Published online 25 September 2017 *Nucleic Acids Research, 2017, Vol. 45, No. 20* ***12039–12053***
doi: 10.1093/nar/gkx843

CRISPR/Cas9-mediated gene knockout is insensitive to target copy number but is dependent on guide RNA potency and Cas9/sgRNA threshold expression level

Garmen Yuen[1], Fehad J. Khan[1,2], Shaojian Gao[3], Jayne M. Stommel[4], Eric Batchelor[5], Xiaolin Wu[6] and Ji Luo[1,*]

[1]Laboratory of Cancer Biology and Genetics, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA, [2]Undergraduate Scholarship Program, National Institutes of Health, Bethesda, MD, USA, [3]Thoracic and Gastrointestinal Oncology Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA, [4]Radiation Oncology Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA, [5]Laboratory of Pathology, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA and [6]Cancer Research Technology Program, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, USA

Received April 12, 2017; Revised September 11, 2017; Editorial Decision September 12, 2017; Accepted September 13, 2017

ABSTRACT

CRISPR/Cas9 is a powerful gene editing tool for gene knockout studies and functional genomic screens. Successful implementation of CRISPR often requires Cas9 to elicit efficient target knockout in a population of cells. In this study, we investigated the role of several key factors, including variation in target copy number, inherent potency of sgRNA guides, and expression level of Cas9 and sgRNA, in determining CRISPR knockout efficiency. Using isogenic, clonal cell lines with variable copy numbers of an EGFP transgene, we discovered that CRISPR knockout is relatively insensitive to target copy number, but is highly dependent on the potency of the sgRNA guide sequence. Kinetic analysis revealed that most target mutation occurs between 5 and 10 days following Cas9/sgRNA transduction, while sgRNAs with different potencies differ by their knockout time course and by their terminal-phase knockout efficiency. We showed that prolonged, low level expression of Cas9 and sgRNA often fails to elicit target mutation, particularly if the potency of the sgRNA is also low. Our findings provide new insights into the behavior of CRISPR/Cas9 in mammalian cells that could be used for future improvement of this platform.

INTRODUCTION

The discovery of the bacterial CRISPR/Cas9 endonuclease system and its adaptation as a mammalian genome editing tool has created a new platform for genome-scale loss-of-function (LOF) screens (1). To generate CRISPR knockout (KO) clones of a single gene in low-throughput assays, deletion efficiency is not a major concern because many clones can be screened to identify the few that are successfully targeted. However, to effectively implement pooled lentiviral CRISPR library screens in a high-throughput format, each guide RNA must be effective at knocking out its target gene in the entire cell population (2,3). Thus, a thorough understanding of the factors that influence CRISPR KO efficiency in mammalian cells would help the future optimization of CRISPR libraries.

Currently, the most widely used CRISPR/Cas9 system is the humanized version of the *Streptococcus pyrogenes* Cas9 endonuclease (*Sp*Cas9, hereafter referred to as Cas9). Cas9 is an RNA-dependent DNA endonuclease that utilizes a single guide RNA (sgRNA) to introduce double-stranded DNA breaks (4). The sgRNA consists of a 20 nt target-specific sequence followed by a constant, structural RNA element that is derived from the *S. pyrogenes* tracrRNA (5). In mammalian cells, sgRNAs can be programmed to target any sequence in the genome which precedes a 'NGG' protospacer-adjacent motif (PAM) (6,7). Binding of the Cas9/sgRNA complex to the target site and subsequent DNA cleavage is a multi-step process (8,9). Cas9 first identifies the PAM before eliciting strand invasion by the sgRNA 20-mer. Stable heteroduplex formation between the sgRNA 20-mer and DNA leads to R-loop expansion and conformational change in Cas9. Finally, coordinated firing of both nuclease domains of Cas9 results in DNA double strand cleavage. In mammalian cells, repair of the DNA break via the non-homologous ending joining (NHEJ) pathway often gives rise to small insertion and deletion (indel) mutations at the cleavage site (4). Thus, when targeted to the exon re-

[*]To whom correspondence should be addressed. Tel: +1 240 760 6931; Fax: +1 240 541 4464; Email: ji.luo@nih.gov

gions of a gene, Cas9 can elicit a high frequency of frameshift mutation and silence gene expression.

To enable high-throughput LOF screens on a genome scale using CRISPR/Cas9, genome-wide CRISPR libraries consisting of pooled lentiviral sgRNAs have been constructed (10). Conceptually, LOF screens using CRISPR libraries are similar to those using shRNA libraries: CRISPR library virus is transduced into cells at a low multiplicity of infection (MOI) such that each cell receives a single sgRNA and thus has one gene knocked out. The resulting population of cells harboring gene KO are selected with the appropriate functional assay to identify genes that are responsible for the desired phenotype (10). CRISPR library screens have been applied to identify genes whose KO leads to lethality in cancer cells and thus could serve as potential drug targets (11–14).

A unique challenge associated with CRISPR KO screen in this context arises from the aneuploid nature of many cancer cell lines. Cancer cell lines, especially those derived from epithelial tumors, often exhibit extensive somatic gene copy number variation (SCNV) as a result of chromosomal instability (15, 16). Thus, a gene could vary from a single copy in one cell line (e.g. heterozygous deletion) to several copies in a second cell line (e.g. aneuploidy) and to many copies in a third cell line (e.g. gene amplification). SCNV poses a unique challenge for CRISPR screen. It has been observed that Cas9 cutting at multiple target sites can lead to DNA damaged-induced cell death (12,13,17). On the other hand, it is not clear to what extent is the KO efficiency of the CRISPR/Cas9 system sensitive to target gene copy number. If the efficiency of Cas9-mediated gene editing is highly sensitive to SCNV, a sgRNA will become progressively ineffective in cell lines with increasing copy numbers of its target gene. This will both increase false-negative rate within an aneuploid cell line and complicate data analyses across cell lines with different SCNV profiles. Thus, the successful implementation of CRISPR library screen in cancer cells requires a better understanding of the behavior of the CRISPR/Cas9 system in relationship to target gene SCNV.

In this study, we systematically investigated the relationship between the KO efficiency of Cas9/sgRNA and the copy number of a target EGFP transgene in isogenic cell lines at the population level. Our result indicates that KO efficiency is largely insensitive to the copy number of the target gene but is primarily determined by the intrinsic potency of the sgRNA 20-mer. Kinetic analysis revealed that the majority of gene deletion occurs between 5 and 10 days, and successful gene knockout requires the expression level of Cas9 to exceed a threshold level in the cell that is dependent on sgRNA potency.

MATERIALS AND METHODS

Cell lines, lentiviral vectors and cell transduction

Human embryonic kidney (HEK) cell line 293T was cultured in Dulbecco's modified Eagle's Medium (Lonza #12-604Q) supplemented with 10% heat inactivated fetal bovine serum (Gibco #10438026) and 1% penicillin/streptomycin (Lonza #17–602E). Human osteosarcoma cell line U2OS was obtained from ATCC (ATCC #HTB-96) and cultured in McCoy's 5A media supplemented with 10% heat-inactivated fetal bovine serum (Gibco #10438026) and 1% penicillin/streptomycin (Lonza #17-602E). The GB1 glioma cell line was culture in DMEM (HyClone #SH30243) with 10% heat inactivated fetal bovine serum. All cell lines were cultured at 37°C in a humidified 5% CO_2 incubator.

Stable clonal cell lines expressing different copy number of the EGFP transgene were created by transducing cells with a retroviral vector expressing the EGFP cDNA (MSCVhygro-EGFP) at a range of MOI. Transduced cells were selected with hygromycin for 5 days and cells were plated at low density for single cell cloning. Single cell colonies were expanded and the copy number of EGFP in cells were evaluated by qPCR and sequencing of genomic DNA.

For lentivirus production, 293T cells grown to 90% confluency in a 10 cm tissue culture plate were transfected with packaging plasmids pMD2.G (Addgene Plasmid #12259) and psPAX2 (Addgene Plasmid #12260) together with the Cas9/sgRNA plasmid at a ratio of 1:1.5:2.5 using TransIT-293 transfection reagent (Mirus #2700) according to the manufacturer's instruction. At 20 h post-transfection, cells were refreshed with 5 ml of growth media and viral supernatant was collected at 48 h after transfection and stored at –80°C until use. Virus titer was measured by transducing known number of 293T cells with various amount of virus for 16 h, followed by drug selection for 3 days (puromycin for Cas9/sgRNA co-expression vector and sgRNA expression vector, blasticidin Cas9 expression vector). Cell viability was measured using CellTiter-Glo (Promega #G8462). Virus titer was calculated based on the percentage of viable cells and the total number of cells at the time of transduction.

For experiments comparing sgRNA potency and the effect of EGFP transgene copy number on knockout efficiency, clonal cells were transduced with lentivirus at MOI $\leq$1 followed by 3 days of puromycin selection. For experiments aimed to test Cas9 and sgRNA dosage effects as described in Figure 5 and Supplementary Figure S10, a higher MOI was used for cell transduction and the MOI for each experiment was indicated. For second-round re-infection experiments as described in Figure 6, cells were transduced with sgRNA with Cas9 at MOI of 5.

For GB1 cell transduction, GB1 cells were transduced with lentiCRISPR-EGFP-sgCCT2 or lentiCRISPR-EGFP-sgCtrl lentivirus, cultured for 13 days, then sorted into thirds comprising EGFP low, medium, and high populations.

Plasmids, sgRNA design, and molecular biology

The EGFP sgRNAs sgEGFP1, sgEGFP2 and sgEGFP4 have been previously described (3). The three additional EGFP sgRNAs (sgEGFP5, sgEGFP6 and sgEGFP7) and the sgRNA targeting the human CCT2 gene were designed using previously published algorithms (18,19). The same algorithms were used to predict the potency score of these sgRNAs. The DNA target sequences of these sgRNAs are listed in Supplementary Figure S1C. DNA Oligos of sgRNAs were cloned into LentiCRISPRv2 (Addgene Plasmid

#52961) and LentiGuide-Puro (Addgene Plasmid #52963) as previously described (20).

The lentiCRISPR-EGFP vector was created by excising the P2A-puromycinR-WPRE-3′LTR DNA fragment between the BamHI and PmeI sites of lentiCRISPRv2 and replacing it with a gBlock double stranded DNA fragment (IDT) that encodes a P2A-EGFP-WPRE-3′LTR fragment. DNA assembly was carried out using the NEBuilder HiFi DNA Assembly kit (New England Biolabs #E2621S).

To measure Cas9 protein expression level in cells, Cells were directly lysed with 1X Laemmli sample buffer (Bio-Rad 1610747) and cell lyate was denatured at 95°C for 10 min. Samples were run on a 6% SDS-PAGE gel and transferred to nitrocellulose membrane. The membrane was blotted with a Cas9 antibody (Millipore #MAC133, 1:1,000 dilution). A vinculin antibody blot (Sigma #V9131, 1:10 000 dilution) was used as loading control. Blots were imaged on a Bio-Rad ChemiDoc MP Imaging System and band densitometry was measured using the associated Image Lab Software (Bio-Rad). To measure CCT2 protein levels in GB1 cells, at five days post EGFP sorting, whole cell lysates were generate using Laemmli sample buffer and were separated on a SDS-PAGE gel and transferred to nitrocellulose membrane. The membrane was immunoblotted with a CCT2 antibody (AbD Serotec #VPA00123, 1:1000 dilution). A RAN antibody (Becton Dickinson #610340, 1:10 000 dilution) blot was used a loading control. Blots were scanned on a Li-Cor Odyssey CLx Imaging System and band densitometry was measured using the Li-Cor Image Studio Software (Li-Cor).

DNA and RNA quantification by real-time PCR (qPCR)

For DNA quantification of the EGFP transgene copy number, genomic DNA of clonal cells stably expressing the EGFP transgene were extracted using QIAamp DNA mini kit (Qiagen #51306) and treated with 4ug/ml of RNAse A (Qiagen #19101). qPCR reactions were performed in triplicates with 50 ng of genomic DNA using Power SYBR Green PCR Master Mix (Applied Biosystems #4309155) on a 7900HT Fast Real-Time PCR System platform (Applied Biosystems) according to manufacturer′s protocol. The primer sets used for qPCR are listed in Supplementary Table S1. The data were normalized to GAPDH as an endogenous control using the ΔC_t method. Because 293T clones A and B, and U2OS clone A were infected with the EGFP retrovirus at MOI <0.1, we assumed these clones to harbor a single copy of EGFP. For 293T clones A and B, we also confirmed by integration site-sequencing that they indeed contained a single integration site. For the other clones, EGFP transgene copy number was estimated by the fold-change in EGFP DNA levels relative to the single copy clones.

For RNA quantification of Cas9 mRNA and sgRNA expression levels, cells were lysed with RiboZol RNA extraction reagent (AMRESCO #N580). Small and total RNAs were extracted using miRNeasy Mini Kit (Qiagen #217004). cDNA synthesis was performed with 1 μg of total RNA using the high-capacity cDNA reverse transcription kit (Applied Biosystems #4368814) according to manufacturer′s protocol. qPCR reactions were performed in triplicates using Power SYBR Green PCR Master Mix (Applied Biosystems #4309155) on a 7900HT Fast Real-Time PCR System platform (Applied Biosystems). The primer sets used for qPCR are listed in Supplementary Table S1. The data were normalized to β-actin as an endogenous control using the ΔC_t method.

Flow cytometry analysis

Cells were trypsinized and collected at the indicated time points for live-cell fluorescence activate flow cytometry analysis. FACS scan analysis was carried out using a BD FACSCalibur cell analyzer (BD Biosciences). Viable single cells were first gated using the forward and side scatter and the EGFP fluorescence signals was measured using a 488 nm laser. FACS sorting was carried out using a BD FACSAria cell sorter (BD Biosciences).

MiSeq analysis for EGFP transgene integration sites

Retroviral integration sites of the retroviral EGFP transgene were mapped using linker-mediated PCR as described previously (21). Briefly, genomic DNA was sheared to ∼300–500 bp fragments that were subsequently end-repaired and dA-tailed. A T-linker was ligated on to the DNA fragment and the retroviral-genomic junction sequence was amplified with one primer specific to the LTR and one primer specific to the linker. A second round of nested PCR was used to ensure the specificity of the target amplicon and to add Illumina sequencing adaptors. Sequencing was carried out using Illumina MiSeq 1 × 150 bp single-end reads and the junction sequence was trimmed and then aligned to host genome (hg19) using BLAT.

Cas9/sgRNA-mediated indel analysis at genomic target sites

To assess the allelic frequency of indel mutation introduced by Cas9/sgRNA at the EGFP target sites, we used both the TIDE analysis (22) of Sanger sequencing chromatograms and the Illumina MiSeq platform. Genomic DNA of clonal cells were extracted at indicated time points post Cas9/sgRNA transduction. For TIDE analysis, a fragment of the EGFP transgene encompassing the sgRNA target sites were PCR amplified using Phusion Hot Start II DNA polymerase (ThermoFisher Scientific #F549L). In the PCR reaction, 50 ng of genomic DNA was used as template and the PCR condition were as follows: an initial hot start step of 98°C for 30 s; 30 cycles of 98°C for 10 s, 60°C for 30 s and 72°C for 30 s; and a final extension step of 72°C for 10 min. Purified PCR products were confirmed by gel electrophoresis and subject to Sanger sequencing. Primers pairs spanning the sgRNA target sites were used for PCR amplification and Sanger sequencing are listed in Supplementary Table S1. EGFP_Seq_F1 was used for samples treated with sgEGFP4. EGFP_Seq_R1 was used for samples treated with sgEGFP1 and sgEGFP2. Sequencing chromatograms were analyzed using the TIDE web tool (https://tide.nki.nl/) to assess the percentage of indel mutant allele vs. WT allele.

For MiSeq analysis, a fragment of the EGFP transgene encompassing the sgRNA target sites were PCR amplified in a similar fashion. To accommodate the

MiSeq read length, the target sites for sgEGFP1 and sgEGFP2 and the target site for sgEGFP4 were amplified as two separate amplicons. The primer pair M13F-MSCVpuro_F1 and M13R-EGFP_indel_check_R1 were used for sgEGFP1 and sgEGFP2 target sites, and the primer pair M13F-EGFP_indel_check_F2 and M13R-EGFP_R1_seq were used for the sgEGFP4 target site (Supplementary Table S1). Purified PCR products were gel purified and prepared for Illumina MiSeq sequencing. Reads were aligned to the WT allele sequence to identify insertion and deletion mutations.

Potency score calculation for EGFP sgRNAs

The potency scores for each EGFP sgRNA was derived using the bioinformatics tool developed by Doench and colleagues (18,19). Briefly, the algorithm considers several factors including single- and di-nucleotide identity at each position, position-independent single- and di-nucleotide count, position of the sgRNA within the protein coding sequence and melting temperature.

Off-target site analysis of sgRNAs

To identify potential off-target sites in the human genome, each sgRNA was BLASTed against the human genome (hg38). Perfect match sequences of specified length to the sgRNA 3′ end that also conform to the NGG PAM were counted.

Modeling and statistical analysis

Modeling analysis was carried out using Prism 7. To model the loss of EGFP full cells, a one-phase exponential decay model with an initial delay and a final plateau was used. This model was chosen to minimize the number of parameters while preserving good fit to the experimental data (as indicated by R^2 values). An initial delay parameter was included in the model to account for the time required for Cas9 and sgRNA to integrate and express in the cell. All model parameters were auto-derived for the best fit to experimental data. For 293T clones E and F transduced with sgEGFP1, auto-fitting did not give a convergent fit. We thus constrained the delay to 3 days, which approximated the average delay time of the other datasets. This resulted in a convergent fit for these two samples.

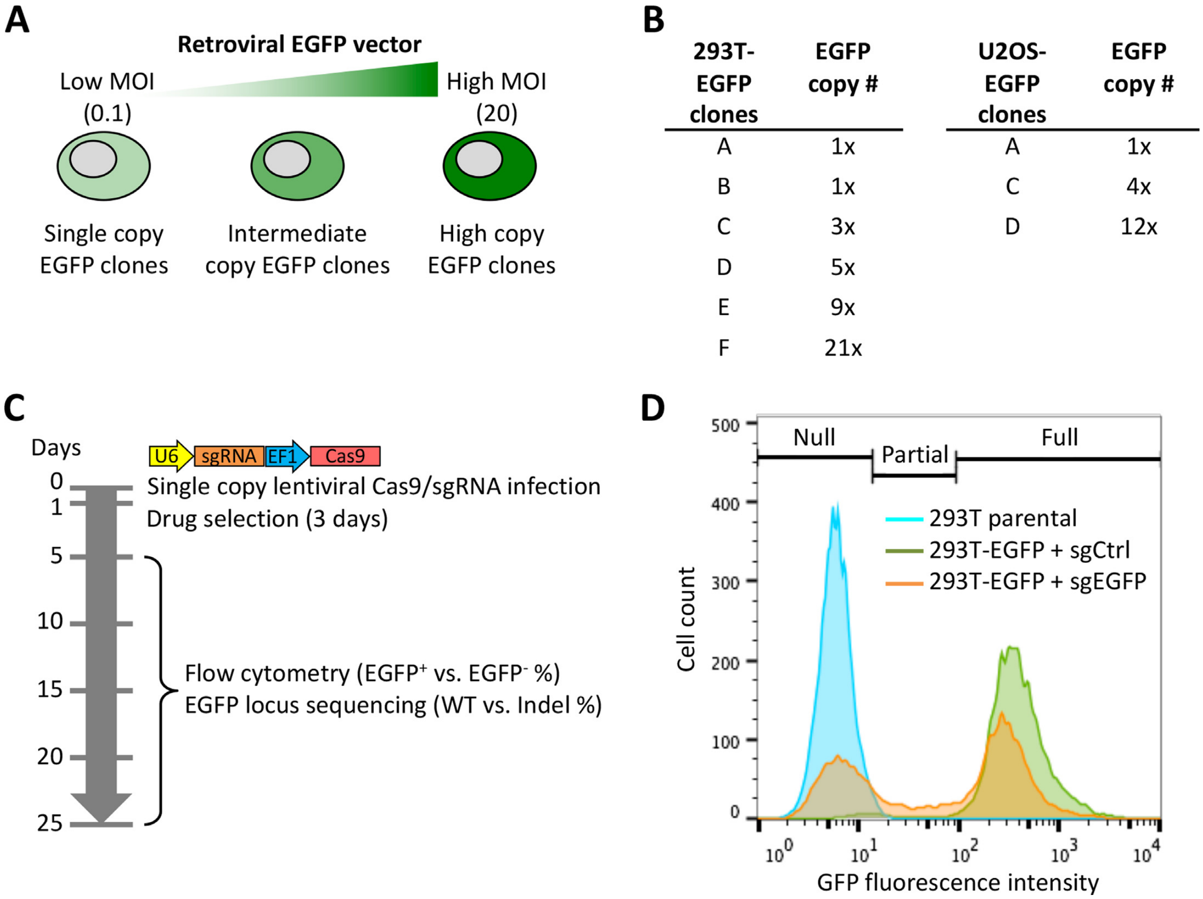
One-way and two-way ANOVA analysis was carried out using Prism 7 to compare the performance of different sgRNAs over time or the behavior of different cell line clones over time. All *P*-values were tabulated in Supplementary Table S2, a *P*-value ≤0.05 was used to define statistically significant differences.

RESULTS

Creation of isogenic cell lines with different copy number of EGFP transgene

Because most cancer cell lines from solid tumors exhibit aneuploid genomes, we sought to determine how SCNV in a target gene impacts CRISPR/Cas9-mediated gene KO efficiency. To this end, we sought to establish clonal lines of human 293T and U2OS cells carrying variable copy number of an EGFP transgene. These cell lines would enable us to study the deletion efficiency of Cas9 in an isogenic setting. We transduced 293T and U2OS cells with a retroviral vector expressing EGFP at different multiplicities of infection (MOI) ranging from 0.1 to 20 (Figure 1A). We next isolated multiple clones from each condition and measured their copy number of EGFP proviral integration by qPCR. We curated a panel of clones that harbored single, intermediate, and high copy number of the EGFP transgene (Figure 1B and Supplementary Figure S1A and B). For the 293T clones, we also carried out integration-site sequencing as an independent means to verify their EGFP transgene copy number. The EGFP transgene copy number were largely consistent between these two measurement methods, albeit qPCR was less sensitive in the high copy clones (Supplementary Figure S1A). These clones therefore are good models for the typical range of gene SCNVs seen in cancer cell lines.

To evaluate the efficacy of Cas9 in our panel of clonal cell lines, we employed three non-overlapping sgRNAs targeting EGFP, hereafter abbreviated as sgEGFP1, sgEGFP2 and sgEGFP4 (Supplementary Figure S1C and D), that were previously used to study EGFP transgene KO (3). Because pooled CRISPR library screens typically require a low MOI of ≤1 to allow single-copy sgRNA integration in each cell, we reasoned that it is most relevant to test the potency of these sgRNAs in the context of single-copy integration. For these experiments, we used an all-in-one lentiviral vector, LentiCRISPRv2, that was designed to co-expressed Cas9 and sgRNA from EF1 and U6 promoters, respectively (20). We transduced each cell line with Cas9/sgRNA at a MOI ≤1 and imposed a brief drug selection to remove uninfected cells. We then measured EGFP expression in stably transduced cells every 5 days by flow cytometry for a period of 25 days. In parallel, we analyzed the EGFP target site mutation frequency by using the TIDE sequence analysis tool (22) which models Sanger sequencing chromatograms to calculate the ratio of indel allele to WT allele (Figure 1C). These two methods were complementary: flow cytometry enabled us to functionally assess the EGFP expression status of individual cells, while TIDE allowed us to estimate the frequency of indel versus WT alleles in the population. In this scheme, we could quantify both the kinetics and the extent of EGFP transgene deletion in cells. For cells with a single copy of EGFP transgene, we expected a binary outcome of either KO or no KO. For cells with multiple copies of EGFP it is possible that only a fraction of the EGFP transgene were knocked out while others remained intact in a cell. Thus, in the flow cytometry analysis we gated cells into three populations (Figure 1D). The ‘EGFP null’ gate represents cells with no EGFP signal and thus all EGFP transgenes would be knocked out. The ‘EGFP full’ gate represents cells with full EGFP signal and thus most of the EGFP transgenes should remain intact. Lastly, the ‘EGFP partial’ gate represents cells with reduced EGFP signal that resulted from one of two scenarios: either a significant fraction of the EGFP transgenes is knocked out in the cell which results in a drop in EGFP signal, or all EGFP transgenes are knocked out but the EGFP protein is not yet fully degraded due to its relatively long half-life. For clones with multiple EGFP



293T-EGFP clones	EGFP copy #
A	1x
B	1x
C	3x
D	5x
E	9x
F	21x

U2OS-EGFP clones	EGFP copy #
A	1x
C	4x
D	12x

Figure 1. Cell line model for evaluating CRISPR/Cas9 knockout efficiency. (**A**) Establishment of stable clonal cell lines with different copy number of the EGFP transgene. (**B**) 293T and U2OS clones with their respective EGFP copy number. (**C**) Schematics of the experimental workflow. A lentiviral vector co-expressing Cas9 and sgRNA was used to target the EGFP transgene. Samples were analyzed at the indicated time points. (**D**) Flow cytometry histogram illustrating how cells were gated for their EGFP signal. Profiles of 293T parental cells (EGFP$^-$), a 293T clone (EGFP$^+$) transduced with Cas9 and a control sgRNA (sgCtrl), and 293T cells transduced with Cas9 and a sgRNA targeting EGFP (sgEGFP) were overlaid. The 'full' gate included cells with EGFP signal intensity similar to the untreated EGFP$^+$ cells of the same clone. The 'null' gate included cells with complete loss of EFGP signal that appeared similar to the EGFP$^-$ 293T parental cells. The 'partial' gate included cells with detectable EGFP signal that was between the EGFP$^+$ and EGFP$^-$ controls.

transgenes these two scenarios are difficult to distinguish without extensive single cell sequencing. We thus primarily focused on the EGFP null and EGFP full populations when analyzing gene KO efficiency. To test the sensitivity of flow cytometry at detecting fractional loss of EGFP alleles in clones with high EGFP copy number, we carried out a mixing experiment using clones B, D E and F that carried 1, 5, 9 and 21 copies of the EGFP transgene, respectively. The peak EGFP signal for clones B and D are clearly distinguishable from E and F, although the latter two showed similar EGFP signal peaks, possibly due to saturating level of EGFP protein in these high copy number clones (Supplementary Figure S1E). When all four clones were pre-mixed with parental cells at equal ratio and the EGFP full gate was set using the signal from clone F, a large fraction of clones B and D were captured in the EGFP partial gate, although some fell within the EGFP full gate (Supplementary Figure S1F). Flow cytometry was therefore sufficiently sensitive to detect changes in EGFP signal when a substantial fraction of alleles are mutated (e.g. reduction of WT EGFP copy number from 9 to 5 or 1, and from 21 to 5 or 1), but it was insufficiently sensitive to detect the fractional knockout of EGFP allele in the extreme case of clone F (e.g. reduction of WT EGFP copy number from 21 to 9). For this reason, we complemented flow cytometry with two sequencing based methods to independently estimate mutational frequency in the EGFP transgene.

CRISPR/Cas9 KO efficiency is primarily determined by sgRNA potency and not by target gene copy number

In 293T cells, the three EGFP sgRNAs exhibited different degrees of KO efficiency and were thus highly informative on the behavior of the system (Supplementary Figures S2 and S3). By flow cytometry, sgEGFP1 was the most efficient at knocking out the EGFP transgene, whereas sgEGFP2 was somewhat less efficient and sgEGFP4 was the least efficient (Supplementary Figure S3A–C). As expected, within each experiment the reduction of EGFP full cells in the population was accompanied by a corresponding, time-delayed increase in EGFP null cells (Supplementary Figure S3D–F). Remarkably, the KO efficiency of each sgRNA was highly consistent across 293T clones with different EGFP copy number. The most potent sgRNA, sgEGFP1, knocked out

EGFP expression with similar kinetics and efficiency in clones with both single and high copy number of EGFP (Figure 2A). In contrast, sgEGFP2 and sgEGFP4 were unable to knock out EGFP in a significant fraction of cells (Figure 2B and C). Despite the reduced potency of these latter two sgRNAs, their kinetics and efficiency were similar across all clones. In 293T clones E and F, which harbored 9 and 21 copies of the EGFP transgene, respectively, sgEGFP1 was still able to efficiently knock out all copies of EGFP, as reflected by the high fraction of EGFP null cells it generated over time (Figure 2D). On the other hand, sgEGFP2 and sgEGFP4 generated fewer EGFP null cells in these clones compared to the single and low copy clones (Figure 2E and F). This suggests that in the presence of multiple copies of the target gene, a less potent sgRNA is less effective at knocking out all of the copies, but instead generates more partially deleted cells. Indeed, whereas the fraction of EGFP partial cells rapidly diminished in all sgEGFP1 transduced clones, in sgEGFP2 and sgEGFP4 transduced clones the EGFP partial population persisted throughout the 25-day period and was more prominent in the high copy clones than the low copy clones (Figure 2G and H).

To verify that the loss of EGFP by flow cytometry reflected the presence of inactivating indel mutations at the sgRNA target sites, we PCR amplified the EGFP target site from three representative 293T clones with 1, 3 and 21 copies of the EGFP transgene and sequenced them by Sanger sequencing. We used the TIDE bioinformatics tool to calculate the relative frequency of WT and indel alleles (22). For each sgRNA, the loss of WT EGFP allele and the accumulation of mutant EGFP allele exhibited similar kinetics and efficiency compared to those observed by flow cytometry (Figure 3). Whereas sgEGFP1 mutated the vast majority of EGFP alleles in all clones (Figure 3A–C), sgEGFP2 and sgEGFP4 only mutated a fraction of the EGFP alleles in each clone, and the fraction of mutant alleles decreased as the copy number of the target gene increased (Figure 3D–I). In 293T cells, the predominant mutation introduced by Cas9 at the cleavage site was a +1 insertion (Supplementary Figure S4A), which would result in a frame-shift mutation in EGFP. To independently verify the TIDE result, we directly counted the read frequency of mutant and WT EGFP alleles in 293T clone A using MiSeq, and obtained similar indel frequency with each sgRNA (Figure 3J–L). Further corroborating the TIDE data, MiSeq also identified +1 insertion as the major mutation in these cells, and each sgRNA target site showed a distinct nucleotide preference (Supplementary Figure S4B and C).

It is possible that difference in KO efficiency among the three sgRNAs might be due to difference in the expression level of different sgRNAs and/or Cas9. This was unlikely because the same vector backbone was used in these experiments. To formally rule out this possibility, we measured Cas9 protein level by western blot in two representative 293T clones A and F. Cas9 expression level was similar for all three sgRNAs (Supplementary Figure S5A and B). Interestingly, we noticed that in both clones the expression level of Cas9 peaked at day 5–10 and declined thereafter. This was confirmed with a shorter time course experiment, which showed that Cas9 expression experienced an initial burst before declining, at least in this system (Supplementary Figure S5C). We also measured the expression level of each sgRNA in cells using RT-qPCR. Again, we did not observe substantial difference in sgRNA expression levels at day 5 and day 10 (Supplementary Figure S5D). These and subsequent experiments indicate that the difference in activity among the three EGFP sgRNAs results from their intrinsic potency rather than their expression level.

To confirm that our findings are not cell-line specific phenomena, we repeated the above KO experiments using clonal U2OS cell lines harboring 1, 4 or 12 copies of the EGFP transgene (Supplementary Figure S1B). Similar to 293T cells, the KO efficiencies of the three EGFP sgRNAs were in the order of sgEGFP1>sgEGFP2>sgEGFP4. Each sgRNA exhibited consistent behavior across all three U2OS clones as measured by flow cytometry analysis (Supplementary Figure S6) and by TIDE sequencing analysis (Supplementary Figure S7).

Taken together, both flow cytometry analysis of single cells and target site mutation analysis at the population level consistently indicate that sgRNA potency is the primary factor for efficient Cas9-mediated gene KO. The copy number of the target gene is largely inconsequential for a potent sgRNA. Less potent sgRNAs, on the other hand, are more sensitive to target copy number, and a higher copy number of target could modestly increase the propensity for partial KO cells while reducing the fraction of complete KO cells. Our findings thus support the notion that sgRNA potency, rather than target copy number, is the major determinant of gene KO efficiency by CRISPR/Cas9.

A number of bioinformatics tools have been developed to predict the potency scores of sgRNAs based on large-scale experimental data (2,3,18,19,23). We noticed that, using the sgRNA potency prediction tool developed by Doench et al. (19), sgEGFP1 was predicted to have a substantially higher potency score than sgEGFP2 and sgEGFP4 (Supplementary Figure S1C). To further test the utility of preselecting sgRNAs based on potency score prediction, we designed three additional sgRNAs, sgEGFP5, sgEGFP6 and sgEGFP7, that were predicted to have good potency scores (Supplementary Figure S1C). Indeed, these three sgRNAs were effective at knocking out EGFP expression regardless of EGFP copy number (Supplementary Figure S8). Thus, rule-based selection of sgRNA sequence could improve the success rate of gene KO under the scenario of high copy number of the target.

One potential explanation for the difference in sgRNA potency is that a less potent sgRNA might have more off-target binding sites in the genome that could serve to 'soak up' the sgRNA and reduce its effective concentration at its on-target site. Previous studies showed that a PAM-proximal 'seed' sequence match of 10–12 nucleotides is critical for sgRNA binding and target cleavage (24,25). We therefore analyzed the number of potential off-target binding sites in the genome for each EGFP sgRNA. We did not observe a clear correlation between sgRNA potency and the number of off-target site with various 3' match lengths (Supplementary Figure S8G and H). Furthermore, when we examined ~800 sgRNAs with experimentally determined potency from a previous study (18), we did not observe a

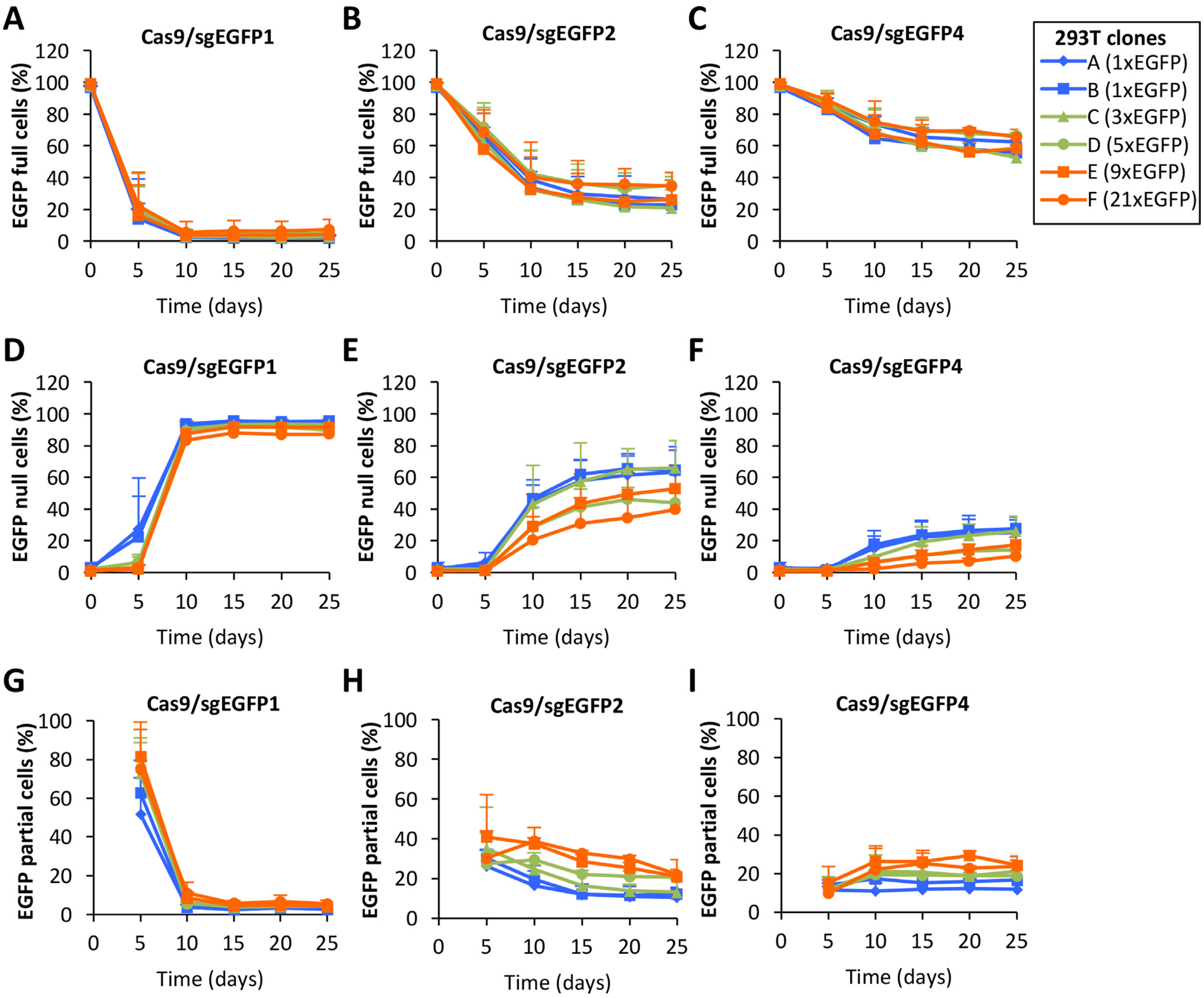


Figure 2. EGFP knockout efficiency by CRISPR/Cas9 is primarily determined by sgRNA potency but not by target gene copy number. (**A–I**) Kinetics of EGFP signal change in cells transduced with Cas9/sgRNA. The plots illustrate the loss of EGFP full cells (**A–C**), the accumulation of EGFP mull cells (**D–F**) and the change in EGFP partial cells (**G–I**) in six 293T clones over time. Data legends are the same for all panels. In all figures, error bars represent S.D. of at least three independent experiments unless otherwise specified. *P*-values from two-way ANOVA are tabulated in Supplementary Table S2A.

strong correlation between sgRNA potency and the number of off-target site with various 3' match lengths (Supplementary Figure S8I). Thus, although difficult to prove directly, our analysis suggests that the number of off-target sites is unlikely to be the explanation for the different potency of the EGFP sgRNAs observed in this study.

Cas9 mediated gene KO efficiency is sensitive to both Cas9 and sgRNA expression level

One unexpected observation from our time course study in both 293T and U2OS cells was that, regardless of a sgRNA′s potency, the fraction of cells with EGFP deletion appeared to plateau after 15 days. In all experiments, both flow cytometry at the single cell level (Figure 2 and Supplementary Figure S6) and sequencing at the population level (Figure 3 and Supplementary Figure S7) showed that the majority of EGFP deletion occurred within 10 days post Cas9/sgRNA transduction. From day 15 - 25, there were only a small amount of additional EGFP deletion events. This finding was surprising because we employed a constitutive lentiviral vector to express Cas9 and sgRNA from mammalian promoters (EF1 and U6 promoters, respectively). It was therefore reasonable to expect that even an inefficient sgRNA would eventually mutate all of its target given sufficient time. Instead, our experimental observation suggests that the window for Cas9-mediated gene KO is transient, and cells become refractory after 10–15 days of Cas9 expression.

To better understand the mechanistic basis of how the rate of target KO declined over time, we modeled the kinetics of EGFP deletion in 293T cells by fitting the data in Figure 2 to an exponential decay model with an initial delay and a final plateau (Figure 4A). In this model,

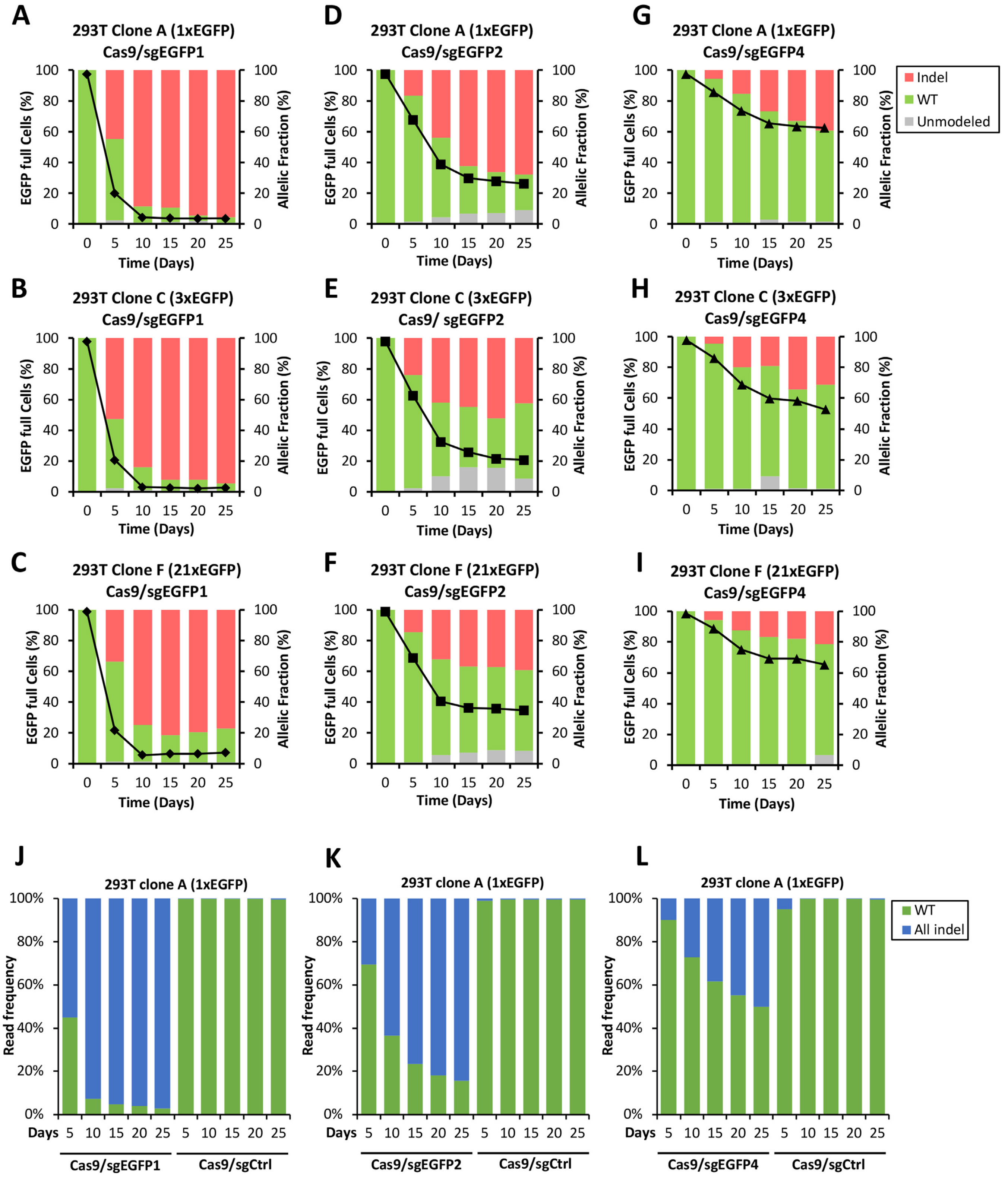


Figure 3. Mutant EGFP allelic frequency is primarily determined by sgRNA potency but not gene copy number. (**A–I**) The frequency of mutant (indel) and WT EGFP alleles in 293T clones at various time points following Cas9/sgRNA transduction. Sanger sequencing chromatogram of each sgRNA target site was modelled using TIDE to identify the fraction of mutant and WT alleles (right axis). For comparison, the fraction of EGFP full cells at the same time point from Figure 2 was plotted in the same graph (left axis, black line). (**J–L**) The frequency of mutant (indel) and WT EGFP alleles in 293T clone A at various time points following Cas9/sgRNA transduction. Allelic read frequency at each sgRNA target site was counted by MiSeq. Cells transduced with negative control sgRNA (sgCtrl) were included as control.

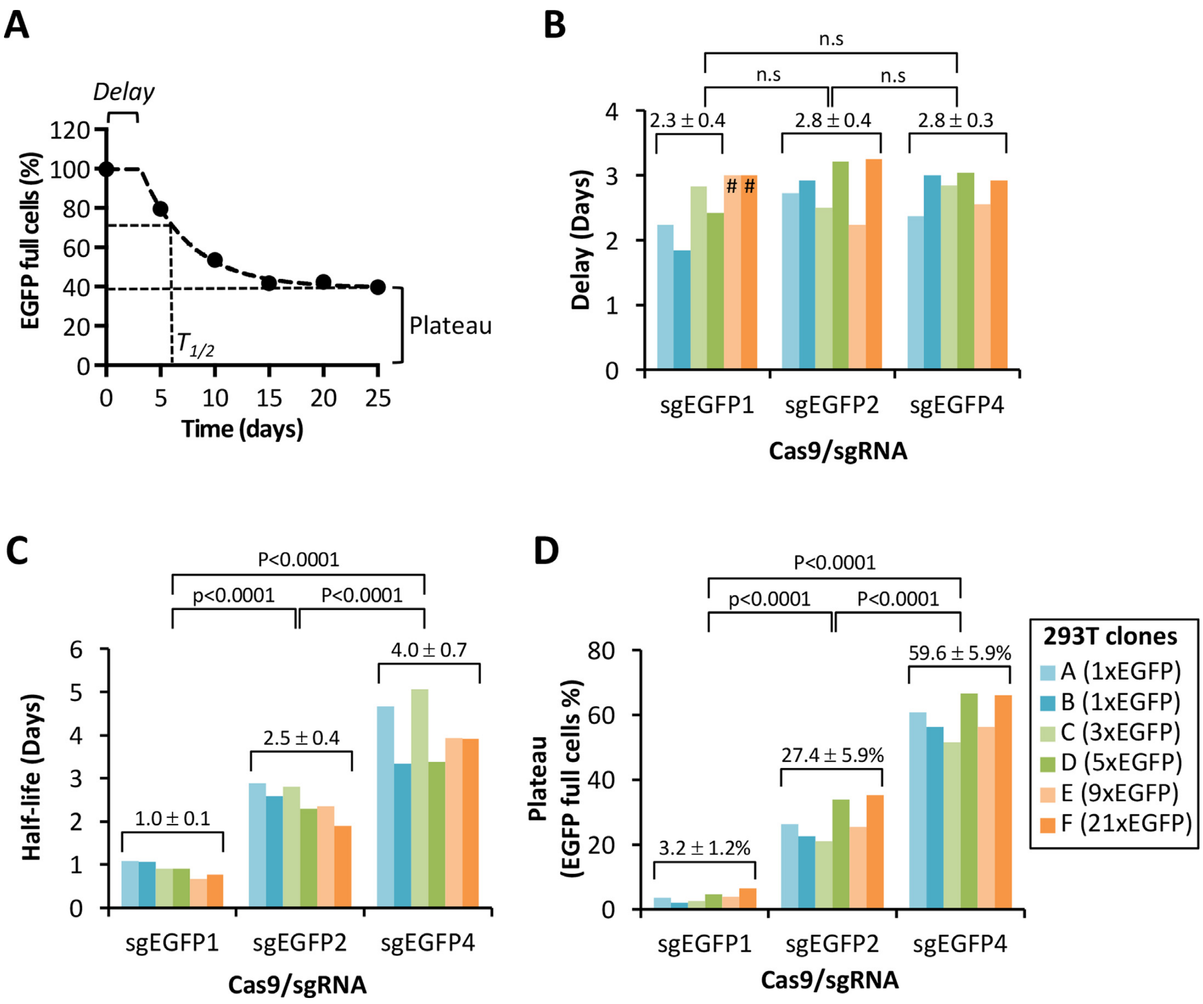


Figure 4. Modeling EGFP gene knockout kinetics. (**A**) Schematics of the one-phase delay model incorporating an initial delay and a terminal-phase plateau that was used to fit the flow cytometry data of EGFP full cells′ decline over time. Experimental data from Figure 2A–C were used for model fitting. (**B**) Modelled delay time for each sgRNA in all six 293T clones with variable EGFP copy numbers (#, delay was constrained to 3 days). Data legends are the same for panels B-D. *P*-values were from one-way ANOVA. (**C**) Modeled half-life of EGFP full cells in the population for each sgRNA in six 293T clones. *P*-values were from one-way ANOVA. (**D**) Modeled terminal-phase plateau for the fraction of persistent EGFP full cells in six 293T clones. *P*-values were from one-way ANOVA.

the initial 'delay' reflected the time required for viral integration, Cas9 and sgRNA expression, and the assembly of functional CRISPR nuclease complex in cells; the 'half-life' measured the rate at which EGFP full cells were lost, thus approximating the rate at which the EGFP transgene was deleted in the population; and the 'plateau' indicated the fraction of cells with no further EGFP deletion after prolonged exposure to Cas9/sgRNA. This model fitted the experimental data well (Supplementary Figure S9). As expected, all three sgRNAs showed similar delay of 2–3 days across all clones (Figure 4B). The half-life of EGFP full cells, however, varied significantly among the three sgRNAs and ranged from 1 day for sgEGFP1 to 4 days for sgEGFP4 (Figure 4C). The plateaus were also substantially different for the three sgRNAs, ranging from 3.2% EGFP full cells for sgEGFP1 to 59.6% for sgEGFP4 (Figure 4D). Importantly, for a given sgRNA, these three parameters were similar for all 293T clones regardless of EGFP copy number. Our modeling thus revealed two important properties of the system. First, the difference in plateau indicates that a less potent sgRNA results in fewer cells experiencing successfully gene KO in the terminal phase. Second, the difference in half-life indicates that, among the cells that do eventually experience gene KO, the rate of gene deletion is slower with a less potent sgRNA. These findings further support the notion that the kinetics of target deletion is largely driven by sgRNA potency rather than target gene copy number.

A simple explanation for the dramatic decline in CRISPR efficiency at later time points is that the lentiviral expression vector we employed became silenced over time. Our examination of Cas9 protein, however, indicated that Cas9 was expressed continuously throughout the 25-day period, although the expression was lower at later time points (Supplementary Figure S5A–C). We thus tested whether increasing the expression of Cas9 and sgRNA by a higher MOI transduction could both improve the efficiency of gene KO for the less potent sgEGFP2 and sgEGFP4. As expected, increasing the MOI of viral transduction from 1 to 10 led to a

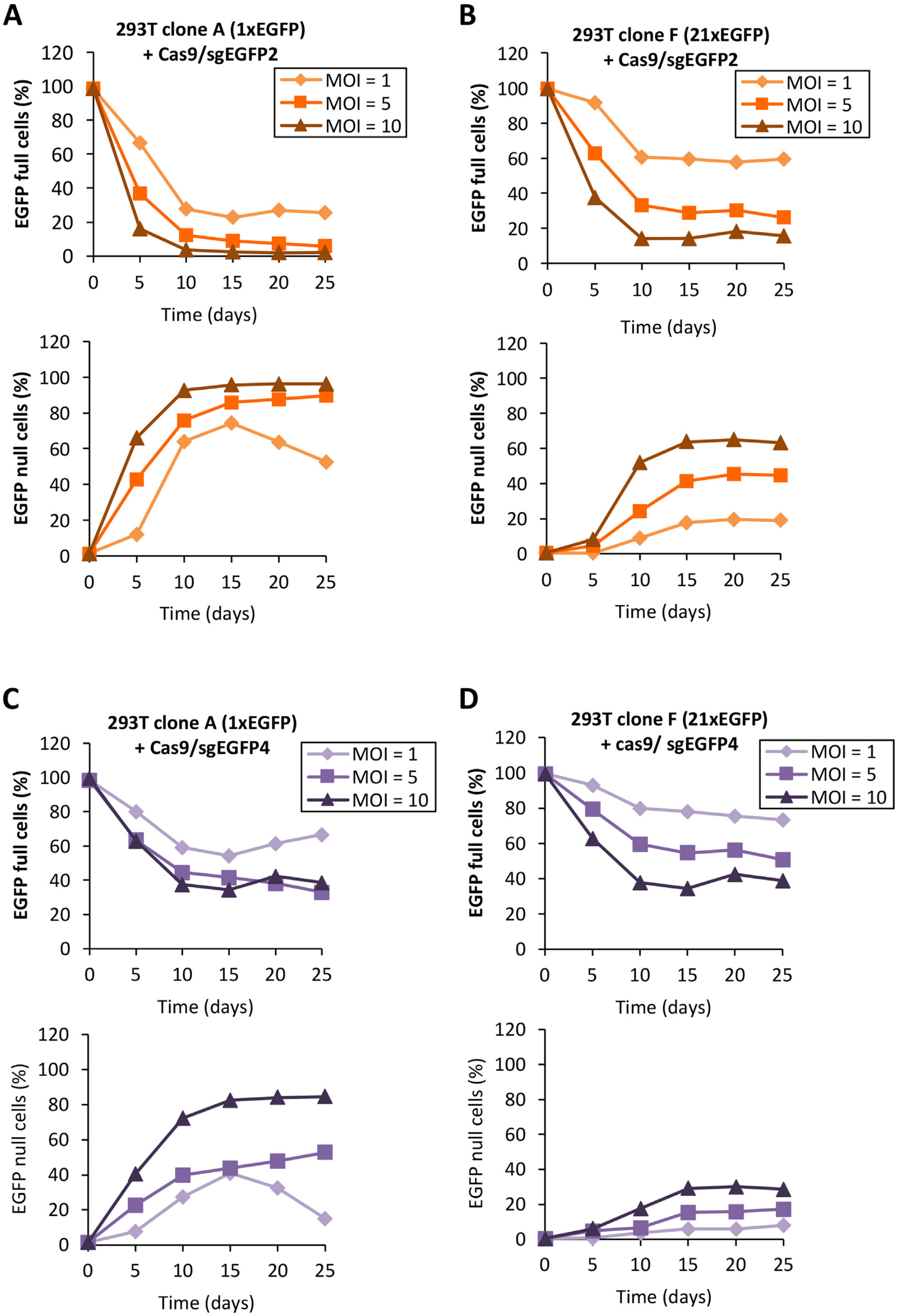


Figure 5. CRISPR/Cas9 KO efficiency increases with higher expression of Cas9 and sgRNA. (A–D) KO efficiency improved with increased MOI of Cas9/sgEGFP2 transduction (**A** and **B**) and Cas9/sgEGFP4 transduction (**C** and **D**) in 293T clones with low and high EGFP copy number. EGFP full and null cell fractions were determined by flow cytometry.

4-fold higher expression of Cas9 in the cell (Supplementary Figure S10A and B). At higher MOI, the KO efficiency was quantitatively improved for both sgEGFP2 and sgEGFP4 in cells with either single or high EGFP copy number (Figure 5). However, even at MOI = 10, both sgEGFP2 and sgEGFP4 failed to achieve the level of KO efficiency as sgEGFP1 achieved at MOI ≤1, especially in cells with high EGFP copy number (Figure 5B and D). Thus, higher expression level was unable to fully compensate for the low potency of sgEGFP2 and sgEGFP4. Importantly, both high and low MOI transductions showed similar rate of EGFP deletion, with the fraction of KO cells eventually plateauing at around day 15 (Figure 5).

A plausible hypothesis to explain the loss of CRISPR efficiency overtime is that there may exist an sgRNA-specific threshold for Cas9 and sgRNA expression that is required for successful gene KO. Cells that express below-threshold level of Cas9 and sgRNA will rarely or never experience gene KO. For a potent sgRNA this expression threshold is lower, whereas for a less potent sgRNA this threshold is higher. Because Cas9/sgRNA expression among a population of cells is likely to be heterogeneous due to lentiviral integration site effect, a less potent sgRNA would have fewer cells achieving the expression threshold required for its activity. Thus, the fraction of cells without gene KO in the terminal phase (day 15–25) reflect those with sub-threshold Cas9/sgRNA expression. This hypothesis would be consistent with the MOI dosage experiment (Figure 5). To test this hypothesis, we took terminal phase 293T clone A cells that stably expressed Cas9/sgEGFP2 or Cas9/sgEGFP4 for 15 days—at a time point when we observed a stable mix of EGFP full and EGFP null cells—and sorted for these two sub-populations by fluorescence activated cell sorting (FACS) (Figure 6A). We observed that Cas9 protein level in the EGFP null cells are higher than the population average whereas Cas9 protein level in the EGFP full cells are substantially lower than the population average (Figure 6B). In parallel, we measured Cas9 mRNA and sgRNA expression in these cells by RT-qPCR and found their expression to be substantially lower in EGFP full cells (Supplementary Figure S10C and D). These findings support the notion that for a less potent sgRNA to be active, Cas9 and sgRNA need to be expressed at a higher level. To further test this idea, we took the FACS-sorted EGFP full cells (i.e. those that failed to delete EGFP at the terminal phase) and transduced them again with lentiviral vectors that expressed either sgRNA alone, Cas9 alone or Cas9 plus sgRNA at high MOI (Schematics in Figure 6A). Increasing sgRNA expression alone in these cells, even when provided with the potent guide sgEGFP1, failed to elicit further deletion (Figure 6C and D). This indicated that the low level of Cas9 in these cells was rate-limiting. Similarly, increasing Cas9 expression alone also failed to elicit further deletion (Figure 6E and F), indicating the low level of sgRNA expression in these cells was also limiting. Only when the expression of both Cas9 and sgRNA was elevated did we elicit further deletion in these cells (Figure 6E and F). Of note, the potency order of the three sgRNAs was preserved even at high MOI transduction.

To further establish the relationship between Cas9/sgR NA expression and target KO efficiency, we modified the original LentiCRISPRv2 vector and created an all-in-one vector that co-expressed Cas9, sgRNA and EGFP where the expression of Cas9 and EGFP was linked through a P2A peptide (Supplementary Figure S11A). We tested this vector with a sgRNA against the CCT2 gene in the glioblastoma cell line GB1. After transduction, cells were FACS sorted based on their EGFP expression level, which mirrored the expression of Cas9 and sgRNA in the cell. Indeed, western blot revealed that higher EGFP expression was correlated with more efficient CCT2 deletion (Supplementary Figure S11B). Taken together, our data indicate that a lack of CRISPR KO in the presence of continuous Cas9 and sgRNA expression is due to both below-threshold level of Cas9 and sgRNA expression, as well as low sgRNA potency.

DISCUSSION

In this study, we investigated the relationship between Cas9, sgRNA and target gene copy number in determining CRISPR-mediated gene KO efficiency in a population of cells. A better understanding of these parameters is critical for the improvement of CRISPR KO library screens where efficient gene targeting at the population level is key to library saturation. By using an isogenic model system, we were able to vary, either alone or in combination, key parameters that influence Cas9/sgRNA KO efficiency and quantify the outcome. Through this analysis, several features of the CRISPR/Cas9 system became salient. Our findings are summarized in Figure 7, and we discuss below the implication of these findings.

First, we found that copy number variation of a target gene plays only a minor role in determining the KO efficiency by CRISPR. Instead, KO efficiency is primarily determined by the intrinsic potency of the sgRNA 20-mer guide sequence. We discovered that for three sgRNAs that targeted an EGFP transgene with different degrees of potency, varying the target number between 1–21 copies in the cell only had a minor effect on mutant allele frequency. This is not entirely surprising as the number of Cas9/sgRNA complexes in the nucleus is likely to exceed the copy number of its target. Thus, the activity of the CRISPR system is mainly driven by the potency of the sgRNA. We employed an EGFP transgene as a model system in this study because we wished to use an isogenic system that is highly tractable and quantifiable. This allows us to hold all other variables constant and focus on the activity of CRISPR/Cas9 in a well-controlled setting. The EGFP transgene was integrated into the genome at various locations as shown by our integration site sequencing data. It should therefore behave like endogenous, euchromatic DNA. The two distinctions are that EGFP expression is driven from a viral long terminal repeat promoter and its mRNA is not spliced. Existing literature does not suggest that promoter usage and mRNA splicing has a strong influence on Cas9 cutting of target DNA. Thus, we believe our model system is appropriate and our findings are generalizable to other endogenous genes. Because retroviral insertion tends to target open regions of the genome (26), our model may not reflect the behavior of CRISPR at heterochromatin target sites. Indeed, it has been shown that Cas9 binding and cutting is less effi-

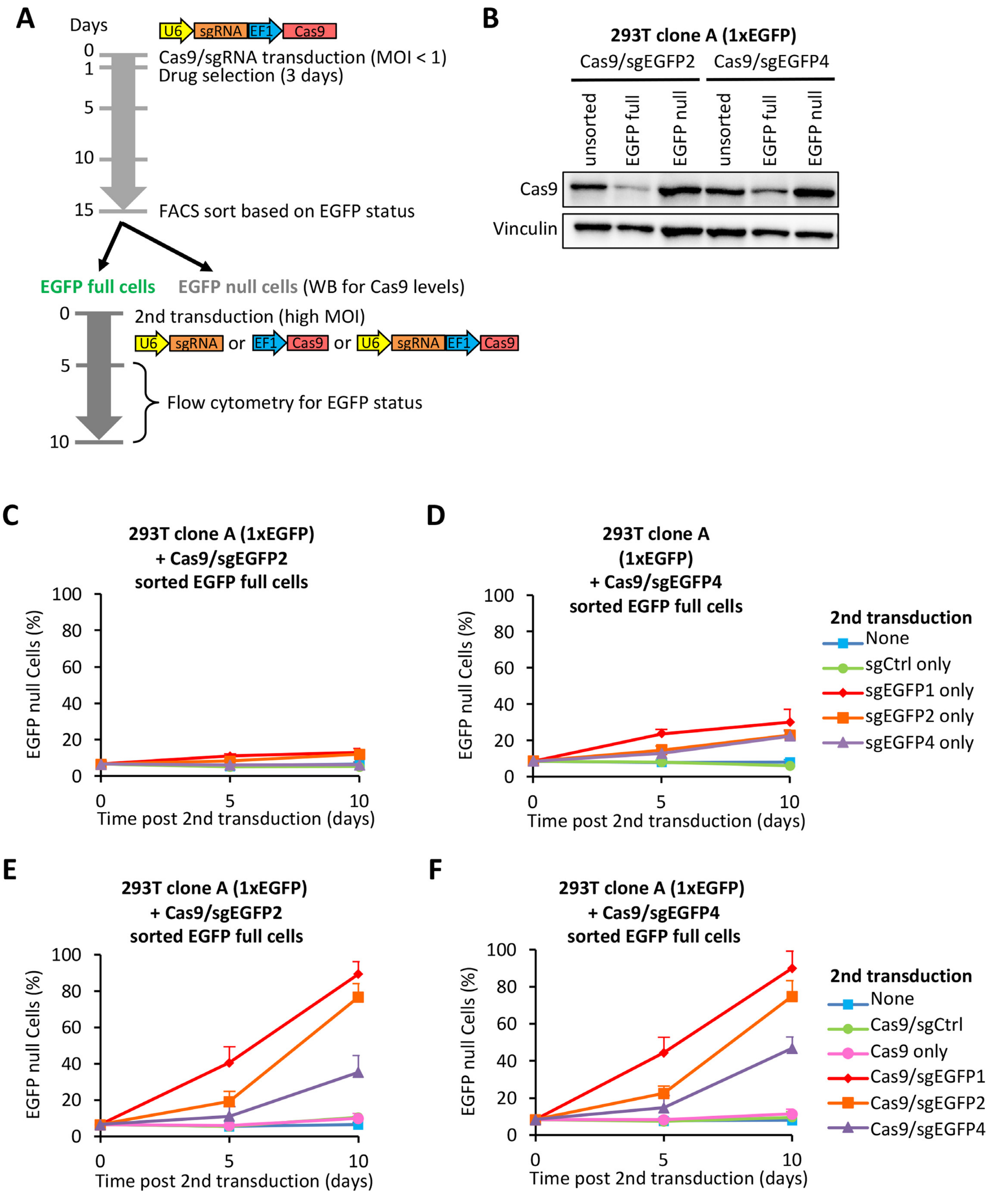


Figure 6. Low CRISPR efficiency is associated with low Cas9/sgRNA expression. (**A**) Schematics for the re-transduction of cells that failed to knock out EGFP. At day 15 after the first round of Cas9/sgRNA transduction (at MOI ≤1), EGFP full and EGFP null cells were separated by FACS sorting. EGFP full cells were re-transduced at high MOI with lentiviral vectors expressing either sgRNA alone, Cas9 alone or both Cas9 and sgRNA together. Cells were analyzed by flow cytometry 10 days after the second transduction. (**B**) Western blot of Cas9 protein level in unsorted, FACS-sorted EGFP full and EGFP null 293T clone A cells at day 15 after the first round of Cas9/sgRNA transduction (MOI ≤1). (**C** and **D**) Second round, high MOI transduction of FACS sorted EGFP full 293T clone A cells (i.e. those without EGFP deletion after the first round of Cas9/sgRNA transduction) with sgRNA alone was insufficient to elicit further EGFP deletion. Data legends are the same for panels C&D. *P*-values from two-way ANOVA are tabulated in Supplementary Table S2C. (**E** and **F**) Second round, high MOI transduction of FACS sorted EGFP full 293T clone A cells (i.e. those without EGFP deletion after the first round of Cas9/sgRNA transduction) with either Cas9 alone or both Cas9 and sgRNA. Further EGFP deletion was seen only when Cas9 and sgRNA were co-expressed at higher level. Data legends are the same for panels E&F. *P*-values from two-way ANOVA are tabulated in Supplementary Table S2C.

		sgRNA potency	
		High	Low
Target copy number	Low	Efficient KO	Inefficient KO
	High	Efficient KO	Inefficient KO
Cas9/sgRNA expression	Low	Efficient KO	Inefficient KO
	High	Efficient KO	Efficient KO

Figure 7. A summary of parameters that determine Cas9/sgRNA knockout efficiency.

cient at heterochromatin (25,27). Thus, gene deletion might be more sensitive to copy number in heterochromatin regions than that in euchromatin regions, and this question warrants further investigation.

In the context of cancer cells, focal amplification could lead to tandem duplication of an oncogene to hundreds of copies (15). This scenario is not modelled by our system, and it is not clear whether a potent sgRNA would remain effective in such an extreme case of gene amplification. Future studies are necessary to corroborate our findings in endogenous genes with different forms of SCNV. Experimentally, such a study currently presents two practical difficulties. One problem is that there are no endogenous genes that show a range of SCNV within an isogenic system. To study endogenous genes with well-documented SCNV, such as *EGFR* and *MYC*, for example, we must resort to analyzing the activity of Cas9/sgRNA in a collection of heterogeneous cancer cell lines. The result could therefore be confounded by cell-line specific factors that are independent of Cas9 activity. Another problem with studying endogenous genes with high copy number is that efficient cutting by Cas9 at sites of gene amplification can often lead to the unintended consequences of cell killing—an effect that is downstream of Cas9 and has to do with differential DNA repair efficiency in different cell lines (12,17). This precludes the accurate assessment of sgRNA efficiency in cells. Fortuitously, cell killing by Cas9 cutting does not appear to be a confounding factor in the isogenic EGFP clones we used in this study. New experimental approaches that combine clonal SCNV analysis within a cell line together with single-cell sequencing could potentially overcome these two problems and enable us to further validate our findings with endogenous genes.

A second feature we discovered is that for a less potent sgRNA, its knockout efficiency is sensitive to the level of both Cas9 and sgRNA expression. In this case, the activity of the system can be quantitatively improved by increasing Cas9 and sgRNA expression. This is in agreement with previous work showing an increased amount of Cas9 and sgRNA plasmid transfection could also lead to better knockout efficiency (24). However, comparison of the activity of a potent sgRNA (sgEGFP1) at low MOI with the activity of two less potent sgRNAs (sgEGFP2 and sgEGFP4) at high MOI suggests that a substantial increase in Cas9 and sgRNA expression was insufficient to fully compensate for the lack of potency in the latter two sgRNAs. Our finding therefore further reinforces the notion that the sgRNA 20-mer sequence is a key determinant in achieving high knockout efficiency in a cell population, such as one would desire in the context of a pooled CRISPR library screen.

Third, we showed that chronic expression of Cas9 and sgRNA at a low level could fail to elicit target mutation even with prolonged Cas9/sgRNA exposure. This phenomenon is attributable to two factors. The first is low sgRNA potency, the second is failure to achieve threshold Cas9 and sgRNA expression. A potent sgRNA (e.g. sgEGFP1) could achieve efficient target knockout in the majority of cells in a polyclonal population even when Cas9 and the sgRNA are expressed at a relatively low level. For the two less potent EGFP sgRNAs, sgEGFP2 and sgEGFP4, efficient knockout is only achieved in cells with higher level of Cas9 and sgRNA expression. This was clearly demonstrated when these sgRNAs were delivered to cells as single-copy integrants, where they caused KO in some cells but not others. Due to integration site effect in the population, sgRNA and Cas9 expression is expected to be heterogeneous in the cell population. Indeed, FACS sorting confirmed that the fraction of cells that experienced KO had higher levels of Cas9 and sgRNA expression, whereas those that did not experience KO had lower levels of Cas9 and sgRNA expression. Importantly, this latter population of cells did not experience KO even after prolonged exposure to Cas9/sgRNA. In these refractory cells, however, elevating the expression of both Cas9 and sgRNA, but not either alone, could effectively induce further KO. Our finding therefore suggests that efficient CRISPR-mediated gene knockout occurs when both Cas9 and sgRNA expression exceeds a critical threshold level, which in turn is determined by the specific potency of a given sgRNA. It is not clear why low levels of continuous Cas9/sgRNA expression fails to mutate the target. Our analysis of off-target site numbers does not support the notion that weaker sgRNAs have more off-target sites and are 'soaked up' by off-target binding. On the other hand, this could be due to the unique enzymatic property of Cas9 (8,9). Further studies are necessary to understand the molecular mechanism underlying this phenomenon.

In the context of a pooled CRISPR library screen, two factors critically influence the saturation of the library: the first being the potency of the sgRNA, the second being the consistency of sgRNA KO efficiency in a polyclonal population of cells. Our findings suggest two important considerations for improving CRISPR KO library.

First, our study strongly suggests that high levels of Cas9 and sgRNA expression could improve target knockout efficiency. Since a CRISPR library is typically delivered as single-copy integration into cells, the expression level of sgRNA and Cas9 could vary significantly from cell to cell due to integration site effect. A highly potent sgRNA is less sensitive to this heterogeneity in expression level and could provide consistent gene KO in the majority of cells. A less potent sgRNA, on the other hand, will be more sensitive to heterogeneous expression level and might only provide efficient KO in high-expression cells. Thus, it would be desirable to improve both Cas9 and sgRNA expression in a CRISPR screen. To improve Cas9 expression, one could either use a clonal cell line that is pre-validated for high Cas9 expression, or a polyclonal population that is pre-selected for high Cas9 expression (e.g. through the use of a Cas9-EGFP expression vector). To achieve higher sgRNA expression level, enhanced or synthetic Pol III promoters (28,29) that can drive higher sgRNA transcription could be incorporated into sgRNA vector design. In addition, barrier insulator element (30) could be incorporated into sgRNA vector design to reduce integration site-dependent repression of sgRNA expression and achieve more homogeneous sgRNA level in a cell population.

Second, our data suggests that improving sgRNA potency in a CRISPR KO library would substantially enhance the penetrance of the library and consequently, the saturation of the screen. Several algorithms have been developed to identify sgRNA sequences that are likely to be potent (2,3,18,19,23). Our limited analysis using one such algorithm (19) supports the utility of rule-based sgRNA sequence selection. Gene tiling studies indicate that only a small fraction of sgRNAs are highly active among all possible sequences (18,19). For a given sgRNA, its potency is determined by a multitude of factors. Several sgRNA-specific sequence features have been found to influence sgRNA activity. These include single and di-nucleotide preference at specific locations within the target site and in the PAM, GC content and melting temperature of the target site, binding sites for epigenetic factors, and potential hairpin structure both within the sgRNA 20-mer and between the sgRNA 20-mer and the constant tracrRNA sequence (18,19,23,31). In addition, the potency of the sgRNA is also influenced by gene-specific features including chromatin structure, nucleosome positioning and DNA accessibility (25,27,31–33). Thus, further refinement of potency prediction algorithms will benefit the selection of highly potent sgRNA sequence for library construction. In addition, large numbers of pre-selected sgRNA sequence can be experimentally tested against their cognate targets using the Sensor assay, which is a massively parallel potency validation approach previously developed for shRNA (34). This will enable the further enrichment of potent sgRNAs for CRISPR library construction.

Third, our data indicate that for a given sgRNA in the library, its behavior is likely to be similar in different cell lines harboring different copy number of its target gene. It has been noted that efficient, on-target cleavage by Cas9 at regions of the cancer genome where focal amplification gives rise to a high copy number of the target gene could lead to cell death due to excessive DNA damage (12,13,17). Beyond this issue, our finding suggests that CRISPR KO screen data using the same library in different cancer cell lines with different SCNVs should be reproducible can be compared across cell lines.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Liam C. Chang, Brandi Carofino, Matthew Brown, David Sun, Karen M. Wolcott and the CCR FACS Core Facility for technical assistance.

FUNDING

US National Cancer Institute Intramural [ZIA BC 011304 to J.L., ZIA BC 011441 to J.M.S.]. Funding for open access charge: National Cancer Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Luo,J. (2016) CRISPR/Cas9: from genome engineering to cancer drug discovery. *Trends Cancer*, **2**, 313–324.
2. Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.
3. Shalem,O., Sanjana,N.E., Hartenian,E., Shi,X., Scott,D.A., Mikkelsen,T.S., Heckl,D., Ebert,B.L., Root,D.E., Doench,J.G. *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.
4. Doudna,J.A. and Charpentier,E. (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*, **346**, 1258096.
5. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
6. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
7. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
8. Sternberg,S.H., LaFrance,B., Kaplan,M. and Doudna,J.A. (2015) Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature*, **527**, 110–113.
9. Sternberg,S.H., Redding,S., Jinek,M., Greene,E.C. and Doudna,J.A. (2014) DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, **507**, 62–67.
10. Shalem,O., Sanjana,N.E. and Zhang,F. (2015) High-throughput functional genomics using CRISPR-Cas9. *Nat. Rev. Genet.*, **16**, 299–311.
11. Hart,T., Chandrashekhar,M., Aregger,M., Steinhart,Z., Brown,K.R., MacLeod,G., Mis,M., Zimmermann,M., Fradet-Turcotte,A., Sun,S. *et al.* (2015) High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell*, **163**, 1515–1526.
12. Munoz,D.M., Cassiani,P.J., Li,L., Billy,E., Korn,J.M., Jones,M.D., Golji,J., Ruddy,D.A., Yu,K., McAllister,G. *et al.* (2016) CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. *Cancer Discov.*, **6**, 900–913.
13. Aguirre,A.J., Meyers,R.M., Weir,B.A., Vazquez,F., Zhang,C.-Z., Ben-David,U., Cook,A., Ha,G., Harrington,W.F., Doshi,M.B. *et al.* (2016) Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.*, **6**, 914–929.
14. Wang,T., Yu,H., Hughes,N.W., Liu,B., Kendirli,A., Klein,K., Chen,W.W., Lander,E.S. and Sabatini,D.M. (2017) Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. *Cell*, **168**, 890–903.
15. Beroukhim,R., Mermel,C.H., Porter,D., Wei,G., Raychaudhuri,S., Donovan,J., Barretina,J., Boehm,J.S., Dobson,J., Urashima,M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.
16. Santaguida,S. and Amon,A. (2015) Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat. Rev. Mol. Cell. Biol.*, **16**, 473–485.
17. Wang,T., Birsoy,K., Hughes,N.W., Krupczak,K.M., Post,Y., Wei,J.J., Lander,E.S. and Sabatini,D.M. (2015) Identification and characterization of essential genes in the human genome. *Science*, **350**, 1096–1101.
18. Doench,J.G., Hartenian,E., Graham,D.B., Tothova,Z., Hegde,M., Smith,I., Sullender,M., Ebert,B.L., Xavier,R.J. and Root,D.E. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.
19. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R. *et al.*

(2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.
20. Sanjana,N.E., Shalem,O. and Zhang,F. (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods*, **11**, 783–784.
21. De Ravin,S.S., Su,L., Theobald,N., Choi,U., Macpherson,J.L., Poidinger,M., Symonds,G., Pond,S.M., Ferris,A.L., Hughes,S.H. *et al.* (2014) Enhancers are major targets for murine leukemia virus vector integration. *J. Virol.*, **88**, 4504–4513.
22. Brinkman,E.K., Chen,T., Amendola,M. and van Steensel,B. (2014) Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.*, **42**, e168.
23. Chari,R., Mali,P., Moosburner,M. and Church,G.M. (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods*, **12**, 823–826.
24. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
25. Kuscu,C., Arslan,S., Singh,R., Thorpe,J. and Adli,M. (2014) Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.*, **32**, 677–683.
26. Lewinski,M.K., Yamashita,M., Emerman,M., Ciuffi,A., Marshall,H., Crawford,G., Collins,F., Shinn,P., Leipzig,J., Hannenhalli,S. *et al.* (2006) Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog.*, **2**, e60.
27. Jensen,K.T., Fløe,L., Petersen,T.S., Huang,J., Xu,F., Bolund,L., Luo,Y. and Lin,L. (2017) Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Lett.*, **591**, 1892–1901.
28. Schwartz,C.M., Hussain,M.S., Blenner,M. and Wheeldon,I. (2016) Synthetic RNA polymeprase III propmoters facilitate high-efficiency CRISPR-Cas9-mediated genome editing in Yarrowia lipolytica. *ACS Synth. Biol.*, **5**, 356–359.
29. Xia,X.G., Zhou,H., Ding,H., Affar,E.B., Shi,Y. and Xu,Z. (2003) An enhanced U6 promoter for synthesis of short hairpin RNA. *Nucleic Acids Res.*, **31**, e100.
30. Emery,D.W. (2011) The use of chromatin insulators to improve the expression and safety of integrating gene transfer vectors. *Hum. Gene Ther.*, **22**, 761–774.
31. Thyme,S.B., Akhmetova,L., Montague,T.G., Valen,E. and Schier,A.F. (2016) Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.*, **7**, 11750.
32. Isaac,R.S., Jiang,F., Doudna,J.A., Lim,W.A., Narlikar,G.J. and Almeida,R. (2016) Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *eLife*, **5**, e13450.
33. Horlbeck,M.A., Witkowsky,L.B., Guglielmi,B., Replogle,J.M., Gilbert,L.A., Villalta,J.E., Torigoe,S.E., Tjian,R. and Weissman,J.S. (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *eLife*, **5**, e12677.
34. Fellmann,C., Zuber,J., McJunkin,K., Chang,K., Malone,C.D., Dickins,R.A., Xu,Q., Hengartner,M.O., Elledge,S.J., Hannon,G.J. *et al.* (2011) Functional identification of optimized RNAi triggers using a massively parallel sensor assay. *Mol. Cell*, **41**, 733–746.