Method

Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq

Daesik Kim,[1,2] Sojung Kim,[1,2] Sunghyun Kim,[1] Jeongbin Park,[3] and Jin-Soo Kim[1,2]

[1]*Center for Genome Engineering, Institute for Basic Science, Seoul 151-747, South Korea;* [2]*Department of Chemistry, Seoul National University, Seoul 151-747, South Korea;* [3]*Department of Chemistry, Hanyang University, Seoul 133-791, South Korea*

We present multiplex Digenome-seq to profile genome-wide specificities of up to 11 CRISPR-Cas9 nucleases simultaneously, saving time and reducing cost. Cell-free human genomic DNA was digested using multiple sgRNAs combined with the Cas9 protein and then subjected to whole-genome sequencing. In vitro cleavage patterns, characteristic of on- and off-target sites, were computationally identified across the genome using a new DNA cleavage scoring system. We found that many false-positive, bulge-type off-target sites were cleaved by sgRNAs transcribed from an oligonucleotide duplex but not by those transcribed from a plasmid template. Multiplex Digenome-seq captured many bona fide off-target sites, missed by other genome-wide methods, at which indels were induced at frequencies <0.1%. After analyzing 964 sites cleaved in vitro by these sgRNAs and measuring indel frequencies at hundreds of off-target sites in cells, we propose a guideline for the choice of target sites for minimizing CRISPR-Cas9 off-target effects in the human genome.

[Supplemental material is available for this article.]

RNA-guided endonucleases (RGENs), derived from the type II CRISPR (clustered regularly interspaced repeat)-CRISPR–associated (Cas) prokaryotic adaptive immune system, enable genome editing in cultured cells and whole organisms (Cho et al. 2013a,b; Cong et al. 2013; Hwang et al. 2013; Jiang et al. 2013; Jinek et al. 2013; Mali et al. 2013b; Kim and Kim 2014). These nucleases, however, can induce off-target mutations, limiting their utility in research and medicine (Cradick et al. 2013; Fu et al. 2013; Hsu et al. 2013; Pattanayak et al. 2013; Cho et al. 2014). Recently, we and other groups independently presented several different methods for profiling genome-wide specificities of RGENs, which consist of gRNAs and the Cas9 protein originated from *Streptococcus pyogenes*, in human cells (Frock et al. 2015; Kim et al. 2015; Ran et al. 2015; Tsai et al. 2015; Wang et al. 2015). All of these methods rely on high-throughput sequencing of human genomic DNA that is cleaved by RGENs in cells or in vitro. High-throughput, genome-wide translocation sequencing (HTGTS) is based on translocations induced by nonhomologous end-joining (NHEJ) repair of two concurrent DNA double-strand breaks (DSBs) in cells (Frock et al. 2015). Both genome-wide, unbiased identification of DSBs enabled by sequencing (GUIDE-seq) (Tsai et al. 2015) and integration-deficient lentiviral vector (IDLV) capture (Wang et al. 2015) rely on NHEJ-mediated insertions of small duplex oligonucleotides or lentiviral vectors, respectively, at cleavage sites. Direct in situ breaks labeling, enrichment on streptavidin, and next-generation sequencing (BLESS) is a method of capturing DSBs in fixed cells (Crosetto et al. 2013; Ran et al. 2015). Digenome-seq (digested genome sequencing) relies on whole-genome sequencing (WGS) of cell-free genomic DNA digested in vitro using a nuclease of interest (Kim et al. 2015). Although each of these methods has successfully identified genome-wide off-target sites in human cells, it is unknown how comprehensive and sensitive each method is: Only one gRNA was analyzed by three different methods thus far (Gabriel et al. 2015; Kim et al. 2015).

In this study, we performed multiplex Digenome-seq to profile genome-wide specificities of up to 11 CRISPR-Cas9 nucleases at once. Multiplex Digenome-seq was more comprehensive than other methods, revealing many bona fide off-target sites that had been missed by GUIDE-seq or HTGTS. After analyzing hundreds of off-target sites, we found a rule of thumb for choosing target sites to minimize genome-wide off-target effects of CRISPR-Cas9.

Results

Improving Digenome-seq

First, we developed a scoring system to computationally identify in vitro cleavage sites across the human genome using WGS data. Although our original Digenome-seq analysis was highly reproducible, some sites with heterogeneous cleavage patterns or with a low sequencing depth were often missed. We found that these sites could be identified by assuming that Cas9 can produce 1- or 2-nucleotide (nt) overhangs in addition to blunt ends. We assigned a DNA cleavage score to each nucleotide position, based on patterns of alignments of sequence reads (Supplemental Fig. 1; Supplemental Scripts 1). Our improved program successfully captured many additional sites that had been missed previously. A genome-wide plot of cleavage scores showed that false-positive sites obtained with undigested genomic DNA were still extremely rare (Fig. 1A): A few false-positive sites identified in the entire genome contained naturally occurring indels in the genomic DNA and could be filtered out with ease. Sequence reads around naturally occurring indel sites cannot be mapped properly, often producing false-positive sites. As shown by two independent Digenome-seq analyses, cleavage scores across the human genome were highly reproducible ($R^2 = 0.89$) (Supplemental Fig. 2).

We also found that sgRNAs transcribed using a plasmid template in a Digenome-seq analysis did not cleave any of the false-positive, bulge-type off-target sites, with a missing nucleotide

Corresponding author: jskim01@snu.ac.kr
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.199588.115. Freely available online through the *Genome Research* Open Access option.

26:406–415 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/16; www.genome.org

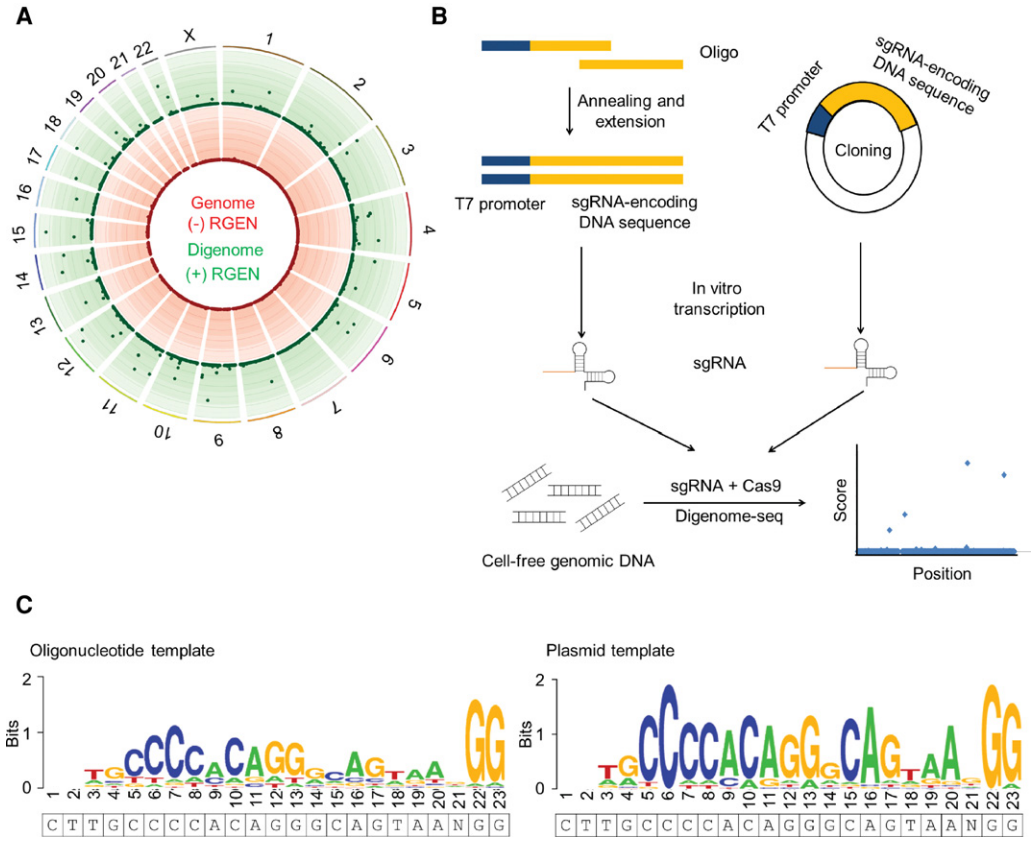


Figure 1. Improving Digenome-seq analysis. (*A*) Genome-wide Circos plots of in vitro DNA cleavage scores. Human genomic DNA (red) or RGEN-digested genomic DNA (green) was subjected to whole-genome sequencing. The few false-positive sites found in intact genomic DNA are not visible because their cleavage scores are not high. (*B*) Schematic overview of Digenome-seq using sgRNA transcribed from an oligonucleotide duplex or a plasmid. (*C*) Sequence logos of cleavage sites obtained using sgRNA transcribed from an oligonucleotide duplex or a plasmid.

compared to the on-target site (Lin et al. 2014), which were captured with those transcribed using an oligonucleotide duplex (Fig. 1B; Supplemental Fig. 3). Apparently, the latter sgRNAs were heterogeneous, containing truncated molecules transcribed from synthesis-failed oligonucleotides. As a result, cleavage sites identified using sgRNAs transcribed from a plasmid template were more highly homologous to its on-target site than those identified using sgRNAs transcribed from an oligonucleotide template (Supplemental Table 1), as shown by sequence logos obtained computationally by comparing DNA sequences around cleavage sites (Fig. 1C). Thus, the use of a new cleavage scoring system and sgRNAs transcribed from plasmid templates substantially reduced the number of false-negative sites and false-positive sites, respectively.

Multiplex Digenome-seq

Unlike other methods, Digenome-seq can be multiplexed without increasing sequencing depth proportionally to the number of nucleases. We chose 10 sgRNAs that had been analyzed individually using GUIDE-seq (Tsai et al. 2015), which is likely to be more sensitive than IDLV capture and other methods. Note that BLESS captures a snapshot of DSBs at a given moment of cell fixation (Crosetto et al. 2013) and that HTGTS relies on two concurrent DSBs, rather than one, induced in a single cell (Chiarle et al. 2011). We digested human genomic DNA with a mixture of the Cas9 protein, the 10 sgRNAs, and one additional sgRNA targeted to the *HBB* gene, which we had analyzed in our previous study (Kim et al. 2015), and carried out two independent WGS analyses (Fig. 2A). Genome-wide in vitro cleavage sites were identified computationally using the scoring system. A total of 964 sites were found in the human genome (Supplemental Table 2). All of these sites were then classified computationally according to the edit distance from the on-target sites (Fig. 2A; Supplemental Table 2; Supplemental Scripts 2), which ranged from zero (on-target sites) to 10. The mean edit distance was 14.4, which is the average edit distance between any two on-target sites among the 11 on-target sites. We recommend choosing on-target sites that differ from each other by at least 11 nt in a multiplex Digenome-seq analysis, to facilitate classification of cleavage sites by edit distance.

Unlike GUIDE-seq and other methods, which require a filtering step to discard up to 90% of captured sites with poor homology to on-target sites, multiplex Digenome-seq does not filter sites and sorts sites based on edit distance. Yet, the 964 sites were divided into 11 groups unambiguously. Furthermore, each of the 11 groups of in vitro cleavage sites was highly homologous to one of the 11 on-target sequences. Thus, de novo motifs or sequence logos obtained by comparing sequences in each group showed matches with respective target sequences at almost every nucleotide position (Fig. 2A). This result suggests that the 5′-terminal 10-nt region in a 23-nt target sequence contributes to the specificity of RGENs, albeit to a lesser extent than do the protospacer-adjacent motif (PAM), recognized by Cas9, and the PAM-proximal

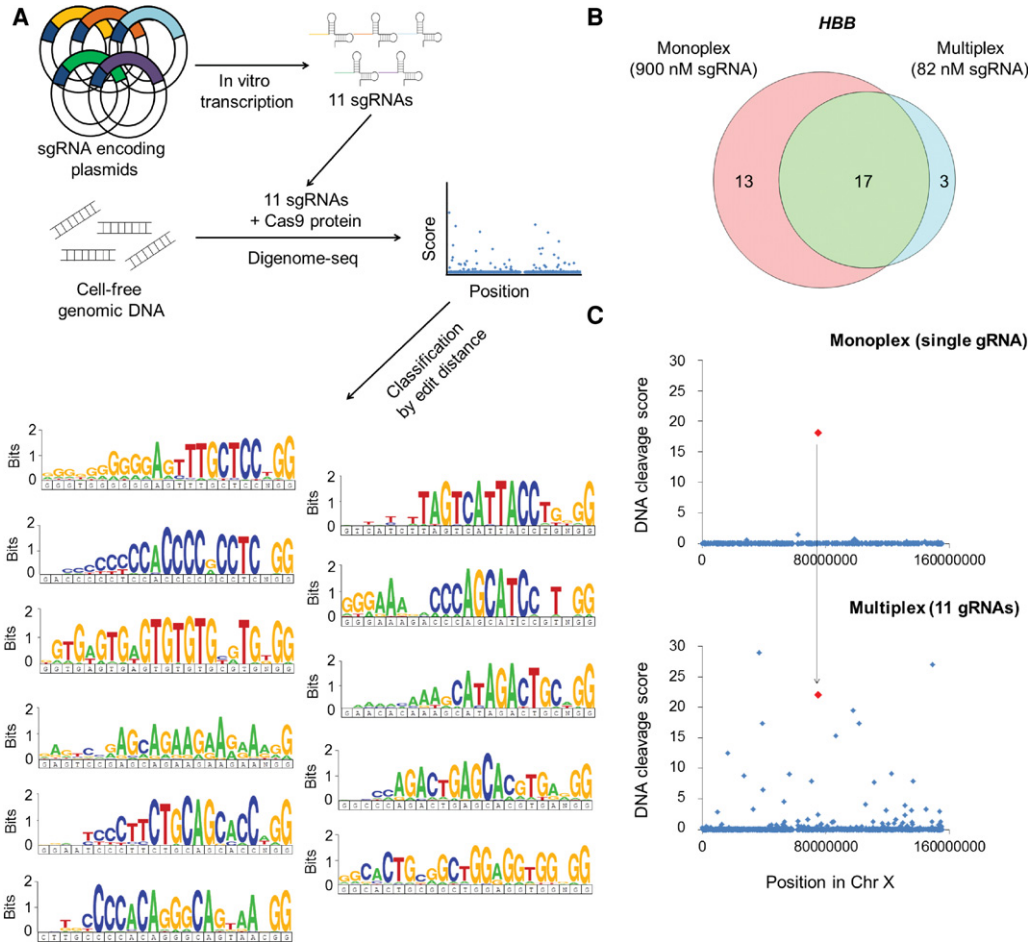


Figure 2. Multiplex Digenome-seq. (*A*) Schematic overview of multiplex Digenome-seq. (*B*) A Venn diagram showing the number of in vitro cleavage sites captured by monoplex and multiplex Digenom-seq analyses. (*C*) In vitro DNA cleavage scores across Chromosome X obtained by monoplex or multiplex Digenome-seq.

10-nt "seed" region (Jinek et al. 2012; Cong et al. 2013; Pattanayak et al. 2013). We noted that all but one of the 964 sites cleaved by 11 RGENs contained a PAM or PAM-like sequence, that is, 5′-NGG-3′ or 5′-NNG-3′/5′-NGN-3′, respectively, 3 nt downstream from cleavage positions, confirming that DSB ends produced in vitro cannot be trimmed by endogenous DNA repair systems, unlike DSBs induced in cells. Thus, in vitro cleavage sites can be pinpointed without a computational search for homologous sequences, a feature not shared with GUIDE-seq and HTGTS.

First, we checked whether an sgRNA in a pool can cleave its on-target and off-target sites. Seventeen out of 30 (57%) sites that were cleaved using the single *HBB*-specific sgRNA alone at high concentration (900 nM) plus Cas9 (300 nM) were also captured by multiplex Digenome-seq using the same sgRNA at low concentration (82 nM) (Fig. 2B,C). Note that more sites are captured at higher concentration of sgRNAs (Kim et al. 2015). Importantly, all four off-target sites and the on-target site that had been validated using targeted deep sequencing in our previous study (Kim et al. 2015) were identified by multiplex Digenome-seq. This result suggests that each sgRNA in a pool of up to 11 sgRNAs can guide Cas9 to most of its on-target and off-target sites independently from each other, supporting the basis of multiplexing.

In vitro cleavage sites

The 11 sgRNAs showed a wide spectrum of genome-wide specificities: The number of cleavage sites per sgRNA in the human genome ranged from 13 to 302 (Fig. 3A; Supplemental Table 2). As expected, all of the 11 on-target sites and most of the sites with one or two mismatches (but with no DNA or RNA bulge), identified in the human genome using Cas-OFFinder (Bae et al. 2014), were captured by Digenome-seq (Fig. 3B). However, sites with more than three mismatches were rarely captured. The fraction of Digenome-captured sites decreased exponentially as the number of mismatches increased from three to six (Fig. 3B). We also found that sites with two or more mismatches in the seed region were much less likely to be cleaved in vitro than those with zero or one mismatch in the seed ($P < 0.01$, Student's *t*-test). Out of the 964 sites cleaved in vitro using the 11 sgRNAs transcribed from plasmid templates, only a single site had a missing nucleotide (an RNA bulge) compared with the on-target site.

Interestingly, we found a strong correlation ($R^2 = 0.93$) between the number of Digenome-captured sites and the number of homologous sites with six or fewer nucleotide mismatches in the human genome (defined as "orthogonality" in Tsai et al. 2015) (Fig. 3C). The six sgRNAs with fewer than 13,000 such homologous sites in the human genome were much more specific

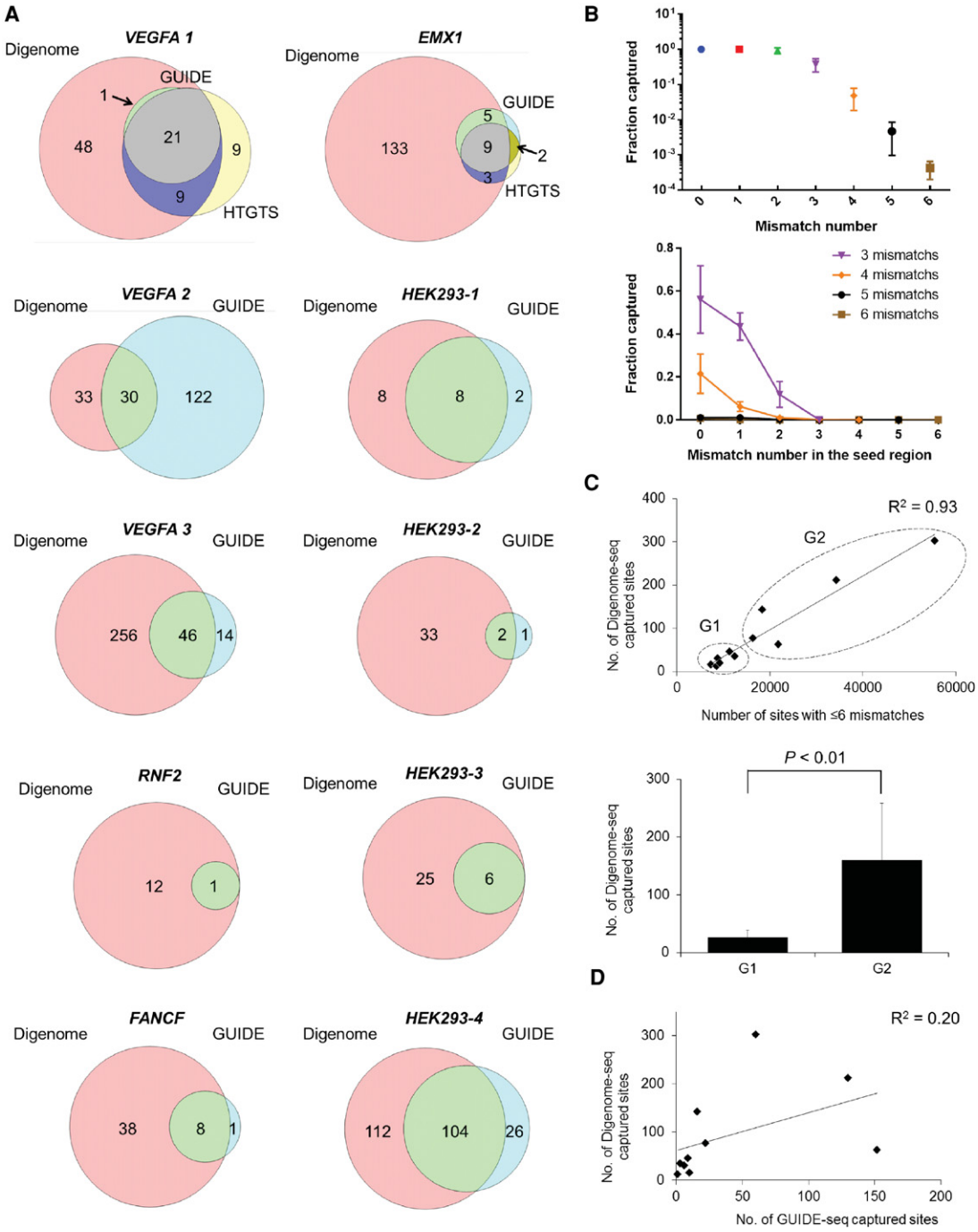


Figure 3. Analysis of multiplex Digenome-captured sites. (*A*) Venn diagrams showing the number of sites captured by Digenome-seq, GUIDE-seq, and HTGTS. (*B*) Fractions of sites captured by Digenome-seq according to the total mismatch number (*top*) and the mismatch number in the seed region (*bottom*). (*C*, *top*) Scatterplot of the number of sites with six or fewer mismatches in the human genome versus the number of Digenome-captured sites. (*Bottom*) Eleven RGEN target sites were divided into two groups, G1 and G2 (those with fewer than 13,000 and 16,000 sites, respectively, harboring six or fewer mismatches in the human genome). Error bars, SEM. The *P*-value was calculated by Student's *t*-test. (*D*) Scatterplot of the number of GUIDE-seq captured sites versus the number of Digenome-seq captured sites.

($P < 0.01$, Student's *t*-test), cleaving 46 or fewer sites in vitro (28 sites/sgRNA, on average), than the other five sgRNAs with more than 16,000 such sites, cleaving 63 or more sites in vitro (161 sites/sgRNA, on average) (Fig. 3C). This result is seemingly in contrast with the poor correlation ($R^2 = 0.29$) observed between the number of GUIDE-seq–positive sites and the orthogonality of the target site relative to the human genome (Supplemental Fig. 4; Tsai et al. 2015). The *VEGFA2* site was an outlier, at least partially causing the poor correlation. We noted, however, that the five most specific sgRNAs revealed by GUIDE-seq, cleaving 10 or fewer sites in cells, were coincident with the most specific sgRNAs revealed by Digenome-seq.

Digenome-seq vs. other methods

On average, multiplex Digenome-seq successfully identified 80 ± 8% of sites captured previously by GUIDE-seq (Fig. 3A). For example, all of the GUIDE-captured sites using the three sgRNAs specific to the *VEGFA1*, *RNF2*, and *HEK293-3* sites were identified by Digenome-seq. In addition, multiplex Digenome-seq captured a total of 703 new sites (70 sites per sgRNA, on average) that had been missed by GUIDE-seq (Fig. 3A). As a result, GUIDE-seq had captured 25 ± 6% of sites identified by multiplex Digenome-seq. The *RNF2*-specific sgRNA was a striking example. Two independent GUIDE-seq analyses had failed to capture any single off-target site, whereas Digenome-seq identified 12 cleavage sites in addition to the on-target site. In fact, we observed a poor correlation ($R^2 = 0.20$) between the number of Digenome-positive sites and that of GUIDE-positive sites (Fig. 3D). It is likely that many additional sites that are cleaved in vitro and, thereby, captured by Digenome-seq are not accessible in cells.

Digenome-seq yielded more candidate off-target sites than GUIDE-seq for nine out of 10 sgRNAs but still was not comprehensive. (The *HBB* sgRNA had not been analyzed by GUIDE-seq.) Thus, in aggregate, GUIDE-seq had captured a total of 168 sites that were missed by Digenome-seq. Two sgRNAs targeted to the *VEGFA1* and *EMX1* sites had also been analyzed by HTGTS (Fig. 3A). Most sites (31 out of 40 sites for *VEGFA1* and 17 out of 19 sites for *EMX1*) captured by at least one of the other two methods were also identified by Digenome-seq, but it missed nine and two sites, respectively. It is possible that some of these sites were false positives that resulted from PCR primer–dependent artifacts or naturally occurring DSBs, intrinsic limitations of GUIDE-seq and HTGTS. Many of these sites, especially the two *EMX1* off-target sites commonly identified by the other two methods, however, would have been missed by multiplex Digenome-seq, owing to a low sequencing depth at these particular sites (Supplemental Fig. 5) or the low concentration (82 nM) of the sgRNA used in this study. These problems could be alleviated by performing WGS at a higher sequencing depth or by using a higher concentration of the sgRNA in a monoplex analysis, respectively.

The *VEGFA2*-specific sgRNA was the only exception to the rule that Digenome-seq captures more candidate sites than GUIDE-seq. Thus, GUIDE-seq had identified 122 sites that were missed by Digenome-seq. The target sequence was unusual, consisting of a stretch of cytosines. Many sequence reads, obtained by WGS, at homopolymer sites can be discarded by a mapping program. GUIDE-seq may still capture these sites because PCR is used to amplify oligonucleotide-captured sites.

We also compared cleavage sites identified in this study with those captured by chromatin immunoprecipitation sequencing (ChIP-seq) using catalytically dead Cas9 (dCas9) (Kuscu et al. 2014). Strikingly, a vast majority of Cas9-cleaved sites (288 sites, 98%) identified by Digenome-seq were not bound by dCas9 (Supplemental Fig. 6). This result suggests that DNA binding by Cas9 is uncoupled from DNA cleavage and that ChIP-seq using dCas9 is inappropriate for assessing genome-wide specificities of Cas9 RGENs (Tsai et al. 2015), although it may still be useful for profiling the specificities of dCas9-based transcription factors and epigenome regulators.

Validation of off-target sites in cells

We then investigated, using a next-generation sequencing (NGS) platform, whether each sgRNA plus Cas9 could induce off-target indels in HeLa cells at some of these Digenome-captured and GUIDE-captured sites (Table 1; Supplemental Table 3). We chose candidate off-target sites with a fewer mismatches, irrespective of DNA cleavage scores in this analysis. Indels were detected over background noise levels caused by sequencing errors at 116 out of 132 (88%) sites commonly captured using Digenome-seq and GUIDE-seq. In contrast, many of the sites captured by Digenome-seq alone and GUIDE-seq alone were not validated by targeted deep sequencing. Indels were induced above noise levels at 21 out of 127 (17%) sites captured by Digenome-seq alone and at 23 out of 45 (51%) sites captured by GUIDE-seq alone, confirming that neither of the two methods was comprehensive. Thus, the overall validation rate was 53% [(21 + 116)/(127 + 132)], with Digenome-seq, or 79% [(23 + 116)/(45 + 132)], with GUIDE-seq.

Indel frequencies at most of these validated sites were <1%, much lower than those at respective on-target sites. For example, the *RNF2*-targeted sgRNA induced indels at the on-target site and two off-target sites identified in this study with a frequency of 68%, 0.25%, and 0.09%, respectively (Supplemental Fig. 7). It still is possible that indels could be induced at NGS-invalidated sites with frequencies below noise levels (0.001%–4%, depending on the site).

Next, we investigated whether off-target sites validated in HeLa cells in this study had been missed by GUIDE-seq performed in U2OS cells or HEK 293T cells, owing to the differential chromatin states. We chose five validated sites that were identified using four sgRNAs in this study but had been missed by GUIDE-seq. Indels were induced at four of these five sites in HEK 293T cells (Supplemental Table 4). These results show that the discrepancy between this study and Tsai et al. (2015) cannot be attributed to the use of different cell lines and that Digenome-seq can capture bona fide off-target sites missed by GUIDE-seq.

To reduce off-target effects, we replaced sgRNAs with versions containing two extra guanines at the 5′ terminus (termed ggX_{20} sgRNAs) (Fig. 4A; Cho et al. 2014). These modified sgRNAs were more specific than their respective GX_{19} sgRNAs by up to 598-fold (Fig. 4B–G). It is of note that off-target indels were not detected above noise levels with the *RNF2*-specific ggX_{20} sgRNA (Fig. 4D).

Indel frequencies at off-target sites

The large number of NGS-validated (160) and NGS-invalidated (144) off-target sites and indel frequencies determined at these sites allowed us to examine off-target effects in detail. A plot of the number of mismatches versus the ratio of indel frequencies at off-target relative to on-target sites showed that off-target sites with up to two mismatches were cleaved efficiently in cells (median indel frequency = 5.38%) and that those with three or more mismatches were poorly cleaved (median indel frequency = 0.14% or lower, respectively) (Fig. 5A). Indel frequencies at on-target sites were 60 ± 7%. Interestingly, mismatches in the validated and invalidated sites were almost evenly distributed in the PAM-distal and PAM-proximal regions. For both validated and invalidated sites with three or more mismatches, the PAM-distal region was as important as the seed region (Fig. 5B,C). Thus, indel frequencies at sites with zero or one mismatch in the seed region were as low as those at sites with two or more mismatches.

In summary, we improved Digenome-seq to reduce the number of false-positive and false-negative sites by developing an in vitro DNA cleavage scoring system and using sgRNAs transcribed from a plasmid template rather than a synthetic oligonucleotide duplex. We multiplexed Digenome-seq by digesting cell-free genomic DNA with a mixture of 11 sgRNAs, which revealed 70

Table 1. Validation of off-target sites in human cells using next-generation sequencing

	Digenome only	Digenome and GUIDE	GUIDE only
VEGFA1			
Total captured sites	57	22	0
No. of NGS-tested sites	15	22	0
No. of validated sites	6	20	0
VEGFA2			
Total captured sites	33	30	122
No. of NGS-tested sites	8	22	14
No. of validated sites	0	22	10
VEGFA3			
Total captured sites	256	46	14
No. of NGS-tested sites	18	27	9
No. of validated sites	4	22	5
EMX1			
Total captured sites	129	14	2
No. of NGS-tested sites	16	12	2
No. of validated sites	3	9	2
FANCF			
Total captured sites	38	8	1
No. of NGS-tested sites	8	8	1
No. of validated sites	1	8	0
RNF2			
Total captured sites	12	1	0
No. of NGS-tested sites	12	1	0
No. of validated sites	2	1	0
HEK293-1			
Total captured sites	8	8	2
No. of NGS-tested sites	3	8	2
No. of validated sites	1	7	2
HEK293-2			
Total captured sites	33	2	1
No. of NGS-tested sites	16	2	1
No. of validated sites	1	2	0
HEK293-3			
Total captured sites	25	6	0
No. of NGS-tested sites	14	6	0
No. of vacated sites	2	6	0
HEK293-4			
Total captured sites	112	104	26
No. of NGS-tested sites	17	24	16
No. of validated sites	1	19	4
Total			
Total captured sites	703	241	168
No. of NGS-tested sites	127	132	45
No. of validated sites	21	116	23

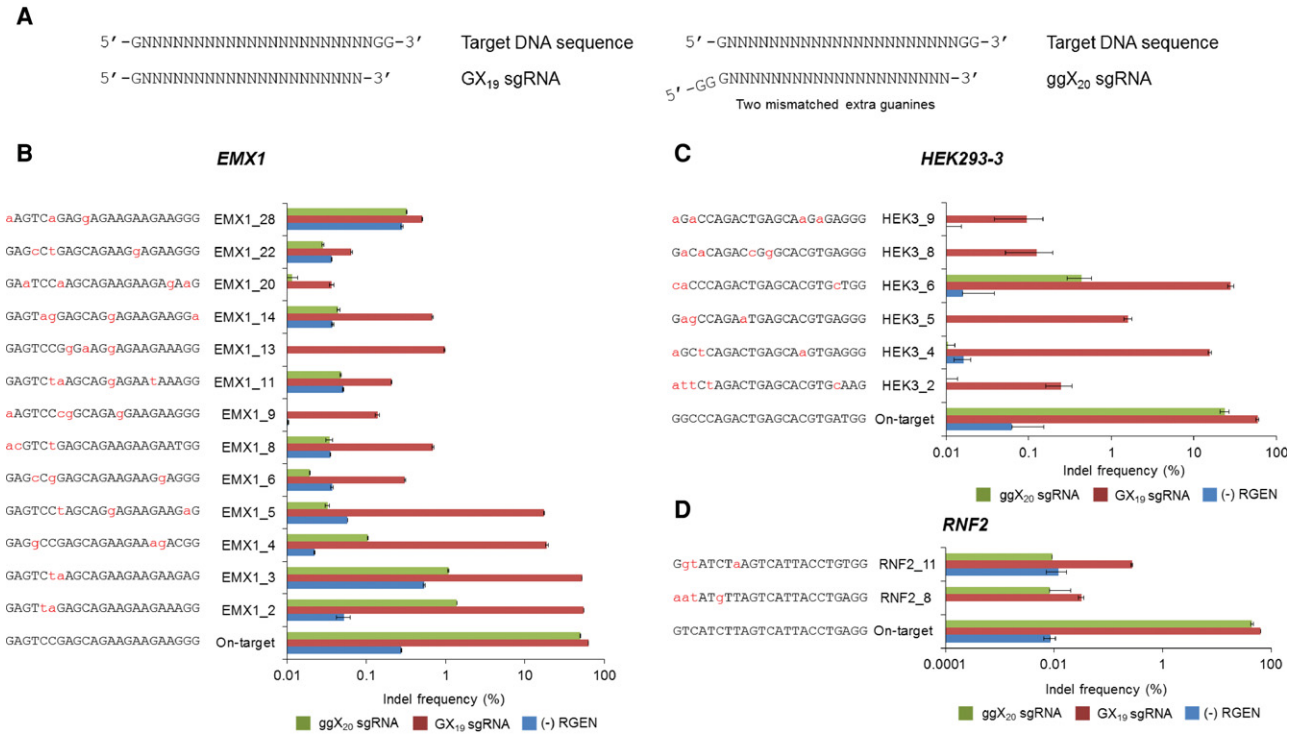
Indel frequencies at off-target sites captured by Digenome-seq and GUIDE-seq were measured in human cells. Validated off-target sites were those with indel frequencies above noise indel frequencies obtained in the absence of RGEN transfection. "Digenome only" and "GUIDE only" sites exclude "Digenome and GUIDE" sites that were commonly identified by the two methods.

additional cleavage sites per sgRNA, on average, that had not been captured by GUIDE-seq. Off-target indels were induced at many of these sites in RGEN-transfected human cells. After carefully examining indel frequencies, mismatch numbers, and mismatch positions at hundreds of off-target sites, we conclude that both the PAM-distal region and the PAM-proximal seed region are important to RGEN specificities. We also noted that sites with two or more mismatches in the seed region are much less likely to be cleaved in vitro than those with no or one mismatch, regardless of the total number of mismatches.

Discussion

To disrupt a gene of interest using RGENs, one should choose target sites with no or few off-target effects. First, a desired target site should have only a few or no off-target sites in the genome. Second, indel frequencies at these off-target sites should be much lower than the frequency at the on-target site. Our results suggest that a unique site that has fewer than 13,000 homologous sites with up to six mismatches in the human genome and that has no homologous sites with up to two mismatches is desirable to minimize off-target effects. Out of 1715 targetable sites containing the 5′-NGG-3′ PAM in the four genes examined in this study, 368 (21.5%) sites satisfy these criteria (Supplemental Table 5). In addition, we present an off-target score (Supplemental Table 6; Supplemental Scripts 3) that accounts for numbers of potential off-target sites in the genome, fractions of these sites captured by Digenome-seq (Fig. 3B), and median indel frequencies at these sites (Fig. 5A). One should choose a low-score site to avoid or reduce off-target effects. A web-based computer program that shows off-target scores in a gene of interest is available at our website (www.rgenome.net/digenome).

A comprehensive analysis of genome-wide off-target sites using Digenome-seq and indel frequencies at these sites using targeted deep sequencing allowed us to define an off-target effect index



E *EMX1*

Target	Specificity ratio (on/off)		
	GX_{19}	ggX_{20}	ggX_{20}/GX_{19}
EMX1_2	1.2	36.0	31.2
EMX1_3	1.2	46.0	37.7
EMX1_4	3.4	477.0	142.0
EMX1_5	3.7	1543.4	420.1
EMX1_6	208.7	2560.4	12.3
EMX1_8	92.9	1454.4	15.6
EMX1_9	461.2	5157.4	11.2
EMX1_11	308.7	1055.8	3.4
EMX1_13	65.6	5406.6	82.4
EMX1_14	94.1	1125.4	12.0
EMX1_20	1750.7	4313.0	2.5
EMX1_22	982.2	1760.9	1.8
EMX1_28	125.0	155.3	1.2

F *HEK293-3*

Target	Specificity ratio (on/off)		
	GX_{19}	ggX_{20}	ggX_{20}/GX_{19}
HEK3_2	240.0	3220.9	13.4
HEK3_4	3.8	2264.1	597.9
HEK3_5	36.9	8716.1	236.3
HEK3_6	2.1	54.8	26.0
HEK3_8	475.5	5618.8	11.8
HEK3_9	628.8	2803.3	4.5

G *RNF2*

Target	Specificity ratio (on/off)		
	GX_{19}	ggX_{20}	ggX_{20}/GX_{19}
RNF2_8	1985.1	5312.7	2.7
RNF2_11	234.9	4749.7	20.2

Figure 4. Indel frequencies determined using targeted deep sequencing at off-target sites. (*A*) Schematic of conventional sgRNAs (GX_{19} sgRNA) and modified sgRNAs (ggX_{20} sgRNA). Indel frequencies at NGS-validated on- and off-target sites for the *EMX1* (*B*), *HEK293-3* (*C*), and *RNF2* sgRNAs (*D*). Specificity ratios calculated by dividing the indel frequency at the on-target site with that at an off-target site for the *EMX1* (*E*), *HEK293-3* (*F*), and *RNF2* sgRNAs (*G*).

(OTI), a ratio of the sum of indel frequencies at off-target sites to the on-target indel frequency. The OTI for the *RNF2*-specific GX_{19} or ggX_{20} sgRNA is 0.005 [(0.25% + 0.09%)/68%] or 0.0005 [(0.01% + 0.01%)/44%], respectively, showing that these sgRNAs are highly specific. The OTIs for two other GX_{19} sgRNAs, determined from our previous data (Kim et al. 2015), are 2.75 (*HBB*) and 4.44 (*VEGFA1*) (Supplemental Note 1), showing that these sgRNAs are not specific. Thus, OTIs can span more than three orders of magnitude, depending on target sites and sgRNA compositions.

To minimize or avoid RGEN off-target effects, we and others have proposed various methods, which include dimeric Cas9 systems (paired Cas9 nickases [Mali et al. 2013a; Ran et al. 2013; Cho et al. 2014] and dCas9-FokI [Guilinger et al. 2014; Tsai et al. 2014]), delivery of RGEN ribonucleoproteins (RNPs) (Kim et al. 2014; Ramakrishna et al. 2014; Zuris et al. 2015), and modified guide RNAs (Cho et al. 2014; Fu et al. 2014). The dimeric systems require two active sgRNAs and two adjacent PAMs, limiting targetable sites. Our Digenome-seq data suggest that the choice of unique or orthogonal target sites is also important and can be sufficient to

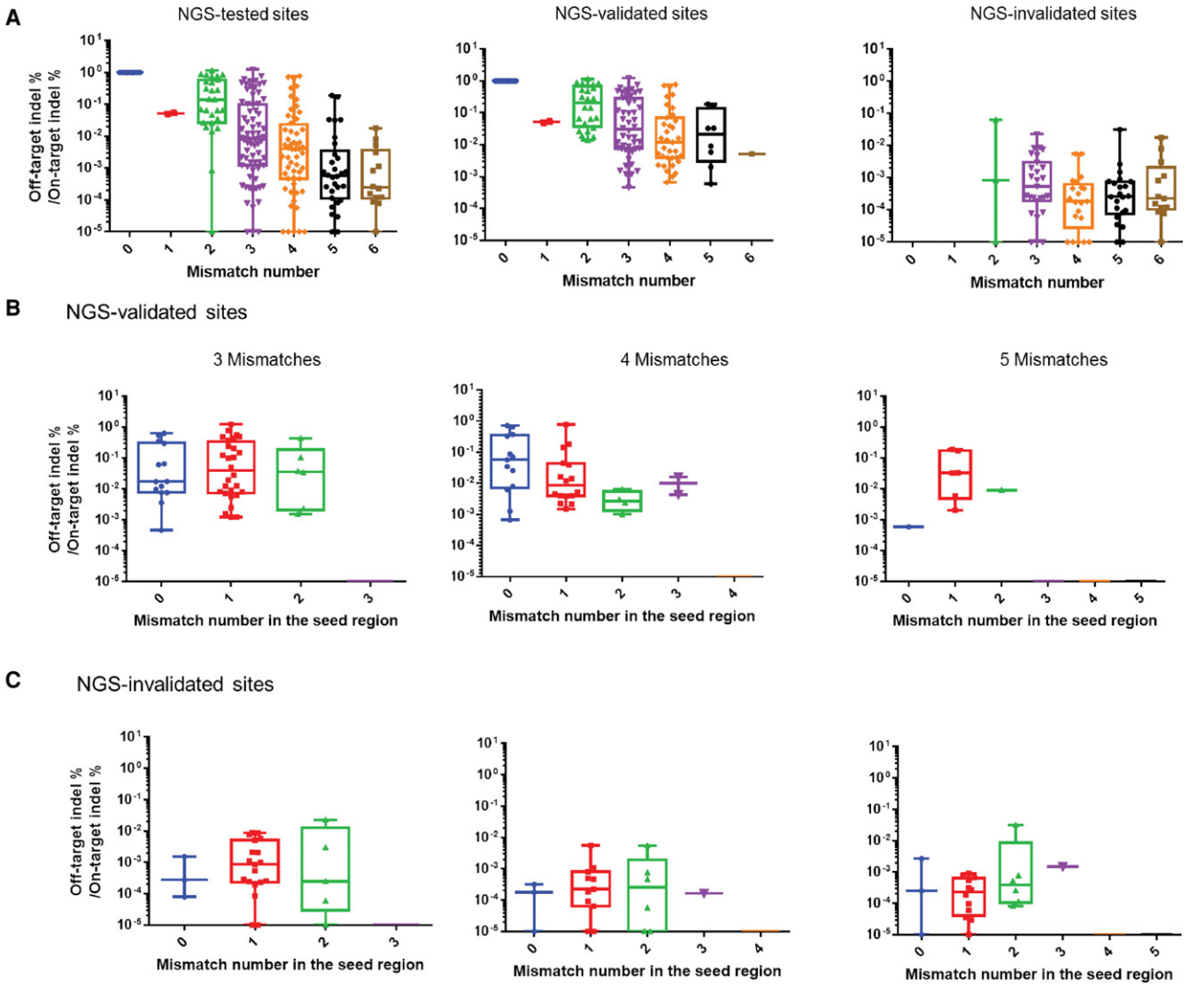


Figure 5. Analysis of NGS-validated and NGS-invalidated off-target sites. Plots of relative indel frequencies (log scale) at off-target sites harboring the number of mismatches indicated in the entire 20-nt sequence (*A*) or the 10-nt seed sequence (*B*,*C*). NGS-tested sites (*A*) were divided into two groups: validated sites (*B*) and invalidated sites (*C*). NGS-validated sites and NGS-invalidated sites were those with indel frequencies above and below, respectively, noise indel levels.

reduce off-target indel frequencies below detection limits even with a Cas9 monomer combined with a modified ggX_{20} sgRNA. Truncated sgRNAs that target unique 17-nt sites that differ by at least 2 or 3 nt from any other site in the genome can also be highly specific, but such sites are much rarer than 20-nt full-length sites.

Digenome-seq differs from other cell-based methods in that it detects DNA cleavages in vitro, rather than in cells, using cell-free genomic DNA. As a result, Digenome-seq is not limited by chromatin, unlike other methods. Note that DSBs cannot be processed or trimmed in vitro by endogenous DSB repair systems. This feature favors pinpointing of in vitro cleavage sites and multiplexing with Digenome-seq, because no computational search for sequences homologous to on-target sites is required. Furthermore, because no PCR amplification steps are involved with Digenome-seq prior to WGS, it is much simpler and easier to carry out than other methods.

Pattanayak et al. (2013) examined in vitro specificities of Cas9-sgRNA using partially randomized DNA substrate libraries containing approximately 10^{12} target sequence variants, rather than human genomic DNA, and reported that five out of 49 sites cleaved in vitro were validated in cells via targeted deep sequencing. This in vitro selection approach is limited by two factors. First, randomized DNA substrates cannot faithfully represent genomic DNA. A vast majority of cleaved sequences do not exist in the genome. Second, potential off-target sites are determined by counting PCR amplicons before and after in vitro cleavage and selection. The PCR step can introduce biases, giving rise to false positives and negatives.

Although multiplex Digenome-seq was more comprehensive and sensitive than GUIDE-seq and HTGTS, in general, it often missed off-target sites captured by these methods. Monoplex Digenome-seq using a high concentration of sgRNA may capture some of these missing sites, although it is cost-inefficient to analyze many sgRNAs individually. We recommend using at least two different methods, one in vitro method, namely, Digenome-seq, and the other cell-based method to profile genome-wide specificities of RGENs comprehensively.

Methods

Cas9 and in vitro sgRNA

Recombinant Cas9 protein was purchased from ToolGen. sgRNAs were synthesized by in vitro transcription using T7 RNA polymerase as described previously (Kim et al. 2014). Briefly, sgRNA templates were generated by annealing two complementary oligonucleotides purchased from Macrogen. These oligonucleotides were reverse-phase-purified using the vendor's MOPC purification method and quality-checked using MALDI-TOF. sgRNA templates were incubated with T7 RNA polymerase in reaction buffer (40 mM Tris-HCl, 6 mM $MgCl_2$, 10 mM DTT, 10 mM NaCl, 2 mM spermidine, NTP, RNase inhibitor, at pH 7.9) for 8 h at 37°C. Transcribed sgRNAs were preincubated with DNase I to remove template DNA, and purified using PCR purification kits (Macrogen).

Cell culture and transfection conditions

HeLa cells were cultured in DMEM media supplemented with 10% FBS. HeLa cells (8×10^4) were cotransfected with the Cas9 expression plasmid (500 ng) and the sgRNA-encoding plasmid (500 ng) using Lipofectamine 2000 (Life Technologies). HEK293T cells were maintained in DMEM media supplemented with 10% FBS. HEK293T cells (1×10^6) were electroporated with Cas9 expression plasmid (5 µg) and the sgRNA-encoding plasmid (5 µg) using Amaxa 4D Nucleofector. Genomic DNA was isolated with the DNeasy tissue kit (Qiagen) according to the manufacturer's instructions after 48 h.

In vitro cleavage of genomic DNA

Genomic DNA was purified from HAP1 cells (Carette et al. 2011) with the DNeasy tissue kit (Qiagen). In vitro cleavage of genomic DNA for Digenome-seq was carried out as described previously (Kim et al. 2015). Briefly, Cas9 protein (40 µg) and 11 sgRNAs (2.7 µg each) were preincubated at room temperature for 10 min to form RNP complexes. Genomic DNA (8 µg) was incubated with RNP complexes in a reaction buffer (100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, 100 µg/mL BSA, at pH 7.9) for 8 h at 37°C. Digested genomic DNA was treated with RNase A (50 µg/mL) to degrade sgRNAs and purified again with the DNeasy tissue kit (Qiagen).

Whole-genome and Digenome sequencing

For whole-genome sequencing, Cas9-digested genomic DNA was fragmented to 400–500 bp using Covaris (Thermo Fischer). Fragmented genomic DNA (1 µg) was ligated with adaptors using TruSeq DNA library prep kit. Libraries were subjected to WGS using an Illumina HiSeq X Ten sequencer at Macrogen. WGS is performed at a sequencing depth of 30× to 40×. We used Isaac aligner to align the sequence file to the human reference genome hg19 (Raczy et al. 2013) or GRCh38 (Cunningham et al. 2015) with the following mapping program and parameters: base quality cutoff, 15; keep duplicate reads, yes; variable read length support, yes; realign gaps, no; and adaptor clipping, yes (adaptor AGATCGGAAGAGC*,*GCTCTTCCGATCT). DNA cleavage sites were identified computationally using a cleavage scoring system described in Supplemental Figure 1. The resulting multiplex Digenome-captured sites were classified into 11 groups by edit distance (Zorita et al. 2015). The computer programs used for identification of in vitro RGEN cleavage sites and classification of these Digenome-captured sites by edit distance are available at our website (www.rgenome.net/digenome). Note that no genomic DNA isolated from CRISPR-Cas9–transfected cells was used for Digenome-seq in this study. It is unnecessary to carry out WGS using Cas9-undigested genomic DNA in future studies because there were essentially no false-positive sites with cleavage scores above the cutoff value of 2.5 as shown in Figure 1A.

Targeted deep sequencing

On-target and potential off-target candidate sites were amplified using Phusion polymerase (New England Biolabs). PCR amplicons were denatured by NaOH and subjected to paired-end sequencing using Illumina MiSeq. Indel frequencies were calculated as described previously (Kim et al. 2014).

Data access

The deep sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP067307.

Acknowledgments

This work was supported by a grant from IBS (IBS-R021-D1). We thank Dr. Sangsu Bae for the creation of our website (www.rgenome.net/digenome) that supports Digenome-seq analysis.

Author contributions: D.K. and Sojung K. carried out the experiments. D.K., Sunghyn K., and J.P. performed bioinformatics analyses. J.-S.K. supervised the research.

References

Bae S, Park J, Kim JS. 2014. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30:** 1473–1475.

Carette JE, Raaben M, Wong AC, Herbert AS, Obernosterer G, Mulherkar N, Kuehne AI, Kranzusch PJ, Griffin AM, Ruthel G, et al. 2011. Ebola virus entry requires the cholesterol transporter Niemann–Pick C1. *Nature* **477:** 340–343.

Chiarle R, Zhang Y, Frock RL, Lewis SM, Molinie B, Ho YJ, Myers DR, Choi VW, Compagno M, Malkin DJ, et al. 2011. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147:** 107–119.

Cho SW, Kim S, Kim JM, Kim JS. 2013a. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat Biotechnol* **31:** 230–232.

Cho SW, Lee J, Carroll D, Kim JS, Lee J. 2013b. Heritable gene knockout in *Caenorhabditis elegans* by direct injection of Cas9–sgRNA ribonucleoproteins. *Genetics* **195:** 1177–1180.

Cho SW, Kim S, Kim Y, Kweon J, Kim HS, Bae S, Kim JS. 2014. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res* **24:** 132–141.

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339:** 819–823.

Cradick TJ, Fine EJ, Antico CJ, Bao G. 2013. CRISPR/Cas9 systems targeting β-globin and *CCR5* genes have substantial off-target activity. *Nucleic Acids Res* **41:** 9584–9592.

Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, Karaca E, Chiarle R, Skrzypczak M, Ginalski K, et al. 2013. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods* **10:** 361–365.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2015. Ensembl 2015. *Nucleic Acids Res* **43(**Database issue**):** D662–D669.

Frock RL, Hu J, Meyers RM, Ho YJ, Kii E, Alt FW. 2015. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* **33:** 179–186.

Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD. 2013. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* **31:** 822–826.

Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. 2014. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32:** 279–284.

Gabriel R, von Kalle C, Schmidt M. 2015. Mapping the precision of genome editing. *Nat Biotechnol* **33:** 150–152.
Guilinger JP, Thompson DB, Liu DR. 2014. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat Biotechnol* **32:** 577–582.
Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31:** 827–832.
Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol* **31:** 227–229.
Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol* **31:** 233–239.
Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337:** 816–821.
Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. 2013. RNA-programmed genome editing in human cells. *eLife* **2:** e00471.
Kim H, Kim JS. 2014. A guide to genome engineering with programmable nucleases. *Nat Rev Genet* **15:** 321–334.
Kim S, Kim D, Cho SW, Kim J, Kim JS. 2014. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res* **24:** 1012–1019.
Kim D, Bae S, Park J, Kim E, Kim S, Yu HR, Hwang J, Kim JI, Kim JS. 2015. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods* **12:** 237–243.
Kuscu C, Arslan S, Singh R, Thorpe J, Adli M. 2014. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol* **32:** 677–683.
Lin Y, Cradick TJ, Brown MT, Deshmukh H, Ranjan P, Sarode N, Wile BM, Vertino PM, Stewart FJ, Bao G. 2014. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res* **42:** 7473–7485.
Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S, Yang L, Church GM. 2013a. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol* **31:** 833–838.
Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013b. RNA-guided human genome engineering via Cas9. *Science* **339:** 823–826.
Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. 2013. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol* **31:** 839–843.
Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, Chuang HY, Kallberg M, Kumar SA, Liao A, et al. 2013. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29:** 2041–2043.
Ramakrishna S, Kwaku Dad AB, Beloor J, Gopalappa R, Lee SK, Kim H. 2014. Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome Res* **24:** 1020–1027.
Ran FA, Hsu PD, Lin CY, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y, et al. 2013. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154:** 1380–1389.
Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, Zetsche B, Shalem O, Wu X, Makarova KS, et al. 2015. *In vivo* genome editing using *Staphylococcus aureus* Cas9. *Nature* **520:** 186–191.
Tsai SQ, Wyvekens N, Khayter C, Foden JA, Thapar V, Reyon D, Goodwin MJ, Aryee MJ, Joung JK. 2014. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat Biotechnol* **32:** 569–576.
Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, Wyvekens N, Khayter C, Iafrate AJ, Le LP, et al. 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33:** 187–197.
Wang X, Wang Y, Wu X, Wang J, Wang Y, Qiu Z, Chang T, Huang H, Lin RJ, Yee JK. 2015. Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat Biotechnol* **33:** 175–178.
Zorita E, Cusco P, Filion GJ. 2015. Starcode: sequence clustering based on all-pairs search. *Bioinformatics* **31:** 1913–1919.
Zuris JA, Thompson DB, Shu Y, Guilinger JP, Bessen JL, Hu JH, Maeder ML, Joung JK, Chen ZY, Liu DR. 2015. Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing *in vitro* and *in vivo*. *Nat Biotechnol* **33:** 73–80.

Received September 15, 2015; accepted in revised form January 6, 2016.

Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq

Daesik Kim, Sojung Kim, Sunghyun Kim, et al.

Genome Res. 2016 26: 406-415 originally published online January 19, 2016
Access the most recent version at doi:10.1101/gr.199588.115