Open access • Journal Article • DOI:10.1038/NMETH.3473

Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach — Source link

Raj Chari, Prashant Mali, Mark Moosburner, George M. Church ...+1 more authors

Institutions: Harvard University, University of California, San Diego, Wyss Institute for Biologically Inspired Engineering

Published on: 01 Sep 2015 - Nature Methods (Nature Research)

Topics: Cas9, CRISPR, Genome engineering and Genome

Title

Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach.

Journal

Nature methods, 12(9)

ISSN

1548-7091

Authors

Chari, Raj
Mali, Prashant
Moosburner, Mark
et al.

Publication Date

2015-09-01

DOI

10.1038/nmeth.3473

Peer reviewed

Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach

Raj Chari[1,5], Prashant Mali[2,5], Mark Moosburner[3] & George M Church[1,4]

We developed an *in vivo* library-on-library methodology to simultaneously assess single guide RNA (sgRNA) activity across ~1,400 genomic loci. Assaying across multiple human cell types and end-processing enzymes as well as two Cas9 orthologs, we unraveled underlying nucleotide sequence and epigenetic parameters. Our results and software (http://crispr.med.harvard.edu/sgRNAScorer) enable improved design of reagents, shed light on mechanisms of genome targeting, and provide a generalizable framework to study nucleic acid–nucleic acid interactions and biochemistry in high throughput.

RNA-guided genome engineering using the clustered, regularly interspaced, short palindromic repeats (CRISPR)-Cas system has yielded an unprecedented ability to perform site-specific editing in a variety of genomes[1,2]. Key modulators of the system include the choice of Cas9 ortholog, spacer sequence composition, sgRNA secondary structure, epigenetic status of the target locus, Cas9–guide RNA complex specificity, use of a double-strand break versus nicking modality and cell type–intrinsic factors. A comprehensive analysis of these aspects will not just enable the ideal design of targeting reagents but also shed light on the mechanisms that underlie genome targeting.

Although studies have begun to reveal some of the rules in sequence composition relating to sgRNA activity[3,4], this has been limited to small numbers of genomic loci and a single CRISPR-Cas9 system. We have developed a new *in vivo* and multiplex library-on-library methodology to assess sgRNA activity across ~1,400 genes for CRISPR-Cas systems from multiple bacterial species, both in the presence and absence of multiple end-processing enzymes. To enable simultaneous assessment across all loci, we performed two independent experiments. In the first, we incorporated the corresponding guide RNA targets into lentiviruses and transduced a population of cells at a high titer (**Fig. 1a**). These synthesized targets had identical flanking sequences, thereby enabling analysis of all synthesized target loci via a single PCR followed by high-throughput sequencing (HTS). In the second method, we did a targeted pulldown of corresponding endogenous loci and performed HTS to assay nonhomologous end joining (NHEJ) profiles (**Fig. 1b**). Both assays enable simultaneous readout across all target loci; however, the results of the former assay correspond to raw targeting rates (because the targets are integrated in expressed lentivirus inserts), whereas the latter results are modulated by underlying epigenetic status of corresponding loci.

We observed strong correlation between independent biological replicates (in HEK293T cells; **Fig. 2a** and **Supplementary Fig. 1**) and saw the expected drastic reduction in NHEJ when wild-type $Cas9_{Sp}$ (from *Streptococcus pyogenes*) was exchanged for a nickase (**Fig. 2b**). We proceeded to use this method in an additional cell line (K562; **Fig. 2c**) as well with another Cas9 ortholog from *Streptococcus thermophilus* ($Cas9_{St1}$; **Fig. 2d**). Comparing mutation rates across the same cell line, we observed a lower average rate for $Cas9_{St1}$ (by approximately sevenfold) than that for $Cas9_{Sp}$, implying the latter was in general a more robust Cas9 ortholog.

We next evaluated the role of four end-processing enzymes and endonucleases on NHEJ-associated mutation frequency using this high-throughput methodology. Earlier experiments performed over a small number of loci[5] had suggested that these enzymes improve mutation rates. Of the four end-processing enzymes we tested, we observed, on average, a 2.5-fold increase in mutagenesis rates associated with TREX2 (**Fig. 2e** and **Supplementary Fig. 2**)—largely attributed to deletions—and a 1.3-fold increase with Artemis, as well as an impact on the size of NHEJ-induced deletions observed (**Supplementary Figs. 3–5**). In contrast, TdT and DdrA exhibited a modest (~1.1-fold) decrease. In a separate experiment, however, we observed that TREX2 also increased off-target mutation rates (**Supplementary Fig. 6**), which suggests that careful sgRNA design is imperative when using TREX2 to increase on-target mutagenesis rates.

We next sought to determine whether specific motifs were enriched or lost in highly active sgRNAs compared to those with little or no activity. For both Cas9s, we identified the sgRNA sequences in the top quartile of activity in each experiment, both in the presence and absence of our four end-processing enzymes. Taking the overlaps of the top quartiles and bottom quartiles in the five experiments, we identified 133 high-activity sgRNAs and 146 low-activity sgRNAs for $Cas9_{Sp}$ and 82 and 69, respectively, for $Cas9_{St1}$ (**Supplementary Data 1**). For each set, we compiled nucleotide frequencies at each position and then compared those frequencies between the high- and low-activity sets. The most drastic difference observed was at position 20, next to the protospacer-adjacent motif (PAM), where there was a

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Bioengineering, University of California, San Diego, La Jolla, California, USA. [3]Scripps Institute of Oceanography, University of California, San Diego, La Jolla, California, USA. [4]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts, USA. [5]These authors contributed equally to this work. Correspondence should be addressed to P.M. (pmali@ucsd.edu) or G.M.C. (gchurch@genetics.med.harvard.edu).

RECEIVED 30 JANUARY; ACCEPTED 26 MAY; PUBLISHED ONLINE 13 JULY 2015; DOI:10.1038/NMETH.3473

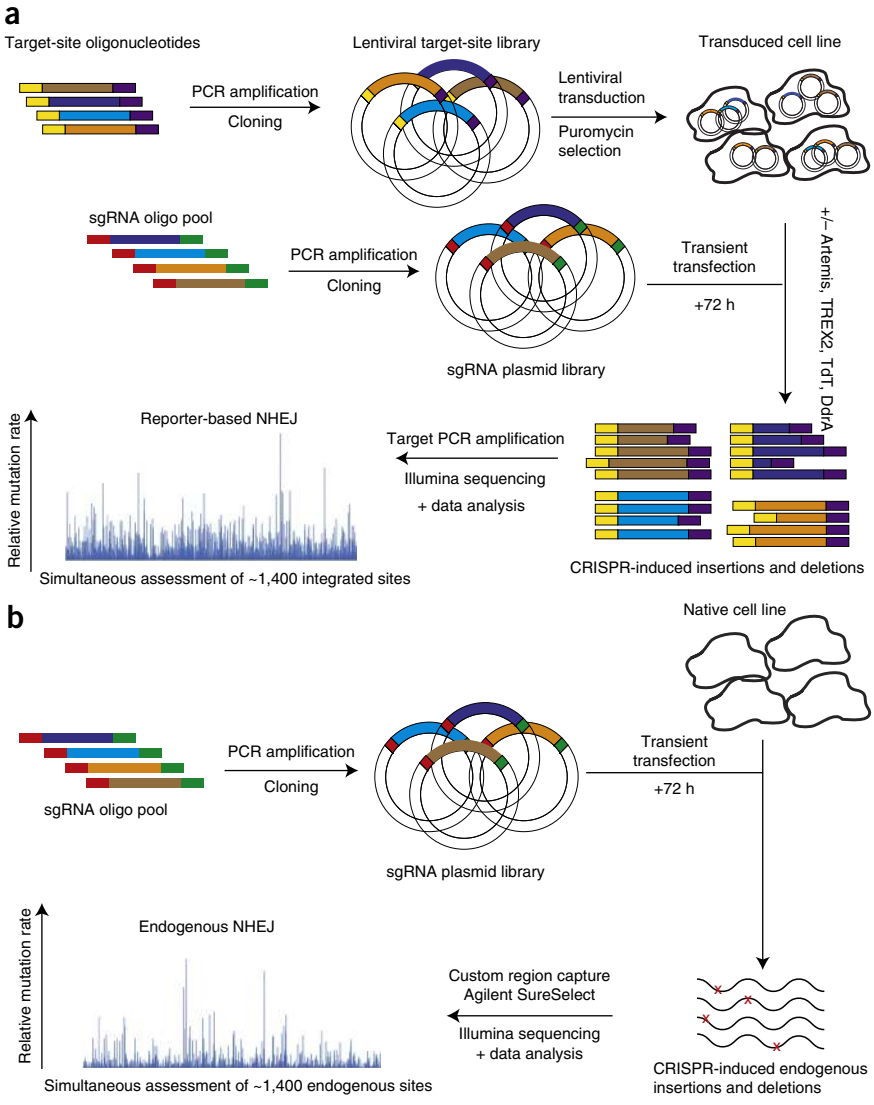

Figure 1 | Schematic of the library-on-library approach employed in our study. (**a**) sgRNA sequences corresponding to ~1,400 genomic loci were tested on ~1,400 synthesized target sites. (**b**) sgRNA plasmid library was transiently transfected into naive cells. Agilent SureSelect enrichment was performed on the genomic DNA with a custom set of probes specific to regions encompassing the ~1,400 endogenous target sites, and libraries were prepared for Illumina sequencing.

strong preference for G and low preference for T (**Fig. 3a**). Although some of these features have been observed previously, we additionally note that the G in position 20 is important to both $Cas9_{Sp}$ and $Cas9_{St1}$ (**Supplementary Fig. 7**), suggesting that this feature may be general to CRISPR-Cas9 systems.

To capture higher-order relationships between the high- and low-activity sgRNAs, we next generated a support vector machine (SVM) model. Tenfold cross-validation of our models achieved an average accuracy of 73.2% for $Cas9_{Sp}$ and 81.5% for $Cas9_{St1}$. For each Cas9, we then selected ten previously untested sgRNAs, five predicted to have high activity and five predicted to have low activity, and assessed them on an individual basis across a set of seven diverse cell lines including pluripotent stem cells (**Supplementary Table 1**). Whereas the difference in mutagenesis rates between predicted high and low sgRNAs was noticeable for $Cas9_{Sp}$, the difference was even more striking for $Cas9_{St1}$. Furthermore, this trend was observed across multiple cell types, thereby suggesting that our SVM model is generalizable (**Supplementary Figs. 8** and **9**).

A recent study using a similarly sized sgRNA library, targeting numerous sites within nine endogenous genes, had also deciphered rules that may govern sgRNA activity[3]. To assess the quality of our SVM, we scored the 1,841 sgRNA sequences used in their study and observed a modest correlation between our predicted scores and their data (**Supplementary Fig. 10**). The observed variability is somewhat expected as (i) the set of sgRNA sequences used to build the model was different, (ii) we assessed impact within 72 h as opposed to 2 weeks, which may limit selection bias, and (iii) our study used a direct sequence-based readout to assay sgRNA activity, whereas the other study employed a phenotype-based readout.

Because lentiviral integrations are generally in transcriptionally active regions, the use of an integrated-target library allows for an interrogation of sgRNA activity that is as purely sequence-composition based as possible. However, when endogenous loci themselves are targeted, this is typically not the case. In a parallel experiment, we transfected our $Cas9_{Sp}$ and $Cas9_{St1}$ sgRNA libraries in target-naive 293T cells, enriched for sequence surrounding all of our target sites and performed HTS. Unsurprisingly, the overall mutagenesis rates were markedly lower than what we had observed using our integrated-target library for both Cas9s (**Supplementary Fig. 11**).

Previous studies using $Cas9_{Sp}$ immunoprecipitation experiments have observed binding preferences of $Cas9_{Sp}$ to areas of higher DNA accessibility[6,7]. To evaluate accessibility in our system, we obtained DNase I hypersensitivity data for the HEK293T cell line from the Encyclopedia of DNA Elements (ENCODE)[8] and compiled DNase-seq values for each site. Comparing the DNase-seq values of the regions encompassing the top quartile of sites with the highest percentage of mutant NHEJ with the values of the bottom quartile, we observed a wider range of values in the top quartile (**Fig. 3b**). In addition, examining histone 3 lysine 4 (H3K4) trimethylation status, a histone mark associated with actively transcribed genes[9], we also saw a similar statistically significant enrichment (**Fig. 3c** and **Supplementary Fig. 6**). Intriguingly, we observed a small set of outliers in the group of sites with low mutagenesis rates, which showed high DNase-seq values, and when we employed our classifier on the corresponding sgRNA sequences, our classifier determined that 76.7% (23/30) of these sequences would be considered poor (**Fig. 3d**). Taken together, the results suggest that both

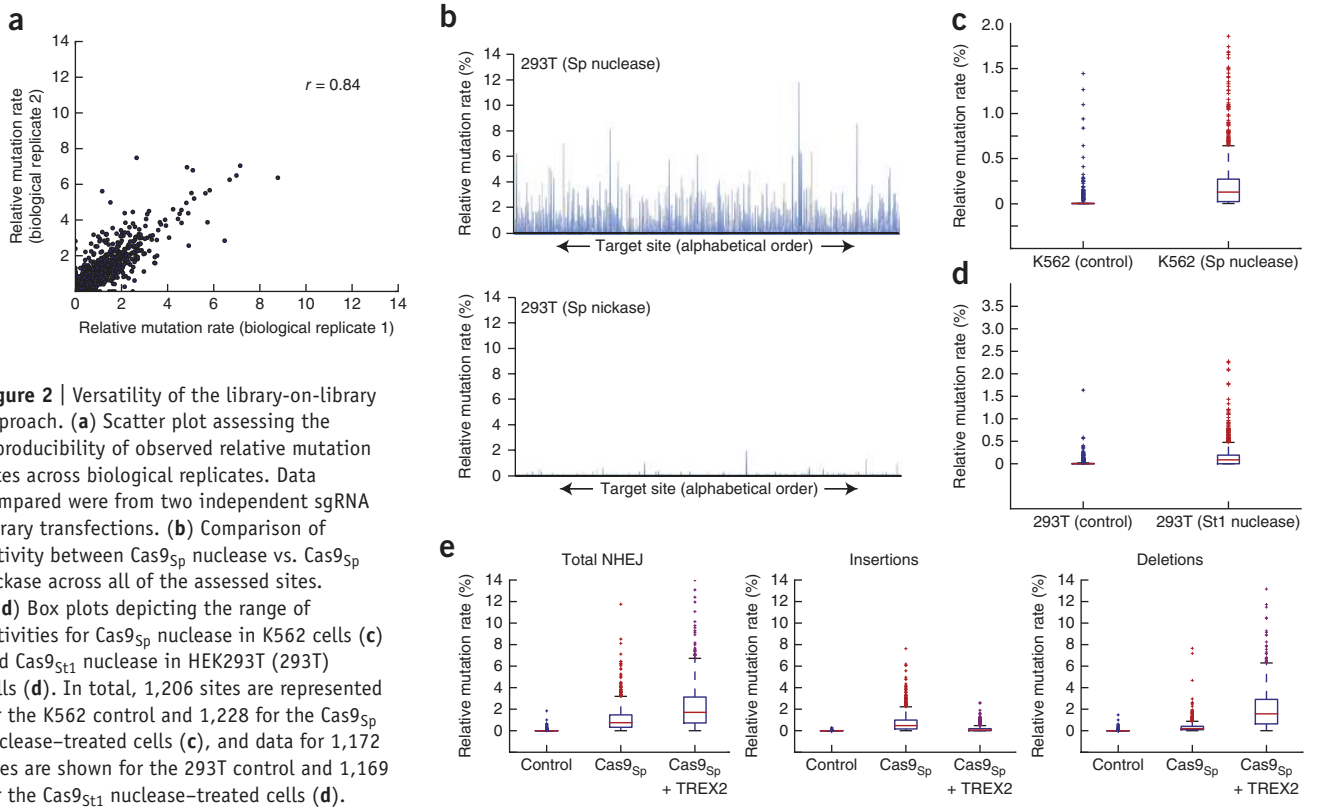

Figure 2 | Versatility of the library-on-library approach. (**a**) Scatter plot assessing the reproducibility of observed relative mutation rates across biological replicates. Data compared were from two independent sgRNA library transfections. (**b**) Comparison of activity between $Cas9_{Sp}$ nuclease vs. $Cas9_{Sp}$ nickase across all of the assessed sites. (**c**,**d**) Box plots depicting the range of activities for $Cas9_{Sp}$ nuclease in K562 cells (**c**) and $Cas9_{St1}$ nuclease in HEK293T (293T) cells (**d**). In total, 1,206 sites are represented for the K562 control and 1,228 for the $Cas9_{Sp}$ nuclease–treated cells (**c**), and data for 1,172 sites are shown for the 293T control and 1,169 for the $Cas9_{St1}$ nuclease–treated cells (**d**). (**e**) Impact of the end-processing enzyme TREX2 on patterns of NHEJ. Data shown are for 293T cells and the $Cas9_{Sp}$ nuclease. For all box plots, the height of the boxes represents the interquartile range (IQR) of the data. Whiskers are drawn to 1.5 × IQR. Red horizontal lines represent the median. '+' represents an outlier data point that is beyond 1.5 × IQR.

locus accessibility and sequence composition of the sgRNA are important in determining sgRNA activity.

We then sought to directly compare the mutagenesis rates at the lentiviral target sites with those at the endogenous sites. Using those sites which had sufficient coverage in both sets of experiments, we found a strong correlation (r = 0.422, $P = 1.6 \times 10^{-53}$). Taking the DNase I values calculated above, we sorted the endogenous mutation rate data set from most to least accessible and performed correlations of the top 100, 200, etc., up to 1,000 most accessible sites to account for the accessibility. As we enriched for more accessible sites, our correlations continually improved (**Supplementary Fig. 12**), suggesting that lentiviral target sites are highly accessible and our parallel lentiviral scheme is important to extract the true underlying sequence features associated with sgRNA activity.

A number of studies have assessed issues related to the CRISPR-Cas9 specificity, from identifying the breadth of the

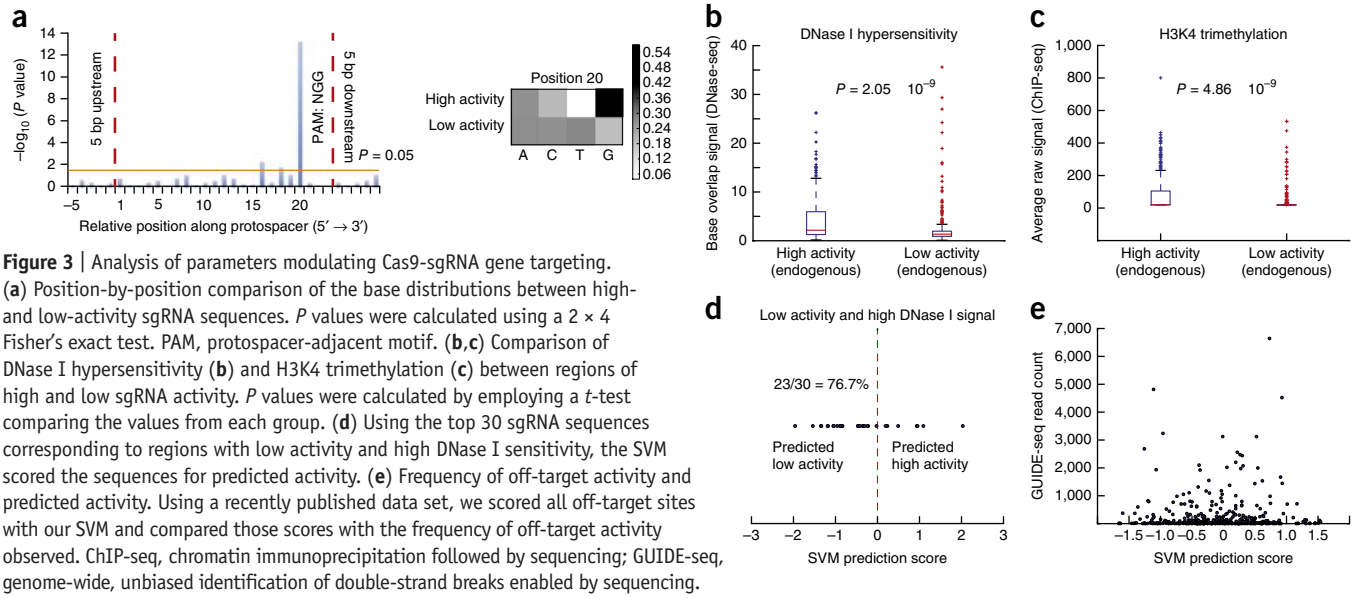

Figure 3 | Analysis of parameters modulating Cas9-sgRNA gene targeting. (**a**) Position-by-position comparison of the base distributions between high- and low-activity sgRNA sequences. P values were calculated using a 2 × 4 Fisher's exact test. PAM, protospacer-adjacent motif. (**b**,**c**) Comparison of DNase I hypersensitivity (**b**) and H3K4 trimethylation (**c**) between regions of high and low sgRNA activity. P values were calculated by employing a t-test comparing the values from each group. (**d**) Using the top 30 sgRNA sequences corresponding to regions with low activity and high DNase I sensitivity, the SVM scored the sequences for predicted activity. (**e**) Frequency of off-target activity and predicted activity. Using a recently published data set, we scored all off-target sites with our SVM and compared those scores with the frequency of off-target activity observed. ChIP-seq, chromatin immunoprecipitation followed by sequencing; GUIDE-seq, genome-wide, unbiased identification of double-strand breaks enabled by sequencing.

off-target activity problem to proposing bioinformatic and methodological solutions[10–17]. Using a recently published off-target data set study[17], we examined the relationship between specificity and activity and observed no tangible relation (**Fig. 3e**). This result highlights a need to account for both specificity and activity in the sgRNA design process. To this end, we have compiled lists of human and mouse exome-wide $Cas9_{Sp}$ and $Cas9_{St1}$ target sites that are predicted to be both highly active and specific, determined by the CasFinder algorithm[18] (**Supplementary Data 2–5**). We have also made available a software package that can identify and/or score sgRNA sites, with user-defined sequences as input (**Supplementary Software**; http://crispr.med.harvard.edu/sgRNAScorer). Overall, our *in vivo* library-on-library approach enables facile assaying of nucleic acid–nucleic acid interactions in high throughput and can be readily extended to assay other modulators of homologous recombination or NHEJ pathways and newly identified CRISPR-Cas systems.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. NCBI Sequence Read Archive: all sequencing data generated in this study are deposited have been deposited under accession SRP048540.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We acknowledge J. Aach for help with CasFinder and useful discussion, B. Turczyk for help with custom array oligonucleotide synthesis, S. Byrne (Harvard Medical School) for providing PGP1 induced pluripotent stem cells, and A. Chavez for useful discussion. This work was supported by US National Institutes of Health grant P50 HG005550. R.C. was supported by a Banting Fellowship from the Canadian Institutes of Health Research. P.M. is supported by University of California, San Diego, startup funds and a Burroughs Wellcome Career Award at the Scientific Interface.

AUTHOR CONTRIBUTIONS

R.C. and P.M. designed the study and performed the experiments. R.C. and P.M. wrote and edited the manuscript. All authors approved the final version of the manuscript. R.C. implemented custom Python software and performed data analysis. M.M. provided technical assistance. G.M.C. supervised the project.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Mali, P. *et al. Science* **339**, 823–826 (2013).
2. Cong, L. *et al. Science* **339**, 819–823 (2013).
3. Doench, J.G. *et al. Nat. Biotechnol.* **32**, 1262–1267 (2014).
4. Gagnon, J.A. *et al. PLoS ONE* **9**, e98186 (2014).
5. Certo, M.T. *et al. Nat. Methods* **9**, 973–975 (2012).
6. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. *Nat. Biotechnol.* **32**, 677–683 (2014).
7. Wu, X. *et al. Nat. Biotechnol.* **32**, 670–676 (2014).
8. ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
9. Koch, C.M. *et al. Genome Res.* **17**, 691–707 (2007).
10. Ran, F.A. *et al. Cell* **154**, 1380–1389 (2013).
11. Mali, P. *et al. Nat. Biotechnol.* **31**, 833–838 (2013).
12. Tsai, S.Q. *et al. Nat. Biotechnol.* **32**, 569–576 (2014).
13. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. *Nat. Biotechnol.* **32**, 279–284 (2014).
14. Fu, Y. *et al. Nat. Biotechnol.* **31**, 822–826 (2013).
15. Guilinger, J.P., Thompson, D.B. & Liu, D.R. *Nat. Biotechnol.* **32**, 577–582 (2014).
16. Pattanayak, V. *et al. Nat. Biotechnol.* **31**, 839–843 (2013).
17. Tsai, S.Q. *et al. Nat. Biotechnol.* **33**, 187–197 (2015).
18. Aach, J., Mali, P. & Church, G.M. Preprint at *bioRxiv* doi:10.1101/005074 (2014).

ONLINE METHODS

Selection of genes and target sites. A list of genes that would be considered of high value, encompassing ion channels, receptors and genes in the cancer gene census[19] was first derived. Next, using the hg19 RefSeq annotations for each gene, we downloaded the exon sequences with 75 nt of flanking sequence from the UCSC Table Browser[20]. Custom Python scripts were written to identify all unique Cas9 *S. pyogenes* ($Cas9_{Sp}$, $N_{20}NGG$) and *S. thermophilus* ($Cas9_{St1}$, $N_{20}NNAGAAW$) sites with the exons. These sites were then aligned against entire Hg19 genome sequence using SeqMap[21], and the sites (i) with no 3-nt off-targets in the genome and (ii) targeting the 5′-most exon were retained. In total, sites were successfully generated for 1,362 genes for $Cas9_{Sp}$ (**Supplementary Data 6**) and 1,449 for $Cas9_{St1}$ (**Supplementary Data 7**), with one site per gene. Sequencing of the original plasmid libraries revealed fairly uniform representation of both targets and sgRNAs (**Supplementary Fig. 13**). For the target sites, 1,344/1,362 of the $Cas9_{Sp}$ target sites and 1,276/1,449 of the $Cas9_{St1}$ were detected from this analysis. For the sgRNA libraries, we were able to detect 100% of the sgRNA sequences that we had synthesized for both Cas9s from the sequencing analysis.

Target-library synthesis. For each target site, we synthesized the target site as part of larger sequence in a 170-base-long oligonucleotide using an in-house CustomArray machine. Within each 170-base sequence, the flanking 25-base sequences were orthogonal sequences suitable for specific PCR amplification[22]. The 10 nt adjacent to each primer were given a unique 10-base barcode sequence that was at least 2 units in Hamming distance away from any other 10-base barcode used. Finally, for the remaining 100 bases, the center of this sequence encompassed the 23-base target site (27 bases for $Cas9_{St1}$) and 39 and 38 bases of endogenous flanking sequence on each side for the $Cas9_{Sp}$ target site (37 and 36 bases of sequence for $Cas9_{St1}$) (**Supplementary Fig. 14**). Sequences were amplified and cloned into a lentiviral vector (Addgene #26777) for subsequent integration experiments. Plasmid libraries were sequenced and assessed for mutations introduced by synthesis errors. These mutations were subsequently used to filter out synthesis-related errors from bona fide mutations observed in experimental samples. Synthesized oligonucleotide sequences are shown in **Supplementary Data 8** and **9** (for $Cas9_{Sp}$ and $Cas9_{St1}$, respectively).

sgRNA library synthesis. For each target site generated, the corresponding sgRNAs were synthesized on CustomArray at the same length as the target site (**Supplementary Fig. 14**). Additional flanking scaffold sequence was added to the synthesized protospacer to enable Gibson assembly–mediated cloning into the destination vector (Addgene #24150). The general methodology and scaffolds used for $Cas9_{Sp}$ and $Cas9_{St1}$ were as previously described[1,23]. sgRNA libraries have been deposited in Addgene.

Library-on-library experiments. Cells were first transduced with a high titer of the lentivirus target library to create a pool of cells bearing the target library. Next, enrichment of successful target integration was performed using puromycin selection (2 μg/mL). These transduced cells (10^6) were then transfected with Cas9 (5 μg), corresponding gRNA library (5 μg) and end-processing enzyme/empty vector (5 μg). Plasmids encoding TREX2 (#40210), Artemis (#40211) and empty backbone (#39991) were obtained from Addgene. Sequences for DdrA and TdT were obtained from previous studies[5,24]. Given the high probability of a single cell achieving multiple NHEJ events owing to the presence of multiple target sites, DNA from cells were harvested 72 h after transfection to minimize loss of cells with large amounts of nuclease activity. For 293T cells, lipofection (using Lipofectamine 2000, Invitrogen) was used; and for K562s, Nucleofection using the 4D Nucleofector (Lonza) was used to deliver the DNA. Cell lines used were obtained from ATCC and tested for mycoplasma.

High-throughput sequencing library preparation. DNA was extracted using Qiagen DNeasy kit with RNase treatment. In order to add the appropriate adaptors for HTS, we prepared libraries in two consecutive PCR reactions. The first was to retrieve the site from genomic DNA, and the second was to add the appropriate adaptor sequences for Illumina sequencing.

To ensure minimal variability due to DNA sampling, we performed 8–10 100-μL PCR reactions for each sample using 1 μg of DNA in each PCR reaction. Briefly, in each reaction, 50 μL of KAPA 2× HiFi ready mix with 3 μL of each primer and 1.2 μL of Sybr green were added together with the appropriate amount of water and template to total 100 μL. Quantitative real-time PCR (qPCR) was used to ensure libraries were not amplified past the linear amplification phase, with PCR reactions terminated accordingly. For each sample, multiple PCR reactions were pooled and then PCR purified with Qiagen columns and then eluted in 40 μL of elution buffer. Owing to the presence of primer dimers, half of each eluate was run on an Invitrogen 2% E-gel EX for 10 min and then subsequently gel purified and extracted using Qiagen columns. In parallel, primers designed to capture the sgRNA sequences were also employed, and in this case, as just the prevalence of each guide was needed, only one 100-μL reaction was used for each sample.

For the second PCR step, 6- to 25-μL PCR reactions were set up for samples. In this case, 12.5 μL of KAPA 2× HiFi ready mix, 0.75 μL of each primer and 0.3 μL of Sybr green were added together. Real-time qPCR was once again used to ensure libraries were not amplified for too long, and in this case, no more than seven cycles were performed. PCR reactions were pooled and purified using Qiagen columns, and these purified PCRs were gel purified as well. Quantification of DNA was performed using Qubit with the high-sensitivity assay. Equal amounts of each sample were then pooled into one tube and sequenced on an Illumina HiSeq 2500 on the rapid run mode with paired-end 150-bp reads; this sequencing was done at the Biopolymers Facility in the Department of Genetics at Harvard Medical School. For the K562 control and treated samples, libraries were prepared the exact same way but were sequenced on an Illumina MiSeq at the Molecular Biology Core Facility at Dana-Farber Cancer Institute using the paired-end 150-bp mode.

For the custom capture sequencing of endogenous targets, probes were designed flanking 200 bp on each of the target sites for Agilent SureSelect and synthesized by the Beijing Genome Institute (BGI). Extracted DNA was then sent to Hong Kong for capture enrichment and sequencing on one lane in an Illumina HiSeq. All data have been deposited in the sequence read archive (SRA) under accession number SRP048540.

Illumina sequencing data processing: target-site analysis. FASTQ files were first desegregated into each sample using in-house–developed Python scripts. For each sample, each set of read pairs were first merged using FLASH[25] into one larger contig and subsequently aligned to the custom reference sequences that were originally designed using BWA[26]. Next, alignments were filtered for uniqueness (no other alignments to any other sequence) and for length (136 bp). 136 bp was chosen as this is the minimum size guaranteed to cover the entire 100-bp payload sequence of our target site. Finally, SAMtools was then used to convert SAM to BAM files and generate pileups[27]. The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.

Illumina sequencing data processing: endogenous-site analysis. FASTQ files obtained from BGI were then mapped to the entire hg19 genome using BWA-MEM in the BWA package[26]. Reads were filtered to those that were successfully mate paired as well as not having multiple hits in the genome. As above, SAMtools was used to obtain pileups from the generated SAM files[27].

Calling NHEJ-associated mutations. Given the prevalence of errors in synthesized oligonucleotides from the custom array, the original target-site plasmid library was sequenced and assessed for insertions and deletions (indels). These indels and those found in the untreated sample were subsequently used to filter indels observed in the treated samples to ensure indels called were truly due to the Cas9 and sgRNAs. Custom Python scripts (available upon request) were used to analyze pileups to generate list of mutated sequences per site, the total number of mutations observed, the types of mutations observed and the total coverage at each. Single-base substitutions were excluded from the analysis owing to the high degree of false positives. Only mutations that spanned any part of the target site were considered. A minimum of 100 mapped reads at a given site was required in order to be considered for mutational analysis. Box plots and scatter plots were produced using Matplotlib[28]. Raw mutation rates observed in 293T cells for $Cas9_{Sp}$ nuclease, $Cas9_{Sp}$ nickase and $Cas9_{St1}$ are provided in in **Supplementary Data 10–12**. Rates observed in K562 with $Cas9_{Sp}$ nuclease are provided in **Supplementary Data 13**.

Sequence motif analysis and support vector machine (SVM) model generation. For each Cas9, the base distributions between the high-activity and low-activity sgRNA sequences were compared using a 2 × 4 Fisher's exact test in R. Five base pairs of sequence on each side of the target site were also analyzed. Because bases at positions 22 and 23 for $Cas9_{Sp}$ and positions 23–26 for $Cas9_{St1}$ are fixed, these positions were excluded in the calculation. *P* values generated were corrected for multiple-hypothesis testing using the Benjamini-Hochberg formula.

To build the SVM classifier model, we encoded the same sets of sequences using a 4-bit binary scheme. The model was generated using SVM^{Light} (ref. 29). To generate the SVM classifier, we categorized the high-activity group of sgRNAs as "+1" and the low-activity group as "–1." For each 23-bp sequence (27 bp for $Cas9_{St1}$), we encoded each character using a 4-bit binary system. For A, the encoding was 0001; for C it was 0010; for T it was 0100 and for G it was 1000. This was done to ensure the distance between any two bases would be equal. Each string of 1s and 0s, one per sequence, was then used as the input for the SVM. Tenfold cross-validation was used to assess the classifier, and then the entire set was used to generate the model that was employed on the list of highly specific sgRNAs obtained from CasFinder[18].

Comparison of the derived SVM model with previously published literature. A supplementary table detailing the observed sgRNA activity and predicted score for 1,841 sgRNA sequences was obtained from a recently published study[3]. Sequences were classified using our model, and raw SVM prediction scores were determined. These scores were plotted against the predicted score by the published model as well as observed sgRNA activity, and a Pearson correlation coefficient was calculated using SciPy. Plots were done using Matplotlib.

Epigenetic data analysis. DNase-seq (GSM1008573) and H3K4 trimethylation (GSM945288) data, generated as part of ENCODE[8], were downloaded from the UCSC Genome Browser[30]. For each site, 225 bp of flanking sequence were used on each side, totaling a region of 473 bp in size for $Cas9_{Sp}$ and 477 bp in size for $Cas9_{St1}$. Data for each region were extracted using the bigWigAverageOverBed tool[31]. Distributions of values were compared using a *t*-test with an unequal variance assumption. The stats module in SciPy was used to perform the *t*-test.

Analyses of specificity and activity. For examining the relationship between specificity and activity, a recently published study whereby a genome-wide analysis of double-strand DNA breaks in the presence of Cas9 and an sgRNA was used[17]. For each observed site, the target sequence was scored with our SVM classifier and a scatter plot was generated comparing the SVM scores versus the number of reads obtained in the GUIDE-seq study.

The lists of target sites in **Supplementary Data 2** and **3** corresponding to $Cas9_{Sp}$ in human and mouse were obtained from http://arep.med.harvard.edu/CasFinder/. Next, the list of 2,712,189 sites identified previously for human and 2,733,854 sites for mouse were processed through the $Cas9_{Sp}$ SVM classifier, prediction scores were determined using SVM^{Light} (ref. 29) and a distribution of these scores was created. Finally, each site generated by CasFinder[4] was then scored with the SVM, and its percentile rank in distribution of scores was calculated and reported.

For **Supplementary Data 4** and **5**, because the PAM for $Cas9_{St1}$ was different from the one used by CasFinder initially, CASValue (part of CasFinder) was rerun with default parameters using a PAM of NNAGAAW on the list of 376,758 sites for human and 350,497 for mouse. In total, CASValue deemed 195,861 human sites and 209,858 mouse sites as highly specific. As with $Cas9_{Sp}$, a score distribution was generated from the larger list sites, and the percentile rank of the CASValue site was calculated and reported.

Analysis of TREX2 on off-target mutagenesis. Three sgRNAs with known off-target sites were used for this analysis[14]. For each on-target site, three off-target sites were assessed for NHEJ-induced mutagenesis. sgRNAs were cloned into the pLKO.1 backbone. Primer sequences and target sites are listed in **Supplementary Table 2**.

doi:10.1038/nmeth.3473

500,000 293T cells were seeded in each well of a six-well plate. Approximately 24 h later, 2 µg of sgRNA and 2 µg of $Cas9_{Sp}$ were cotransfected using Lipofectamine 3000 at a ratio of 2:1. For the wells that included TREX2, 2 µg of pExodus-TREX2 were included as well. 72 h after transfection, cells were harvested and DNA was extracted with 200 µL of Quick Extract solution.

5 µL of extracted DNA were then used as template in a 100-µL KAPA HiFi reaction, and target sequences were amplified using qPCR. PCR products were then purified using SPRI-bead purification and subsequently used as template for a second round of PCR to add Illumina adaptor sequences. Final products were run on a 2% E-gel EX and then extracted and purified using the Zymoclean gel DNA recovery kit. All final PCR products were barcoded and pooled for Illumina sequencing analysis on the Illumina MiSeq. Paired-end sequencing data generated from the MiSeq were first merged using FLASH and then mapped to amplicon sequences using BWA-MEM. Pileups were generated using SAMtools, and NHEJ-induced mutagenesis was determined using custom Python scripts. To minimize the influence of Illumina error rates, we excluded single-base substitutions and included only insertion and deletion events.

Validation of predicted sgRNA activity. In total, 20 independent target sites were selected for individual sgRNA targeting. The 20 sites corresponded to 5 sites predicted for high activity and 5 sites predicted for low activity for both $Cas9_{Sp}$ and $Cas9_{St1}$ and were obtained from **Supplementary Data 2** and **4**, respectively. sgRNAs were cloned into the pLKO.1 backbone. Target-site sequences and primers used to amplify regions encompassing target sites are listed in **Supplementary Data 11**.

As in the TREX2 experiments, 500,000 293T cells were seeded, and approximately 24 h later, 2 µg of sgRNA and 2 µg of $Cas9_{Sp}$ (or 2 µg of $Cas9_{St1}$) were cotransfected using Lipofectamine 3000 at a ratio of 2:1, and DNA was harvested 72 h after transfection. DNA extraction, library preparation and data analysis were performed exactly the same as in the TREX2 experiments.

For experiments in A549, U2OS, HepG2 and SK-NAS cell lines, cells were seeded in a 24-well plate, and for each sgRNA, 500 ng of sgRNA and 500 ng of Cas9 plasmid were cotransfected using Lipofectamine 3000 at a ratio of 2:1 and DNA was harvested 72 h after transfection. For K562 and PGP1 induced pluripotent stem (iPS) cells, DNA was transfected using the Lonza 4D Nucleofector. These cell lines were not tested for mycoplasma contamination or authenticated for these experiments.

Code availability. All custom Python scripts used for the data analysis in this study are available on GitHub at https://github.com/rchari/Library-on-library.Scripts/.

19. Futreal, P.A. *et al. Nat. Rev. Cancer* **4**, 177–183 (2004).
20. Karolchik, D. *et al. Nucleic Acids Res.* **32**, D493–D496 (2004).
21. Jiang, H. & Wong, W.H. *Bioinformatics* **24**, 2395–2396 (2008).
22. Xu, Q., Schlabach, M.R., Hannon, G.J. & Elledge, S.J. *Proc. Natl. Acad. Sci. USA* **106**, 2289–2294 (2009).
23. Esvelt, K.M. *et al. Nat. Methods* **10**, 1116–1121 (2013).
24. Harris, D.R. *et al. PLoS Biol.* **2**, e304 (2004).
25. Magoč, T. & Salzberg, S.L. *Bioinformatics* **27**, 2957–2963 (2011).
26. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).
27. Li, H. *et al. Bioinformatics* **25**, 2078–2079 (2009).
28. Droettboom, M. *et al.* Matplotlib version 1.4.0. doi:10.5281/zenodo.11451 (2014).
29. Schölkopf, B., Burges, C.J.C. & Smola, A.J. *Advances in Kernel Methods: Support Vector Learning* (MIT Press, 1999).
30. Karolchik, D. *et al. Nucleic Acids Res.* **42**, D764–D770 (2014).
31. Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S. & Karolchik, D. *Bioinformatics* **26**, 2204–2207 (2010).