 score was calculated for the designed sgRNA sequence. The rule set is between 0 and 1 with anything above 0.5 generally having good on-target efficiency. This score, unfortunately did not help solve the issue of poor activity. The LFCs of sgRNAs constructs that shared at least 9 nucleotides with the original target constructs were evaluated to see if off target activity was occurring. Compared to the targeted construct, nine out of the fifteen constructs showed LFC less than 0. This could suggest some off-target activity of the sgRNA, but without further investigation cannot completely make that claim. Overall this experiment highlighted that removal of a 20-nucleotide sgRNA construct could be removed out of a pooled CRISPR library by using sgRNA-Cas9 complexes.

After sgRNA-Cas9 cleaves the 20-nucleotide region of the plasmid, the open plasmid and 20-nucleotide strand contain blunt-ends. The final experiment looked into seeing if the plasmid, and cleaved out region could be ligated back together. T4 ligase was used to bind the two blunt-ended fragments back together. A control where the

ligation could not occur was added. In order to prevent the ligation from occurring, half of the sgRNA-Cas9 complex digested DNA had CIP treatment. CIP treatment removes the phosphates off the ends of the DNA, preventing ligation from happening. After CIP treatment, both reactions had the T4 enzyme and buffer added. Results were analyzed on a gel, along with the DNA with sgRNA-Cas9 complex, sgRNA and Cas9 only to help aid in understanding the banding pattern. The banding pattern in the CIP treatment showed the linearized plasmid around 8.3kb. The non-CIP treated had a bright band around 8.3kb, indicating that not all the DNA ligated back together. A band higher than the ladder that was slightly lighter in color was also present in the non-CIP treated lane. This band is not present in the CIP treated lane, but present in the lanes containing DNA with sgRNA-Cas9 complex, sgRNA and Cas9 only. Based off of this observation, the upper expected band is likely to be relaxed circular DNA, which runs much slower in an agarose gel. All lanes contained a band around 5kb, which represented un-cleaved DNA. Results were only analyzed via gel, and therefore cannot comment on the percentage of DNA that was actually ligated back together.

Conclusion

These experiments were able to demonstrate that sgRNA-Cas9 complexes can target 20-nucleotide sgRNA sequences. Most importantly these results showed that multiple targeted sgRNA constructs present in a CRISPR library could be reduced up to 8-fold by using the CRISPR-Cas9 system as a restriction digest. This application could be very useful when certain sgRNA constructs need to be removed from an existing library.

As previously mentioned, due to errors in constructing the library certain sgRNA constructs may want to be removed. Alternatively, based on results from a pooled CRISPR screen, certain sgRNA hits may be skewing the data. If the presence of those sgRNA constructs were reduced up to 8-fold, then potentially the depletion of those unwanted constructs could help improve screening data, once assay was repeated with the updated pooled library.

Results from all experiments showed that not all of the DNA containing the 20-nucleotide sgRNA sequence was cleaved. Further testing of different ratios would need to be completed in the hopes to increase the amount of cleaved plasmid. Additionally, more studies into how the on-target score can affect removal of sgRNA constructs from libraries would need to be done. The removal of only eight sgRNA's was studied in a large library. Comparing how well sgRNA cleaves DNA, based on the size of the library could be a helpful study, because pooled CRISPR libraries vary in size from 200 oligos up to 160,000 oligos. Further looking into how many sgRNA constructs can be removed from a library at a given time would be informative as well.

Outside of just removing 20-nucleotide sgRNA sequences, seeing if the results would be similar removing a different region within a plasmid that contains a PAM sequence would be inciteful. And not only removing regions within a sgRNA plasmid, but also shRNA plasmids as well, because all plasmids contain regions that have an NGG site. This could significantly help cleave regions of DNA that restriction enzymes are unable to.

Appendix 1.

List of Tables

Table 1. Eight sgRNA oligo sequences and their 20-nucleotide target sequence.

20-nucleotide target sequence	sgRNA oligo sequence (Without PAM)
TGCAGTGGAAACACGTCTTG	CAAGACGTGTTTCCACTGCA
TCGCCAACATCACCTATGCC	GGCATAGGTGATGTTGGCGA
GCCCAATGCCTGCACTACAG	CTGTAGTGCAGGCATTGGGC
CCGTATGTGAGTAGGCCTCG	CGAGGCCTACTCACATACGG
CTCAATCTGGTCCAGCACCA	TGGTGCTGGACCAGATTGAG
GACATTGCCGTCTATCCCAG	CTGGGATAGACGGCAATGTC
AGACGGAGTAAATAGCTCTG	CAGAGCTATTTACTCCGTCT
AGGAAGATAGAGGAGGAACA	TGTTCCTCCTCTATCTTCCT

A total of eight sgRNA constructs were chosen within the Brunello Library, that displayed different read counts in the sequencing data of the library. Sequence GCCCAATGCCTGCACTACAG, was also a singular plasmid, used in the proof of principal. The reverse compliment of each target was designed and then additional sequences to help aid in sgRNA synthesis were added.

Table 2. LFCs of the 8 targeted constructs from Exp.1 and Exp.2

20-nucleotide target sequence	Log Fold Change Experiment 1	Log Fold Change Experiment 2
TGCAGTGGAAACACGTCTTG	-5.0423419	-3.5891529
TCGCCAACATCACCTATGCC	-2.9497377	-2.5932003
GCCCAATGCCTGCACTACAG	-2.7915823	-2.3563291
CCGTATGTGAGTAGGCCTCG	-4.0017322	-1.98564
CTCAATCTGGTCCAGCACCA	-0.7127583	-1.6768796
GACATTGCCGTCTATCCCAG	-2.8996441	-1.5869326
AGACGGAGTAAATAGCTCTG	-0.7807695	-0.9573703
AGGAAGATAGAGGAGGAACA	-0.5635049	0.2948561

The log2-fold Change (LFC) for each targeted sgRNA construct was calculated for both Experiment 1 and 2. Overall LFC values were similar across experiments.

Table 3. Average LFCs of each library in Experiments 1 and 2.

	Average LFC Experiment 1	Median LFC Experiment 1	Average LFC Experiment 2	Median LFC Experiment 2
8 Targeted sgRNA Constructs	-2.4677589	-2.8456132	-2.1065007	-1.98564
Whole Library	-0.6854096	-0.0896424	0.00140502	0.0005376
Whole Library minus 8 constructs	-0.6852255	-0.0896418	0.00159179	0.0005376

The average log2-fold change (LFC) of the 8 sgRNA constructs was overall higher in experiment 1 compared to experiment 2. However, the whole library LFC in experiment 1 was negative compared to experiment 2, indicating that overall read counts in the treated library were reduced in experiment 1.

Appendix 2.

List of Figures

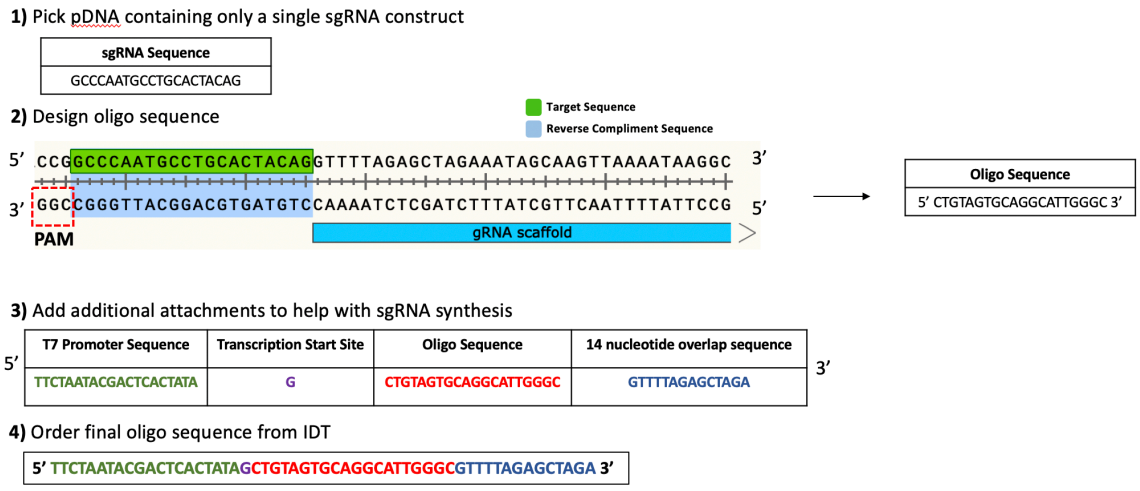

Figure 1. Schematic of sgRNA design.

A plasmid DNA (pDNA) was chosen that contained a single sgRNA construct; sequence shown above in figure. Image in step 2 was a screenshot generated by using snapgene, that shows the sgRNA sequence cloned into the vector backbone. The oligo sequence was designed by taking the 20-nucleotide reverse compliment sequence (shown in blue) of the target sequence (shown in green). Sequence was generated in the 5' to 3' of the reverse compliment strand, with the last nucleotide being next to the PAM site on the 3' end. T7 promoter sequence and transcription start site (G) were added onto the 5' end of the oligo sequence. A 14 nucleotide overlap sequence was added onto the 3' end of the oligo sequence. Final sequence was ordered through IDT. Additional 7 sgRNAs went through same workflow as sgRNA shown in schematic.

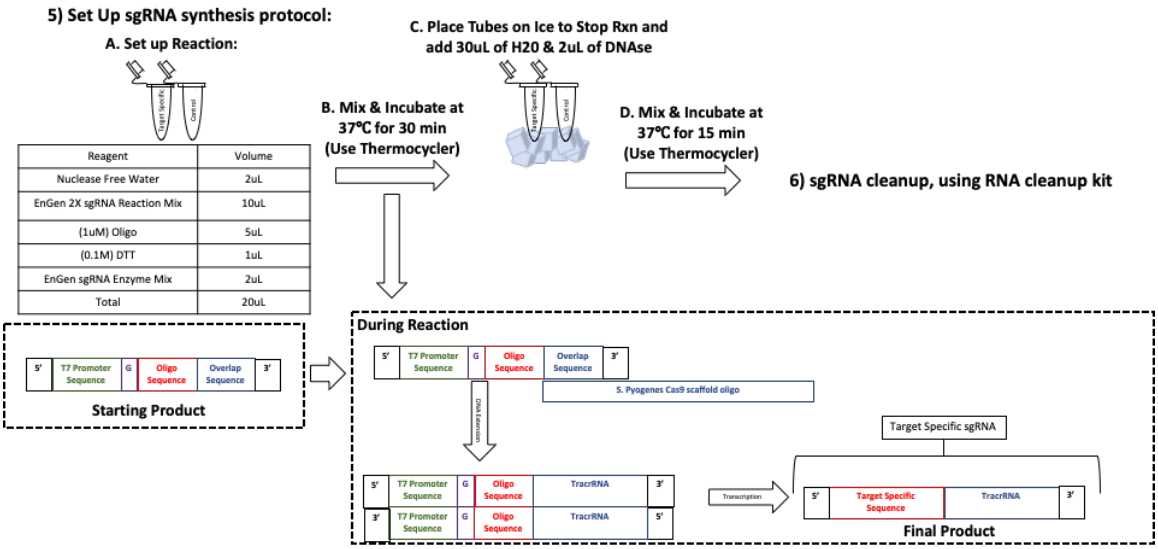

Figure 2. Schematic of sgRNA synthesis

Protocol provided in sgRNA synthesis kit (NEB) was used. Two reactions were set up, one containing the oligo we ordered and the other containing the control oligo provided in the kit. One reaction mix was set up and added to the oligos, they were mixed and then incubated at 37°C for 30 minutes in a thermocycler. During the reaction the S.Pyogenes Cas9 scaffold oligo anneals to the 14-nucleotide scaffold region of the oligo sequence. DNA polymerase will then create the DNA to become double stranded DNA. RNA polymerase then recognizes the double stranded T7 promoter sequence and begins transcription at the transcription start site. Transcription was stopped by adding DNAse to the reaction and then incubating for an additional 15 minutes at 37°C. The sgRNA was then purified with an RNA cleanup kit (NEB). The final product is sgRNA that contains only the target 20-nucleotide sequence, crRNA, and the TracrRNA.

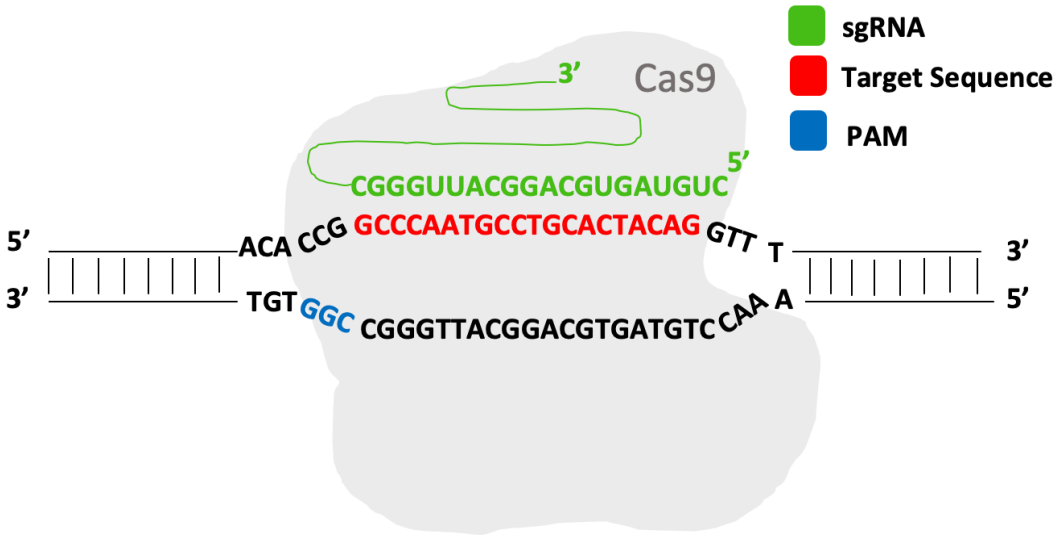

Figure 3. Representation of sgRNA-Cas9 complex binding to target sequence.

Once sgRNA-Cas9 complex has formed, plasmid DNA is then digested with the plasmid. The 20-nucleotide sgRNA sequence will guide the Cas9 to the targeted region. The Cas9 will recognize the PAM site (NGG), bind to the DNA, and cleave both strands 4 nucleotides upstream from the PAM region.

Lane	1	2	3	4	5	6	7	8	9	
1 kb+ Ladder	-	-	-	+	+	+	+	-	-	sgRNA
	-	-	-	+	+	-	-	+	+	Cas9
	+	+	+	+	+	+	+	+	+	DNA
	-	-	+	-	+	-	+	-	+	Exonuclease Treatment
	-	+	+	-	-	-	-	-	-	NotI Digest

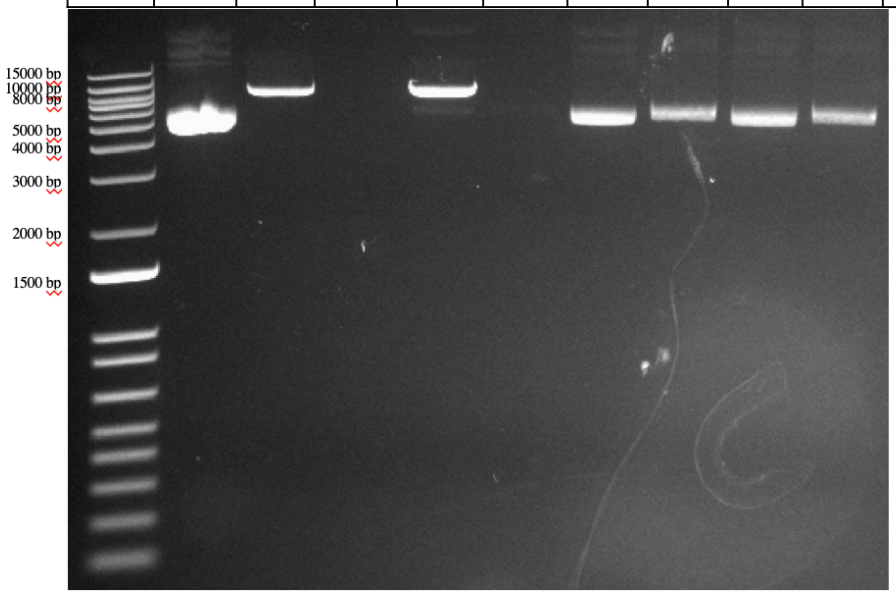

Figure 4. sgRNA-Cas9 complex cleaved 20-nucleotide within plasmid.

Lane 1 was run with supercoiled DNA and band was ~6kb. Lanes 2 and 3 were digested with NotI-HF, which resulted in linear DNA around expected band size of 8.3kb. As expected, after exonuclease treatment, no band was identified. Lanes 4 and 5 contained DNA digested with sgRNA-Cas9 complex. Lane 4 showed band ~8.3 kb, which confirmed linear DNA by cleavage. Lane 5 as expected showed no band, highlighting the DNA was indeed linear. Lanes 6 and 7 contained sgRNA and DNA only, and lanes 8 and 9 contained Cas9 and DNA only. All four lanes showed a band ~6kb, which highlights supercoiled DNA. This proved that both sgRNA and Cas9 are required to cleave the 20-nucleotide sgRNA construct.

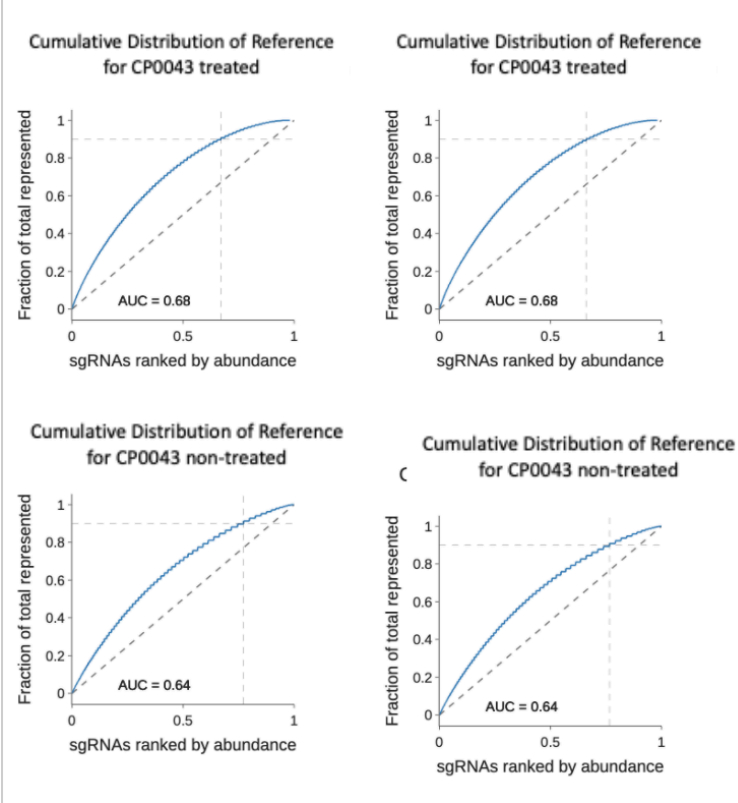

Figure 5. AUC for treated library was slightly higher than non-treated library.

Area under the curve (AUC) graphs show the two replicates for both the treated and non-treated library. The graphs display the cumulative fraction of reads per sgRNA, in relation to the fraction of reads representation. A perfect AUC is 0.5, but due to large library size the original library had an AUC of 0.64, and the treated library had an AUC of 0.68, indicating poorer sgRNA distribution when treated library was prepped.

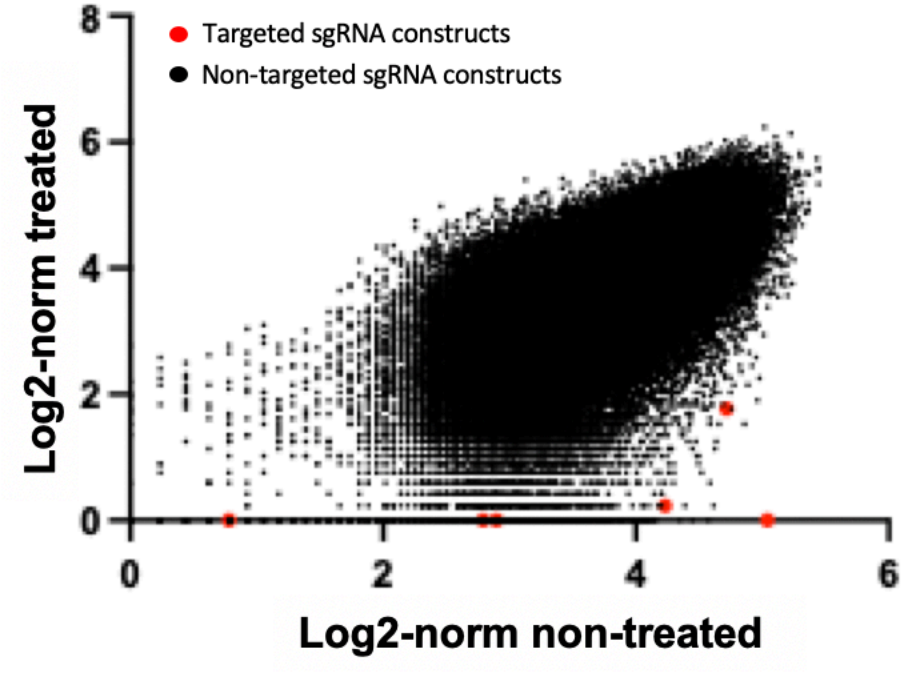

Figure 6. Both libraries show read count depletion in Experiment 1.

Calculated log2 to normalize the read counts both the non-treated and treated library were plotted. Red dots represent the 8 targeted sgRNA constructs, and the black dots represent all remaining sgRNA constructs in the library. Graph shows that both targeted and non-targeted constructs show reduced read counts in treated library.

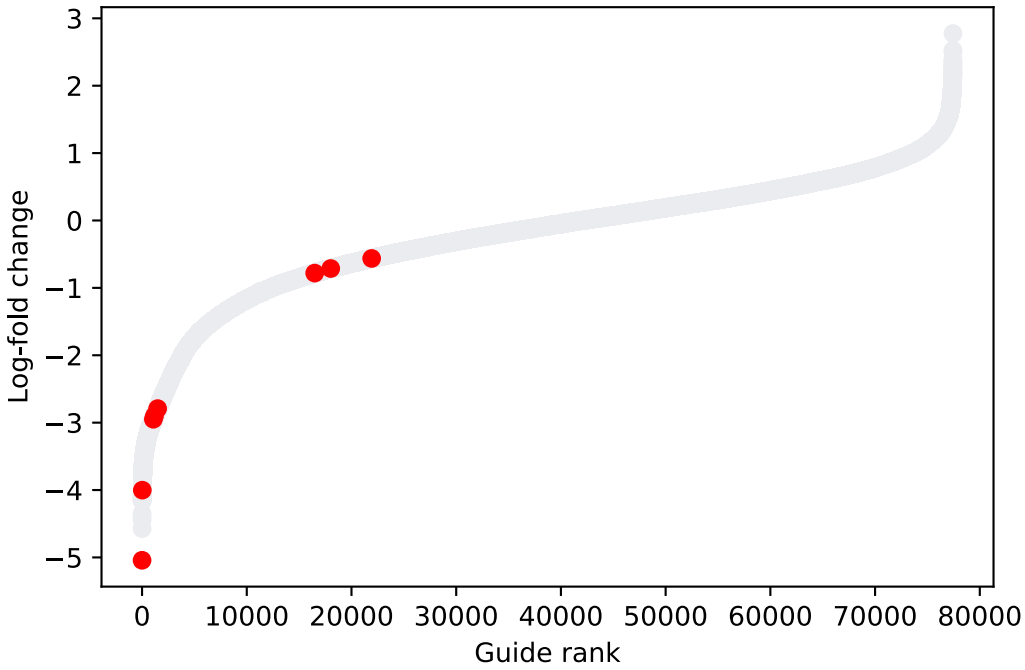

Figure 7. Targeted constructs were not top ranked across all constructs in library.

The log-fold change (LFC) was calculated for each sgRNA construct in the library and then ranked, based on highest (most negative) LFC and plotted on a graph. The 8 sgRNA constructs that were targeted are colored in red and remaining constructs in library are grey. Graph shows that most of the targeted constructs are not ranked high amongst all constructs. Additionally, results demonstrate that there are many non-targeted constructs that have negative LFC values.

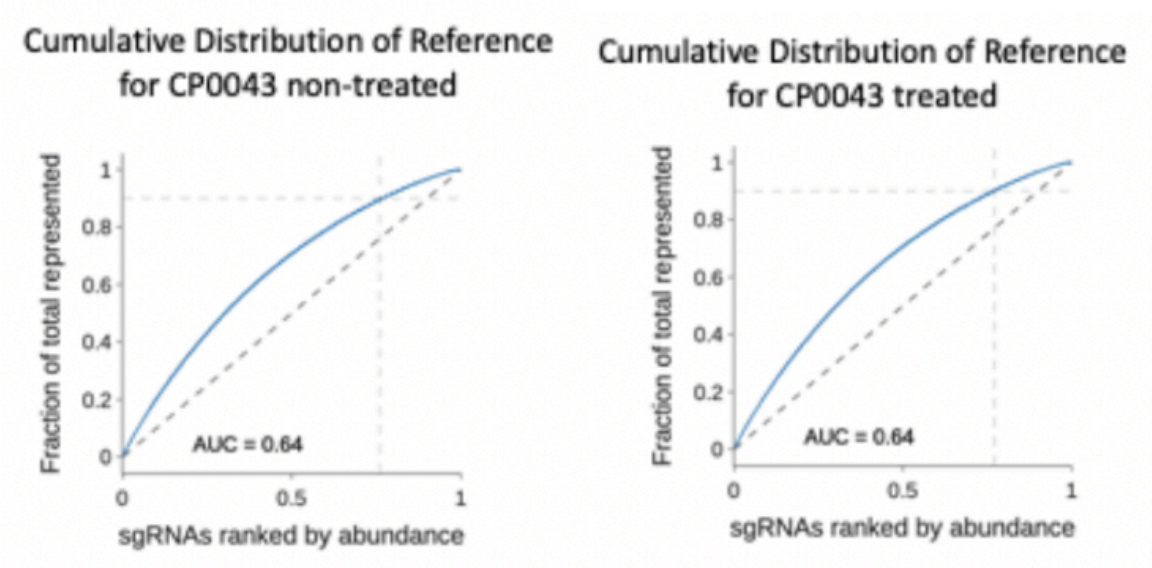

Figure 8. AUC values for both libraries, were the same for Experiment 2.

The area under the curve (AUC) calculated for both the treated and non-treated library was 0.64. This AUC value is the same as the non-treated library from experiment 1, indicating similar distribution of sgRNA constructs within both prepped libraries.

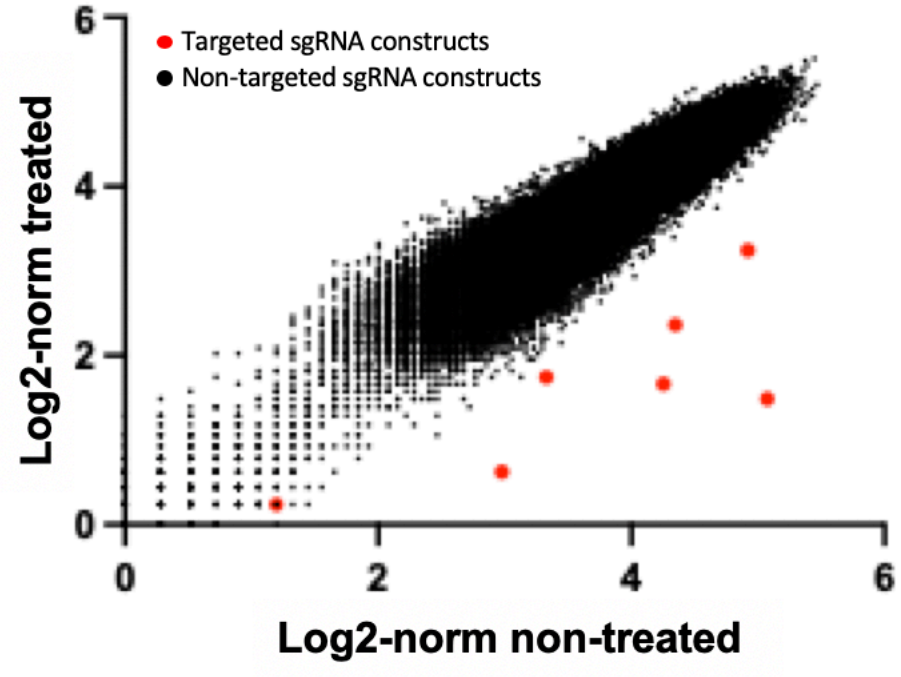

Figure 9. Eight sgRNA constructs show read count depletion in Experiment 2.

Log2 was calculated to normalize read counts for both libraries. The red dots represent the targeted sgRNA constructs, and the black dots represent the other constructs in the library. Graph shows a tighter correlation between both the treated and non-treated library for non-targeted constructs, and 7 out of 8 targeted constructs drop out indicating cleavage in treated library with appropriate sgRNA-Cas9 complexes.

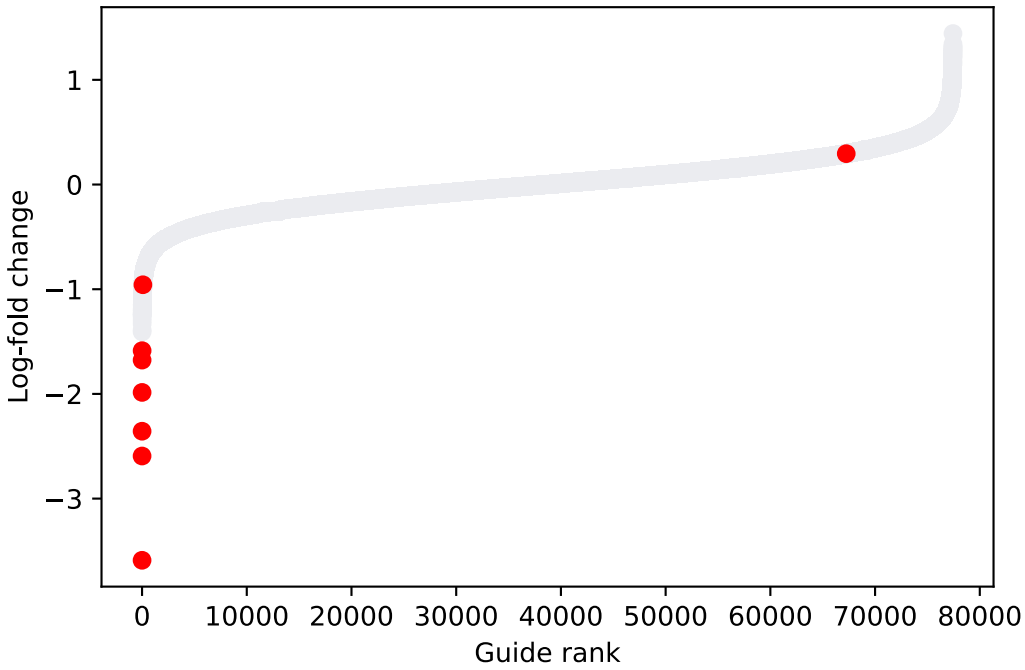

Figure 10. Six of the targeted constructs ranked in top for highest LFC

The log2-fold change (LFC) was calculated for each sgRNA construct in the library and then ranked, based on highest (most negative) LFC and plotted on a graph. The 8 sgRNA constructs that were targeted are colored in red and remaining constructs in library are grey. This graph shows that the 8 targeted sgRNA constructs ranked high compared to non-targeted constructs. Additionally, fewer non-targeted constructs had negative LFCs.

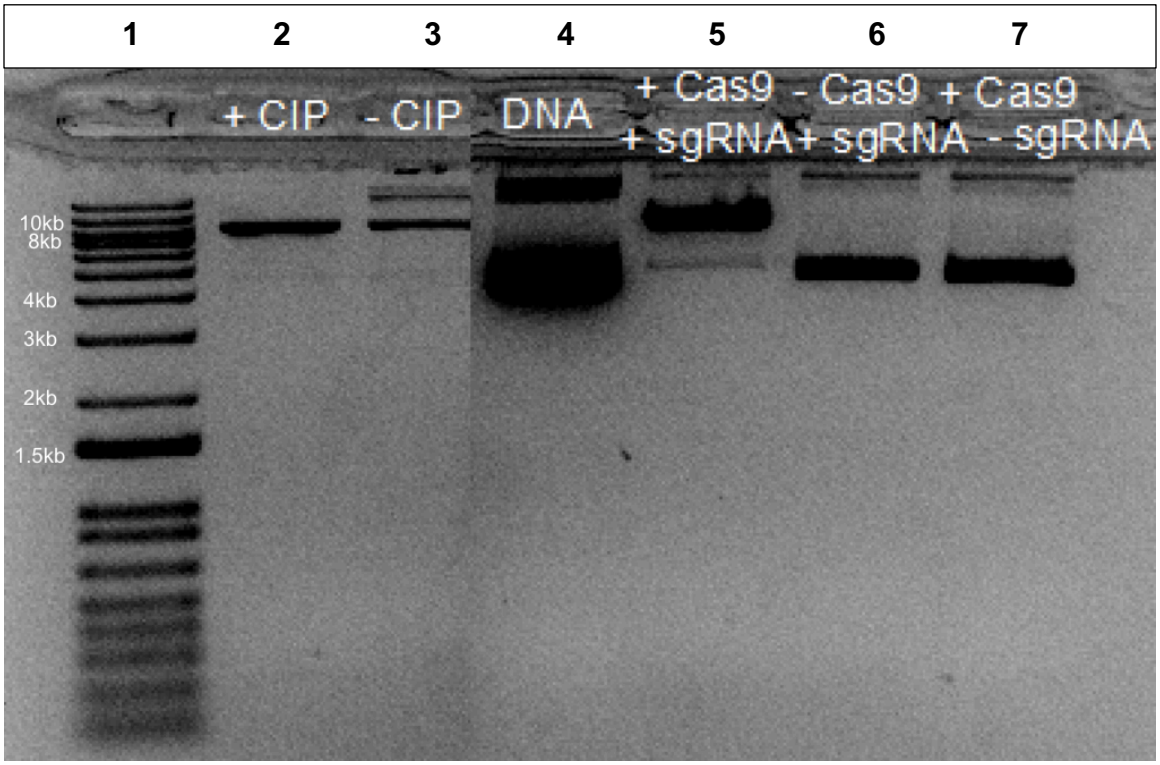

Figure 11. Partial ligation of 20-nucleotide sequence back into backbone vector.

Removing both the 3' and 5' phosphates from the linear plasmid DNA, prevented the DNA from re-ligating back together. This is observed in lane 2, where only a band ~8.3kb is present, which is indicative of the linear plasmid. Lane 3 shows a linear band at ~8.3kb, showing the DNA was not ligated. However, comparing lane 2 to lane 3, there is an additional band slighty above the ~15kb mark, that is also observed in lanes 4-7 indicating that there was partial ligation of the plasmid DNA, and most likely is relaxed circular DNA. 1kb plus ladder was run in lane 1.

References

Britannica, T. Editors of Encyclopaedia (2020) Restriction enzyme. Encyclopedia Britannica. https://www.britannica.com/science/restriction-enzyme

Broad Institute. (2019) PoolQ. Retrieved from https://gpplims.broadinstitute.org/screening/poolq/create.

Brown T.A. (2002). Studying DNA . Genomes. Oxford: Wiley-Liss. 2 (4). https://www.ncbi.nlm.nih.gov/books/NBK21129

Doench J. G. (2018). Am I ready for CRISPR? A user's guide to genetic screens. Nature reviews. Genetics, 19(2), 67–80.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., & Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature biotechnology, 34(2), 184–191. https://doi.org/10.1038/nbt.3437

Hartenian, E., & Doench, J. G. (2015). Genetic screens and functional genomics using CRISPR/Cas9 technology. The FEBS journal, 282(8), 1383–1393. https://doi.org/10.1111/febs.13248

Hu, N., Xian, Z., Li, N. *et al.* (2019). Rapid and user-friendly open-source CRISPR/Cas9 system for single- or multi-site editing of tomato genome. *Hortic Res* 6**,** 7.

Hughes, R. A., & Ellington, A. D. (2017). Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harbor perspectives in biology*, *9*(1), a023812.

Jia, G., Wang, X. & Xiao, G. (2017) A permutation-based non-parametric analysis of CRISPR screen data. BMC Genomics 18, 545. https://doi.org/10.1186/s12864-017-3938-5

Jun, S., Lim, H., Chun, H., Lee, J. H., & Bang, D. (2020). Single-cell analysis of a mutant library generated using CRISPR-guided deaminase in human melanoma cells. *Communications biology*, *3*(1), 154.

Kebschull, J. M., & Zador, A. M. (2015). Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic acids research*, *43*(21), e143.

Kelley, M. L., Strezoska, Ž., He, K., Vermeulen, A., & Smith, A. v. (2016). Versatility of chemically synthesized guide RNAs for CRISPR-Cas9 genome editing. *Journal of biotechnology*, *233*, 74–83.

Liu, Y., Tao, W., Wen, S., Li, Z., Yang, A., Deng, Z., & Sun, Y. (2015). In Vitro CRISPR/Cas9 System for Efficient Targeted DNA Editing. *mBio*, *6*(6), e01714–e1715. https://doi.org/10.1128/mBio.01714-15

Lundin S, Jemt A, Terje-Hegge F, Foam N, Pettersson E, et al. (2015) Endonuclease Specificity and Sequence Dependence of Type IIS Restriction Enzymes. PLOS ONE 10(1): e0117059. https://doi.org/10.1371/journal.pone.0117059

Piccioni, F., Younger, S.T. & Root, D.E. (2018) Pooled Lentiviral-Delivery Genetic Screens. *Current Protocols in Molecular Biology*. 121: 32.1.1-32.1.21.

Pingoud, A., & Jeltsch, A. (2001). Structure and function of type II restriction endonucleases. Nucleic acids research, 29(18), 3705–3727. https://doi.org/10.1093/nar/29.18.3705

Pray, L.A. (2008) Restriction Enzymes. Nature Education. 1 (1): 38. https://www.nature.com/scitable/topicpage/restriction-enzymes-545/#

Sanson, K.R., Hanna, R.E., Hegde, M. *et al.* (2018) Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat Commun* 9, 5416.

Scott, T., Urak, R., Soemardy, C., & Morris, K. V. (2019). Improved Cas9 activity by specific modifications of the tracrRNA. *Scientific Reports, 9*(1). doi:10.1038/s41598-019-52616-5

Sun, L., Lutz, B. M., & Tao, Y. X. (2016). The CRISPR/Cas9 system for gene editing and its potential application in pain research. Translational perioperative and pain medicine, 1(3), 22–33. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4971521/

Vincze, T., Posfai, J., & Roberts, R. J. (2003). NEBcutter: A program to cleave DNA with restriction enzymes. Nucleic acids research, 31(13), 3688–3691. https://doi.org/10.1093/nar/gkg526

Wu, X., Kriz, A. J., & Sharp, P. A. (2014). Target specificity of the CRISPR-Cas9 system. *Quantitative biology (Beijing, China)*, *2*(2), 59–70.

Zhang, Dingbo, et al. "Perfectly Matched 20-Nucleotide Guide RNA Sequences Enable Robust Genome Editing Using High-Fidelity SpCas9 Nucleases." *Genome Biology*, vol. 18, no. 1, 2017, https://doi.org/10.1186/s13059-017-1325-9.