Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells

Daesik Kim[1,2,8], Sangsu Bae[1,2,7,8], Jeongbin Park[2], Eunji Kim[3], Seokjoong Kim[3], Hye Ryeong Yu[3], Jinha Hwang[4], Jong-Il Kim[4–6] & Jin-Soo Kim[1,2]

Although RNA-guided genome editing via the CRISPR-Cas9 system is now widely used in biomedical research, genome-wide target specificities of Cas9 nucleases remain controversial. Here we present Digenome-seq, *in vitro* Cas9-digested whole-genome sequencing, to profile genome-wide Cas9 off-target effects in human cells. This *in vitro* digest yields sequence reads with the same 5′ ends at cleavage sites that can be computationally identified. We validated off-target sites at which insertions or deletions were induced with frequencies below 0.1%, near the detection limit of targeted deep sequencing. We also showed that Cas9 nucleases can be highly specific, inducing off-target mutations at merely several, rather than thousands of, sites in the entire genome and that Cas9 off-target effects can be avoided by replacing 'promiscuous' single guide RNAs (sgRNAs) with modified sgRNAs. Digenome-seq is a robust, sensitive, unbiased and cost-effective method for profiling genome-wide off-target effects of programmable nucleases including Cas9.

© 2015 Nature America, Inc. All rights reserved.

Programmable nucleases, which include zinc-finger nucleases[1–5], transcription activator–like effector nucleases[6–8] and RNA-guided engineered nucleases (RGENs)[9–15] derived from the type II clustered, regularly interspaced, short palindromic repeats (CRISPR)–CRISPR-associated (Cas) prokaryotic adaptive immune system, are now widely used for genome editing in cultured cells and whole organisms[16]. Unfortunately, these nucleases can induce off-target mutations. For example, RGENs—which consist of the Cas9 protein derived from *Streptococcus pyogenes* and sgRNAs, and which recognize 22-bp target DNA sequences composed of a 20-bp sequence that hybridizes with the sgRNA and a 5′-NGG-3′ protospacer-adjacent motif (PAM) sequence recognized by Cas9—can tolerate mismatches at up to several nucleotide positions[17–22]. This suggests that there are up to thousands of off-target sites for a given nuclease in the human genome. Off-target DNA cleavages can lead to mutations at unintended genomic loci and to chromosomal rearrangements such as translocations[20,23], deletions[7,24] and inversions[25]; such effects raise concerns about the application of this technology in research and medicine.

Various strategies have been reported to reduce RGEN off-target effects, including the use of sgRNAs with two extra guanine nucleotides at the 5′ end[20], truncated sgRNAs[26], paired Cas9 nickases[20,27,28] and a catalytically dead Cas9 (dCas9)-FokI fusion[29,30], as well as delivery of purified Cas9 protein[14,31,32]. Although these different approaches have been shown to reduce the frequency of off-target mutations by at least an order of magnitude at several sites, it remains unknown whether these RGEN variants are free of off-target effects in the entire genome. To address this critical issue, researchers must develop methods to interrogate the specificities of RGENs on a genomic scale in an unbiased manner.

In this study, we used whole-genome sequencing (WGS) to profile genome-wide RGEN off-target mutations in clonal gene knockout (KO) cells created using RGENs. Furthermore, we identified RGEN off-target sites in a bulk population of cells by sequencing nuclease-digested genomic DNA. Our results showed that RGENs can be highly specific, inducing off-target mutations at only a handful, rather than thousands, of sites in the entire genome and that these off-target effects can be avoided by using modified sgRNAs.

RESULTS

Whole-genome sequencing of human haploid cells

We chose a human haploid cell line, HAP1 (ref. 33), to isolate clonal populations of gene KO cells that carried RGEN-induced mutations. In haploid cells, the same number of reads translates to a sequencing depth double that of diploid cells, and a filter can remove heterozygous variants. We isolated five different KO HAP1 cell lines, each with a single disruption in a kinase gene (*ABL1*, *EPHB2*, *ERBB3*, *FGFR2* and *FGFR4*; **Supplementary Fig. 1**). Genomic DNA isolated from these cells and wild-type cells was subjected to WGS. To test the reproducibility of WGS, we sequenced wild-type cells and the *ABL1*$^{-}$ cells in duplicate.

[1]Center for Genome Engineering, Institute for Basic Science, Seoul, South Korea. [2]Department of Chemistry, Seoul National University, Seoul, South Korea. [3]ToolGen, Inc., Byucksan Kyoungin Digital Valley 2-Cha, Seoul, South Korea. [4]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea. [5]Department of Biochemistry, Seoul National University College of Medicine, Seoul, South Korea. [6]Genomic Medicine Institute (GMI), Medical Research Center, Seoul National University, Seoul, South Korea. [7]Present address: Department of Chemistry, Hanyang University, Seoul, South Korea. [8]These authors contributed equally to this work. Correspondence should be addressed to J.-S.K. (jskim01@snu.ac.kr).

RECEIVED 2 DECEMBER 2014; ACCEPTED 31 DECEMBER 2014; PUBLISHED ONLINE 9 FEBRUARY 2015; DOI:10.1038/NMETH.3284

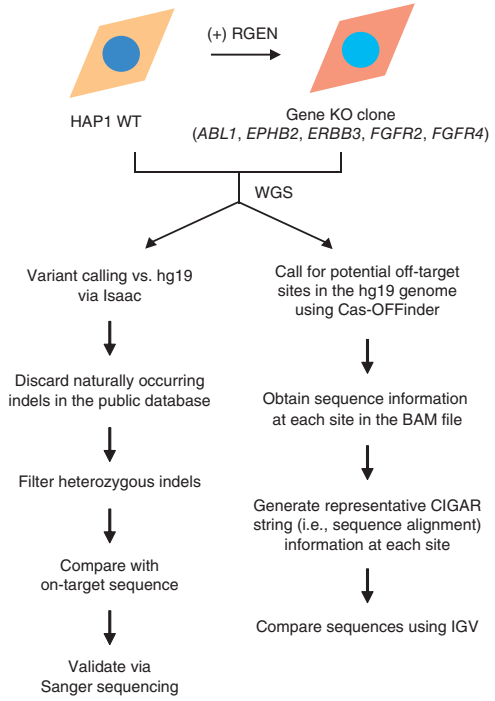


Figure 1 | Workflow of off-target analysis of gene KO clones via WGS. WT, wild type.

We used Isaac, a variant-calling program, to identify small insertions or deletions (indels) rather than point mutations, relative to the hg19 reference genome (**Fig. 1**), because programmable nucleases rarely produce substitutions[34]. After applying bioinformatics filters, we obtained 2,026–3,250 indels that were unique in each genome (**Supplementary Table 1** and **Supplementary Note 1**). We then compared RGEN target sites with these indel locations. Only 9–84 indel sites contained a PAM sequence and had at least 10-nt matches with respective on-target sequences (**Supplementary Table 1**). Only one of these indels was validated by Sanger sequencing (**Supplementary Fig. 2a**), and it was caused by spontaneous mutation (**Supplementary Fig. 2b**). We confirmed all five on-target mutations with either Isaac or the Integrative Genomics Viewer (IGV)[35] (**Supplementary Fig. 2c**,**d**).

Examining potential off-target sites

Next, we used Cas-OFFinder[36] (http://www.rgenome.net) and its upgraded version to list >100,000 homologous sites that differed from on-target sites by up to 8 nt or that differed by 2 nt with a DNA or RNA bulge (**Supplementary Fig. 3a**), and we examined whether there were nuclease-induced indels at these potential off-target sites in each genome sequence. We developed a computer program to compare sequence reads aligned around each of the >100,000 homologous sites with the reference sequence (**Supplementary Fig. 3b**). Notably, this program successfully identified all five on-target mutations in the five KO clones (**Supplementary Fig. 3c**). Still, no off-target indels were found.

These results suggest that RGENs are highly specific, leaving no off-target mutations in single cell–derived clones. However, off-target mutations that occur at frequencies below 10% in a bulk population of cells cannot be detected by analyzing several clones, which could pose a problem in gene and cell therapy applications. Furthermore, it is still unknown whether RGENs can induce low-frequency indels at hundreds or even thousands of off-target sites.

Digenome-seq: straight versus staggered alignment

We reasoned that we could identify off-target mutations induced by RGENs in a bulk population of cells by sequencing *in vitro* nuclease-digested genomes (digenomes). These digests should produce many DNA fragments with identical 5′ ends, which give rise to sequence reads that are vertically aligned at cleavage sites. In contrast, all other sequence reads would be aligned in a staggered manner.

We chose an *HBB* gene–specific RGEN that has been shown to induce off-target mutations at a highly homologous site (referred to as the OT1 site)[19]. We also analyzed three other homologous sites (OT3, OT7 and OT12) that differed from the on-target site by 3 nt. This RGEN cleaved the on-target, OT1 and OT3 sites efficiently in a genomic context *in vitro* and *in vivo* (**Fig. 2** and **Supplementary Note 2**).

Four different sets of genomic DNA were subjected to WGS to investigate whether *in vitro* digestion of genomic DNA using RGENs can produce sequence reads with straight alignment at cleavage sites. Genomic DNA isolated from RGEN- and

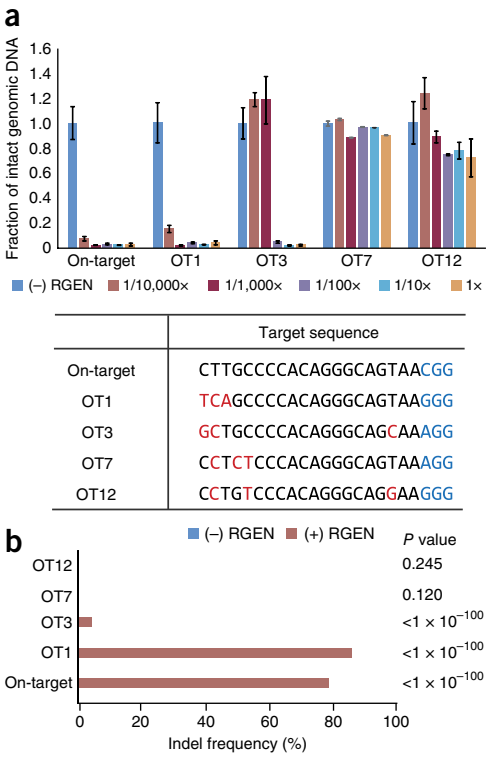


Figure 2 | RGEN-mediated genomic DNA digestion *in vitro*. (**a**) Top, fraction of intact genomic DNA not cleaved by the *HBB*-targeting RGEN at on-target and four potential off-target sites. 1×–1/10,000× denote serial dilutions of sgRNA and Cas9. The fraction of uncleaved DNA was measured by real-time quantitative PCR. Error bars, s.e.m. from at least three independent experiments. Bottom, DNA sequences with mismatched nucleotides (red) and PAM sequence (blue) of the on-target and the four potential off-target sites. (**b**) Measuring RGEN-driven mutation frequencies with targeted deep sequencing. *P* values were calculated by the Fisher exact test.

© 2015 Nature America, Inc. All rights reserved.

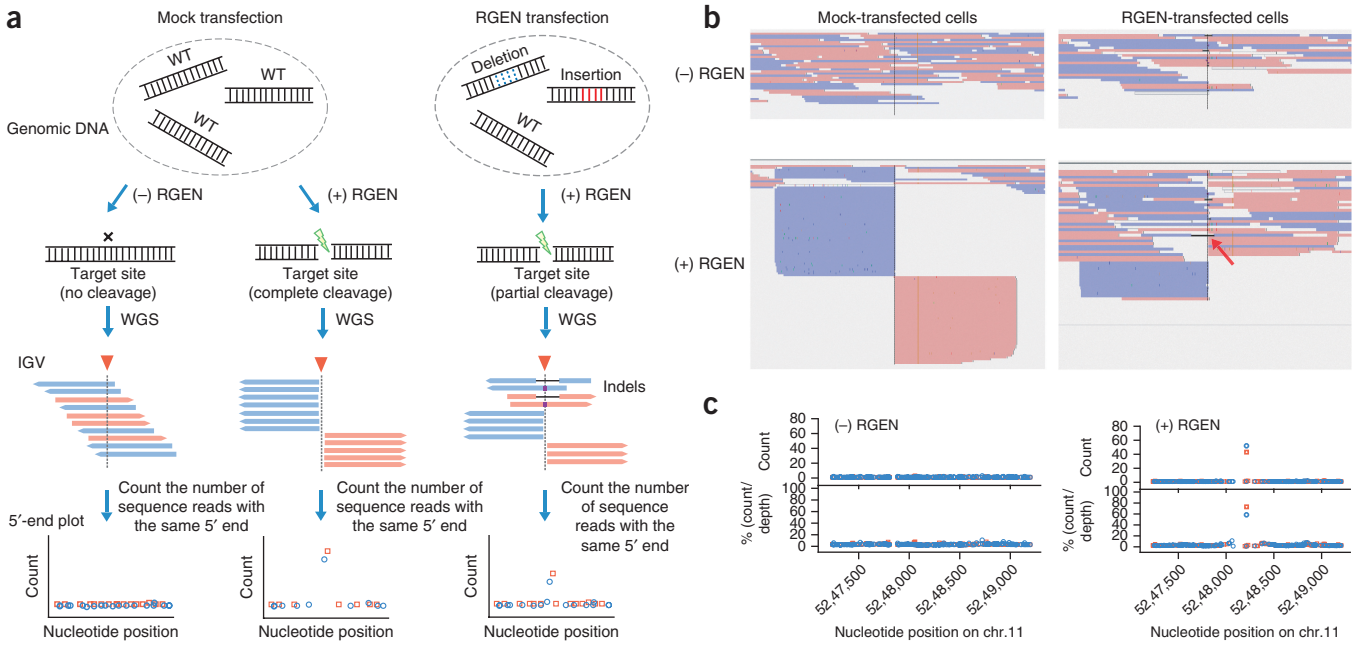


Figure 3 | RGEN-induced 'digenome' sequencing to capture off-target sites. (**a**) Overview of Digenome-seq. Forward and reverse sequence reads are shown in pink and blue, respectively. Red triangles and vertical lines indicate cleavage positions. WT, wild type. (**b**) Representative IGV images obtained using the *HBB*-specific RGEN at the on-target site. An indel is indicated by the arrow. (**c**) 5′-end plots showing the absolute and relative number of sequence reads with the same 5′ end across nucleotide positions.

mock-transfected HAP1 cells was digested *in vitro* with the *HBB* RGEN at 300 nM to ensure complete cleavage. In parallel, intact genomic DNA isolated from transfected and mock-transfected cells was subjected to WGS without RGEN digestion *in vitro* (**Fig. 3a**). After mapping sequence reads to the reference genome, we used IGV to observe patterns of sequence alignments at the on-target and the four homologous sites.

First, we examined the digenome isolated from mock-transfected cells. At the on-target, OT1 and OT3 sites, unusual patterns of straight alignments were observed (**Fig. 3b** and **Supplementary Fig. 4a**,**b**). Sequence reads that spanned the cleavage sites were very rare. Without RGEN treatment *in vitro*, no such straight alignments were observed. At the OT7 and OT12 sites, most sequence reads spanned the potential cleavage site (3 bp upstream of the PAM), resulting in a staggered alignment (**Supplementary Fig. 4c**,**d**).

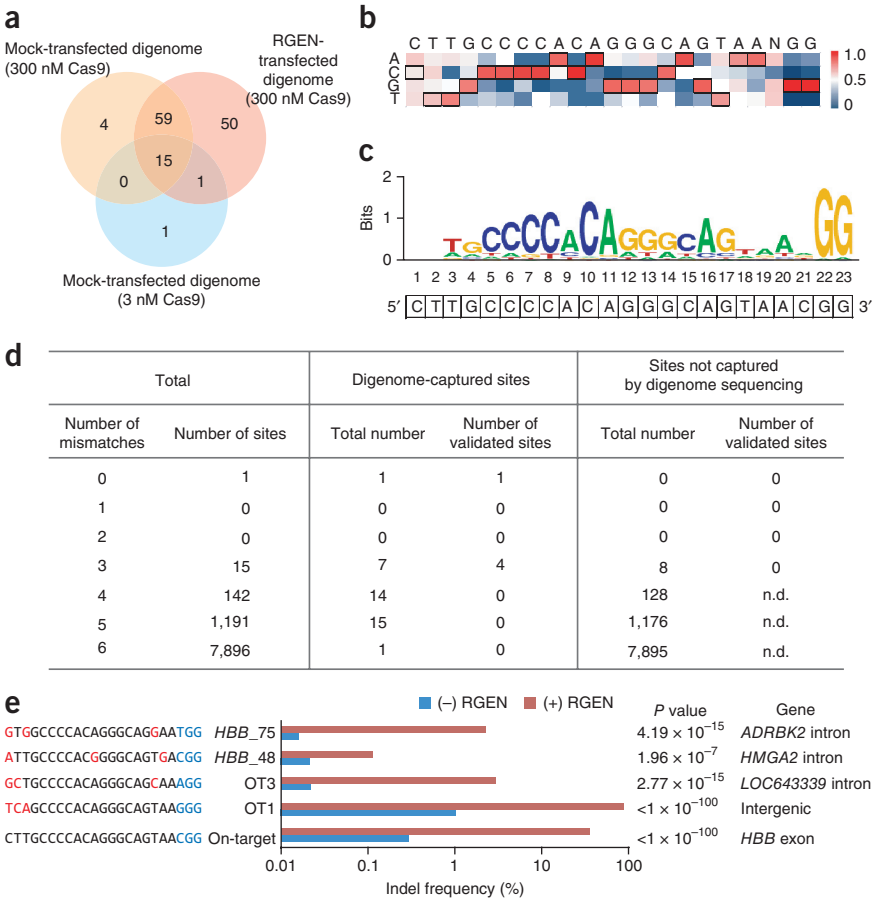
Second, we compared the digenome isolated from RGEN-transfected cells with the respective intact genome. At all five sites, the intact genome gave rise to typical patterns of staggered alignments (**Fig. 3b** and **Supplementary Fig. 4**). In contrast, the digenome showed both straight and staggered alignments at the on-target and OT1 sites. At these two sites, almost all overlapping sequence reads responsible for staggered alignments contained indels. Apparently, these indel sequences were not cleaved by the RGEN *in vitro*[37]. No indels were found with overlapping sequence reads that spanned the OT7 and OT12 sites. The OT3 site was an interesting case: the digenome showed a straight pattern with a few overlapping sequence reads. Notably, one overlapping sequence contained an indel, most likely induced by the RGEN (**Supplementary Fig. 4b**). These results suggest that Digenome-seq is sensitive enough to allow identification of rare off-target mutations and that a straight or vertical alignment of sequence reads is a unique signature of RGEN cleavage *in vitro*, although not all sites with straight alignments are genuine off-target sites.

Off-target sites at genomic scale

We developed a computer program that searches for straight alignments of sequence reads across the genome. First, we plotted the count of sequence reads whose 5′ ends started at the nucleotide position near the *HBB* on-target and two validated off-target sites at single-nucleotide resolution (**Supplementary Fig. 5a**). We assumed that an almost equal number of sequence reads, corresponding to either the Watson or Crick strand, should be observed right next to each other at a cleavage site, producing double peaks. As expected, the digenome gave rise to double peaks at the three cleavage sites (**Fig. 3c** and **Supplementary Fig. 5b**,**c**). The intact genome that had not undergone RGEN treatment *in vitro* did not produce such double-peak patterns at these sites.

Next, we applied this approach to the entire RGEN-transfected digenome, mock-transfected digenome, intact RGEN-transfected genome and intact mock-transfected genome. In addition, mock-transfected genomic DNA was treated with Cas9 protein *in vitro* in the absence of sgRNA or with a 100-fold–lower concentration of RGEN (3 nM Cas9) and then subjected to WGS and digenome analysis. We computationally searched for sites where the count of sequence reads with the same 5′ end was greater than 10 in both strands and where at least 20% of sequence reads were aligned vertically. A total of 17 and 78 sites, including the on-target and two validated off-target sites, were identified in the mock-transfected digenome treated with 3 nM and 300 nM RGEN, respectively (**Fig. 4a**), which showed double-peak patterns in a 5′-end plot and straight alignments in an IGV image. In the RGEN-transfected digenome, we observed such patterns at 125 sites, including the three validated on- and off-target sites. The OT7 and OT12 sites did not show double-peak patterns in

© 2015 Nature America, Inc. All rights reserved.



Total		Digenome-captured sites		Sites not captured by digenome sequencing	
Number of mismatches	Number of sites	Total number	Number of validated sites	Total number	Number of validated sites
0	1	1	1	0	0
1	0	0	0	0	0
2	0	0	0	0	0
3	15	7	4	8	0
4	142	14	0	128	n.d.
5	1,191	15	0	1,176	n.d.
6	7,896	1	0	7,895	n.d.

P value	Gene
4.19×10^{-15}	*ADRBK2* intron
1.96×10^{-7}	*HMGA2* intron
2.77×10^{-15}	*LOC643339* intron
$<1 \times 10^{-100}$	Intergenic
$<1 \times 10^{-100}$	*HBB* exon

Figure 4 | Off-target sites of the *HBB* RGEN captured by Digenome-seq and validated by targeted deep sequencing. (**a**) Venn diagram showing the number of digenome-captured sites using the *HBB* RGEN in mock-transfected or RGEN-transfected cells. (**b**) Heat map comparing digenome-captured sites with the on-target site. Dark red and dark blue correspond to 100% and 0% matches, respectively, at a given position. (**c**) Sequence logo obtained via WebLogo using DNA sequences at digenome-captured sites. (**d**) Summary of Digenome-seq and targeted deep sequencing. n.d., not determined. (**e**) Off-target sites validated by targeted deep sequencing. Blue and red bars represent indel frequencies obtained using mock-transfected HAP1 cells and the *HBB* RGEN–transfected HAP1 cells, respectively. Left, DNA sequences of on-target and off-target sites. Mismatched bases are shown in red. The PAM is shown in blue. Right, *P* values calculated by the Fisher exact test. Additional deep sequencing results can be found in **Supplementary Table 3**.

these three digenomes. Most sites were commonly identified in the three digenomes, demonstrating the high reproducibility of Digenome-seq. Thus, 15 (94%) of the 16 candidate sites (excluding the one false positive site) found in the mock-transfected digenome (3 nM RGEN) were also identified in the other two independent digenomes (**Fig. 4a**). But none of the 122 sites (excluding the 3 validated sites) was accompanied by indels in the RGEN-transfected digenome, suggesting that mutations at these candidate sites occurred rarely, if at all. Such double-peak patterns were observed at only a few positions in the three intact genomes (**Supplementary Table 2**). All of these positions identified in the intact genomes were false positives that resulted from naturally occurring indels in the HAP1 genome relative to the reference genome (**Supplementary Fig. 6**). Thus, double-peak patterns or straight alignments of sequence reads were unique features found in the three digenomes.

We compared DNA sequences at the 74 common sites identified in the RGEN-transfected and mock-transfected digenomes with the 20-bp on-target site and found that, of the 20 nt, all but the one at the 5′ end were conserved (**Fig. 4b**). Interestingly, the 5′ base in the sgRNA was a mismatched guanine, required for transcription under the control of the *U6* promoter. The 5′ base in the target sequence was a cytosine. Furthermore, the *de novo* motif computationally obtained by comparing the DNA sequences at the 74 sites with one another, rather than with the on-target sequence, clearly showed matches with the on-target sequence at all positions other than the first 2 nt (**Fig. 4c**). Notably, 70 (95%) of these double-peak positions were accompanied by the 5′-NGG-3′ PAM, exactly 3 nt downstream from the expected cleavage position. Some sites were matched to the on-target site by allowing a DNA/RNA bulge or assuming 5′-NGA-3′ or 5′-NAG-3′ as a noncanonical PAM. The other sites had no apparent sequence homology with the on-target sequence, suggesting that they were false positives.

We also found that the fewer mismatches there were in homologous sites, the more likely they were to be captured by Digenome-seq. Thus, 21 out of 157 (13%) homologous sites that differed from the on-target site by 3 or 4 nt were captured, but only 15 out of 1,191 sites (1.2%) and 1 out of 7,896 sites (0.013%) that differed by 5 and 6 nt, respectively, were captured (**Fig. 4d**). Taken together, these results indicate that most of the double-peak patterns are caused by RGEN digestion *in vitro* and that Digenome-seq can capture nuclease cleavage sites in a genomic context.

Validating off-target effects at candidate sites

We performed targeted deep sequencing to validate or invalidate off-target effects at the 74 common sites identified in the two independent digenomes (**Fig. 4e** and **Supplementary Table 3**). In addition, we tested the other eight sites that differed from the on-target site by 3 nt but were not captured by Digenome-seq. No off-target indels were detected at these eight sites with a frequency of at least 0.1% and greater than that of the negative control (Fisher exact test, $P < 0.01$) (**Fig. 4d** and **Supplementary Table 3**). Indels were observed at five sites—including already-validated on-target, OT1 and OT3 sites—among the 74 sites, with frequencies ranging from 0.11% to 87% (**Fig. 4e** and **Supplementary Fig. 7**). At the other two newly validated off-target sites, termed *HBB*_48 and *HBB*_75, indels were detected with a frequency of 0.11% and 2.2%, respectively. These two sites differed from the on-target site by 3 nt. There were three nucleotide mismatches at the *HBB*_48 site and two mismatches at the *HBB*_75 site, relative to the 20-nt sgRNA sequence, which differed from the on-target site by 1 nt at the 5′ terminus. None of these validated off-target sites harbored

© 2015 Nature America, Inc. All rights reserved.

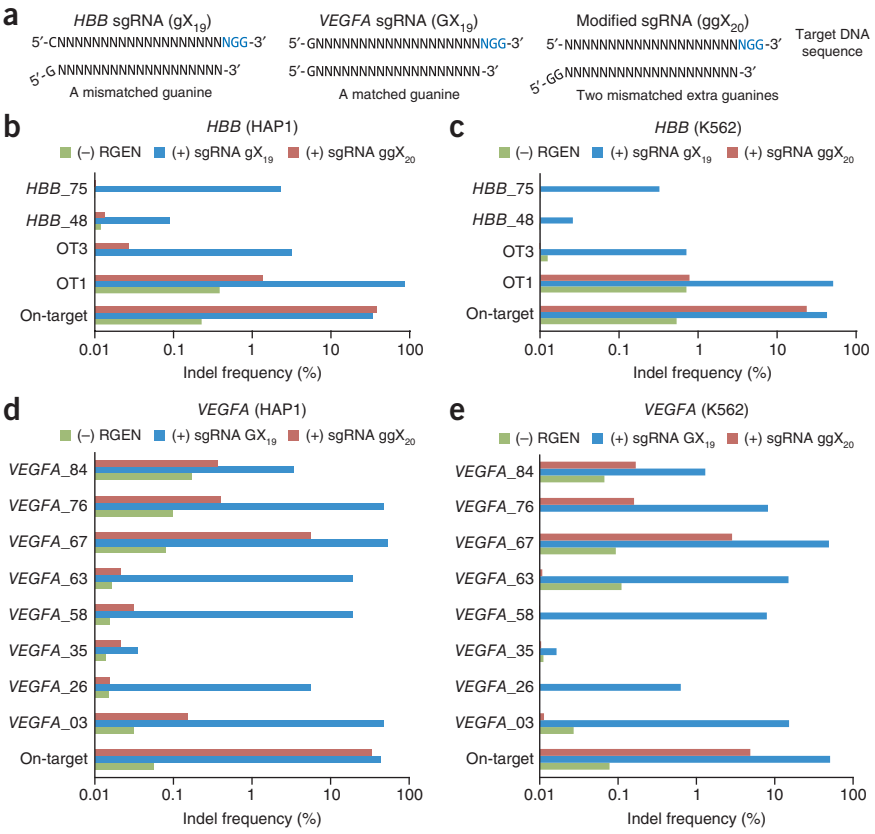


Figure 5 | Comparison of conventional sgRNAs with modified sgRNAs that include two extra guanine nucleotides. (**a**) Schematic of gX_{19} sgRNA, GX_{19} sgRNA and ggX_{20} sgRNA; g and G (preceding X) represent a mismatched guanine and matched guanine, respectively. (**b**,**c**) Indel frequencies at the *HBB* on-target and the four validated off-target sites in HAP1 cells (**b**) and K562 cells (**c**) measured via targeted deep sequencing. (**d**,**e**) Indel frequencies at the *VEGFA* on-target and the eight validated off-target sites in HAP1 cells (**d**) and K562 cells (**e**) measured via targeted deep sequencing. Numeric high-throughput sequencing data are shown in **Supplementary Table 6**.

a DNA/RNA bulge compared to the 20-nt sgRNA sequence, nor were they accompanied by a noncanonical PAM (5′-NGA-3′ or 5′-NAG-3′). Note that these two new off-target sites and the other three sites were captured independently in each of the three digenomes, illustrating the high sensitivity and reproducibility of Digenome-seq.

Digenome-seq with another 'promiscuous' RGEN

We carried out Digenome-seq with another RGEN that had been shown to induce on-target mutations at a *VEGFA* locus and off-target mutations at four homologous sites[18]. A total of 81 sites, including the on-target and four validated off-target sites, were captured that showed double-peak patterns (**Supplementary Figs. 8** and **9**). All of the DNA sequences at these 81 sites contained the canonical 5′-NGG-3′ PAM. Comparison of these sequences with the on-target sequence showed matches at every nucleotide position. We also compared these sequences with one another to obtain a *de novo* motif: the resulting sequence logo also showed matches with the target sequence at almost every nucleotide position, suggesting that each nucleotide in the 20-nt sgRNA sequence contributed to the specificity (**Supplementary Fig. 8b**,**c**).

We then used targeted deep sequencing to confirm on-target and off-target effects at the 81 sites captured by digenome analysis and 28 sites that differed from the on-target site by 3 or fewer nucleotides but were not captured. This RGEN produced indels at the target site and the four previously validated off-target sites with frequencies that ranged from 0.32% to 87%. In addition, four new off-target sites that had been captured by Digenome-seq were validated at which indels were induced with frequencies that ranged from 0.065% ± 0.021% to 6.4% ± 1.2% (± s.e.m.) (**Supplementary Figs. 8e** and **10**). These off-target sites contained 1–6 nucleotide mismatches with the 20-nt target sequence and at least one mismatch in the PAM-proximal seed region. There are 13,892 sites with 6-nt mismatches in the human genome, but only six sites (0.043%) were captured by Digenome-seq; among them, only one site was validated by deep sequencing (**Supplementary Fig. 8d**,**e**). None of these validated off-target sites contained a DNA/RNA bulge, although 40 out of 81 sites captured by Digenome-seq contained a missing or extra nucleotide compared to the 20-nt target sequence. At all the other sites, including those not captured by Digenome-seq, indel frequencies were below 0.05% or were smaller than or not statistically different from those obtained using an empty vector control.

Avoiding RGEN off-target effects via modified sgRNAs

sgRNAs with two extra guanines at the 5′ terminus (termed ggX_{20} sgRNAs) can efficiently discriminate on-target sites from homologous sites that differ by ≥2 nt, thus reducing off-target effects by orders of magnitude without sacrificing on-target effects[20]. We replaced the two promiscuous gX_{19} (*HBB*) and GX_{19} (*VEGFA*) sgRNAs with respective ggX_{20} sgRNAs ('g' and 'G' represent a mismatched guanine and matched guanine, respectively) (**Fig. 5a**) and measured on-target and off-target mutation frequencies in HAP1 and K562 (human leukemia) cells at the validated sites. Strikingly, indels were barely detectable above deep-sequencing error rates at the several off-target sites in the two genes when we used the ggX_{20} sgRNAs (**Fig. 5b**–**e** and **Supplementary Table 4**). These sgRNAs were almost equally active, compared to the gX_{19} and GX_{19} sgRNAs, at the respective on-target sites in HAP1 cells, although they were less active than the two conventional sgRNAs at the target sites in K562 cells. On the basis of the specificity ratio of on-target to off-target indel frequencies, the two modified sgRNAs showed up to 660-fold greater specificity than the conventional sgRNAs (**Supplementary Table 5**).

DISCUSSION

The genome-wide specificity of the CRISPR-Cas system is of broad interest in the field of RNA-guided genome editing. However, seemingly conflicting results have been reported about RGEN specificity. We had used, first, the T7E1 assay[15] and, later, deep

© 2015 Nature America, Inc. All rights reserved.

sequencing[20] to show that RGENs do not leave off-target footprints even at sites that differ from on-target sites by 2 nt in human cells. In line with these results, both whole-exome sequencing[20] and WGS[38,39] had shown that no off-target indels are induced by RGENs in the entire human exome and genome, respectively, in a clonal population of cells. In sharp contrast, several other groups had reported that RGENs can induce off-target indels at sites that differ by up to 5 nt (refs. 17–21) or that lack or contain extra nucleotides compared to the target sequences[22]. These results suggest that a given nuclease might recognize and cleave up to thousands of off-target sites in the genome. Measuring off-target effects at all of these potential sites one by one is an almost impossible task. RGENs may induce indels at hundreds or thousands of off-target sites with frequencies below 1% or 0.1%, respectively, which cannot be detected by sequencing only a few clones[40]. It is also possible that sgRNAs have widely different specificities: some may be highly specific, whereas others are promiscuous. To address these issues, we chose two promiscuous sgRNAs that had been shown by others to induce high-frequency off-target mutagenesis in human cells and examined their genome-wide off-target effects via Digenome-seq.

For profiling genome-wide off-target effects of engineered nucleases in an unbiased manner, SELEX (systematic evolution of ligands by exponential enrichment) and chromatin immunoprecipitation–sequencing have been used[41–43]. These methods rely on DNA binding rather than DNA cleavage. Unfortunately, these events are often uncoupled: most off-target DNA-binding sites recognized by dCas9 are not cleaved at all by Cas9 in cells. Integrase-defective lentiviral vector (IDLV) capture and *in vitro* selection are two different methods that detect nuclease cleavage sites rather than binding sites[26,27]. These two methods are complementary, but neither is comprehensive. Because the efficiency of IDLV capture at a given site is much lower than the mutation frequency at that site, many bona fide off-target sites cannot be captured by this method. *In vitro* selection uses RGENs to cleave a biased DNA substrate library that consists of $>10^{12}$ variants and determines cleaved library members by deep sequencing. Unfortunately, most cleaved sequences do not exist in the genome[17]. Furthermore, this approach cannot examine off-target cleavage at sequences that can be hybridized with a DNA/RNA bulge.

Here we showed that Digenome-seq is a reproducible and sensitive method for profiling nuclease off-target effects in an unbiased manner. First, Digenome-seq relies on DNA cleavage rather than binding. Second, unlike *in vitro* selection approaches, Digenome-seq is performed in a genomic context. Notably, potential off-target sites with a DNA/RNA bulge are captured by Digenome-seq. Third, Digenome-seq is sensitive enough to detect off-target sites at which indels are induced with a frequency of 0.1% or lower, near the detection limit of high-throughput sequencing platforms. Fourth, Digenome-seq is cost efficient: with the Illumina HiSeq X Ten, WGS now costs 1,000 US$. Last, this method is reproducible, as shown by three independent digenome analyses.

Very recently, two groups reported methods, termed HTGTS[44] and GUIDE-seq[45], of capturing double-strand breaks (DSBs) induced by Cas9 in cells. We noted that *VEGFA* sgRNA was analyzed in these studies as well. Interestingly, eight out of nine validated on- and off-target sites identified by Digenome-seq were also captured by both HTGTS and GUIDE-seq, demonstrating that these three methods produced comparable results. Unlike HTGTS and GUIDE-seq, Digenome-seq is not limited by chromatin accessibility. Furthermore, HTGTS and GUIDE-seq require (i) filtering out captured sites with poor homology to on-target sites to discard the many false positive sites that arise from randomly occurring DSBs in cells or from PCR artifacts and (ii) searching for homologous sites around captured sites, owing to indels of variable lengths that accompany DSB repair in cells; both of these can introduce bias.

In this study we applied Digenome-seq to conventional Cas9, but this method should also allow profiling genome-wide off-target effects caused by dCas9-FokI, paired Cas9 nickases or other programmable nucleases. Before such nucleases are used in a gene or cell therapy application, their off-target effects could be monitored carefully via Digenome-seq to avoid unwanted mutations. This method will also provide a litmus test in the development of next-generation genome-editing tools.

METHODS

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

Accession codes. NCBI Sequence Read Archive: SRX743626.

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

ACKNOWLEDGMENTS

This research was supported by grants from the Institute for Basic Science (IBS-R021-D1) to J.-S.K., Korea Health Industry Development Institute (HI14C1277) to J.-I.K., and Korea Institute of Planning and Evaluation for Technology of Food, Agriculture, Forestry and Fisheries (311011-05-3-SB010) to S.K.

AUTHOR CONTRIBUTIONS

D.K., E.K. and H.R.Y. performed the experiments. D.K., S.B., J.P., J.H. and J.-I.K. performed bioinformatics analyses. J.-S.K. and S.K. supervised the research.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Porteus, M.H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (2003).
2. Bibikova, M., Beumer, K., Trautman, J.K. & Carroll, D. Enhancing gene targeting with designed zinc finger nucleases. *Science* **300**, 764 (2003).
3. Urnov, F.D. *et al.* Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**, 646–651 (2005).
4. Kim, H.J., Lee, H.J., Kim, H., Cho, S.W. & Kim, J.S. Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. *Genome Res.* **19**, 1279–1288 (2009).
5. Kim, H., Um, E., Cho, S.R., Jung, C. & Kim, J.S. Surrogate reporters for enrichment of cells with nuclease-induced mutations. *Nat. Methods* **8**, 941–943 (2011).
6. Miller, J.C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.* **29**, 143–148 (2011).
7. Kim, Y. *et al.* A library of TAL effector nucleases spanning the human genome. *Nat. Biotechnol.* **31**, 251–258 (2013).
8. Kim, Y.K. *et al.* TALEN-based knockout library for human microRNAs. *Nat. Struct. Mol. Biol.* **20**, 1458–1464 (2013).
9. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
10. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
11. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).

© 2015 Nature America, Inc. All rights reserved.
npg

12. Hwang, W.Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229 (2013).
13. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
14. Cho, S.W., Lee, J., Carroll, D., Kim, J.S. & Lee, J. Heritable gene knockout in *Caenorhabditis elegans* by direct injection of Cas9-sgRNA ribonucleoproteins. *Genetics* **195**, 1177–1180 (2013).
15. Cho, S.W., Kim, S., Kim, J.M. & Kim, J.S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* **31**, 230–232 (2013).
16. Kim, H. & Kim, J.S. A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.* **15**, 321–334 (2014).
17. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
18. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
19. Cradick, T.J., Fine, E.J., Antico, C.J. & Bao, G. CRISPR/Cas9 systems targeting β-globin and *CCR5* genes have substantial off-target activity. *Nucleic Acids Res.* **41**, 9584–9592 (2013).
20. Cho, S.W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014).
21. Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
22. Lin, Y. *et al.* CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).
23. Brunet, E. *et al.* Chromosomal translocations induced at specified loci in human stem cells. *Proc. Natl. Acad. Sci. USA* **106**, 10620–10625 (2009).
24. Lee, H.J., Kim, E. & Kim, J.S. Targeted chromosomal deletions in human cells using zinc finger nucleases. *Genome Res.* **20**, 81–89 (2010).
25. Lee, H.J., Kweon, J., Kim, E., Kim, S. & Kim, J.S. Targeted chromosomal duplications and inversions in the human genome using zinc finger nucleases. *Genome Res.* **22**, 539–548 (2012).
26. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.* **32**, 279–284 (2014).
27. Ran, F.A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
28. Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* **31**, 833–838 (2013).
29. Tsai, S.Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32**, 569–576 (2014).
30. Guilinger, J.P., Thompson, D.B. & Liu, D.R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.* **32**, 577–582 (2014).
31. Kim, S., Kim, D., Cho, S.W., Kim, J. & Kim, J.S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).
32. Ramakrishna, S. *et al.* Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. *Genome Res.* **24**, 1020–1027 (2014).
33. Carette, J.E. *et al.* Ebola virus entry requires the cholesterol transporter Niemann-Pick C1. *Nature* **477**, 340–343 (2011).
34. Kim, Y., Kweon, J. & Kim, J.S. TALENs and ZFNs are associated with different mutation signatures. *Nat. Methods* **10**, 185 (2013).
35. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
36. Bae, S., Park, J. & Kim, J.S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **35**, 1473–1475 (2014).
37. Kim, J.M., Kim, D., Kim, S. & Kim, J.S. Genotyping with CRISPR-Cas-derived RNA-guided endonucleases. *Nat. Commun.* **5**, 3157 (2014).
38. Smith, C. *et al.* Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. *Cell Stem Cell* **15**, 12–13 (2014).
39. Veres, A. *et al.* Low incidence of off-target mutations in individual CRISPR-Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing. *Cell Stem Cell* **15**, 27–30 (2014).
40. Tsai, S.Q. & Joung, J.K. What's changed with genome editing? *Cell Stem Cell* **15**, 3–4 (2014).
41. Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat. Biotechnol.* **32**, 677–683 (2014).
42. Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).
43. Gabriel, R. *et al.* An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat. Biotechnol.* **29**, 816–823 (2011).
44. Frock, R.L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* doi:10.1038/nbt.3101 (15 December 2014).
45. Tsai, S.Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* doi:10.1038/nbt.3117 (16 December 2014).

© 2015 Nature America, Inc. All rights reserved.

ONLINE METHODS

Cas9 and *in vitro* sgRNA. Recombinant Cas9 protein was purified from *Escherichia coli* as described previously[37] or purchased from ToolGen. sgRNAs were synthesized by *in vitro* transcription using T7 RNA polymerase as described previously[37]. Briefly, sgRNA templates were mixed with T7 RNA polymerase in a reaction buffer (40 mM Tris-HCl, pH 7.9) containing 6 mM $MgCl_2$, 10 mM DTT, 10 mM NaCl, 2 mM spermidine, NTP, and RNase inhibitor. The reaction mixture was incubated at 37 °C for 8 h. Transcribed sgRNAs were purified using PCR purification kits (Macrogen) and quantified using NanoDrop (Thermo Fisher).

Cell culture and transfection conditions. HAP1 cells were obtained from Haplogen and cultured in IMDM medium supplemented with 10% FBS. HAP1 cells (1×10^6) were cotransfected with the Cas9 expression plasmid (400 ng), sgRNA-encoding plasmid (400 ng), and the plasmid (200 ng) encoding the blasticidin-resistance gene using FuGENE (Promega). Transfected cells were enriched by treatment with 20 μg/ml blasticidin. Single cell–derived gene knockout cells were obtained by limiting dilution. K562 cells (ATCC) were maintained in RPMI medium supplemented with 10% FBS. K562 cells (2×10^5) were electroporated with the Cas9 expression plasmid (1 μg) and sgRNA-encoding plasmid (1 μg) using Nucleofector (Lonza). Cells were not tested for mycoplasma contamination.

***In vitro* cleavage of genomic DNA.** Genomic DNA was purified with the DNeasy Tissue kit (Qiagen) according to the manufacturer's instructions. To digest target sequences in the genome, we mixed Cas9 protein (0.004–40 μg), which had been preincubated with sgRNA (0.003–30 μg) at room temperature for 10 min, with genomic DNA (8 μg) in a reaction volume of 400 μl (100 mM NaCl, 50 mM Tris-HCl, 10 mM $MgCl_2$, and 100 μg/ml BSA) and incubated at 37 °C for 8 h. Digested genomic DNA was purified again with DNeasy Tissue kit (Qiagen) after RNase A (50 μg/ml) was added to remove sgRNA. Purified digested genomic DNA was mixed with 2× Sybr Green Master Mix and analyzed by real-time quantitative PCR (qPCR). DNA sequences of primers used for qPCR are shown in **Supplementary Table 6**. The percentage of target site cleavage was measured using the comparative C_T method[46].

T7E1 assay. Genomic DNA was isolated using DNeasy Tissue kit (Qiagen) according to the manufacturer's instructions. The target site was amplified by PCR. The T7E1 assay was performed as described previously[4]. Briefly, amplified PCR products were denatured by heating and annealed slowly using a thermocycler. Annealed products were incubated with T7 endonuclease I (ToolGen) for 20 min at 37 °C and size separated by agarose gel electrophoresis.

Targeted deep sequencing. Genomic DNA segments spanning the on-target and potential off-target sites were amplified using Phusion polymerase (New England BioLabs). The resulting PCR amplicons were subjected to paired-end sequencing using Illumina MiSeq. Primer information can be found in **Supplementary Table 6**. Indels located 3 bp upstream of the PAM were considered to be the mutations induced by RGENs. The deep sequencing data are available at the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRX743626.

Whole-genome and digenome sequencing. Genomic DNA (1 μg) was fragmented using the Covaris system (Life Technologies) and polished to generate blunt ends using End Repair Mix. Fragmented DNA was ligated with adaptors to produce libraries, which were then subjected to WGS using an Illumina HiSeq X Ten Sequencer at Macrogen. We used the following mapping program and parameters: Isaac aligner, ver. 01.14.03.12; Human genome reference, hg19 from UCSC (original GRCh37 from NCBI, Feb. 2009); Base quality cutoff, 15; Keep duplicate reads, yes; Variable read length support, yes; Realign gaps, no; and Adaptor clipping, yes (adaptor: AGATCGGAAGAGC*,*GCTCTTCCGATCT)[47]. Indels were called by Isaac software.

The computer program used for capturing *in vitro* RGEN cleavage sites in sequenced genomes is available at https://github.com/snugel/digenome-toolkit. To identify double-peak positions, we used sequence reads with a mapping quality of 1 or greater without considering base sequence quality. The threshold of minimum ten reads at a given position was used to avoid too many false positives, while capturing bona fide off-target sites.

Analysis of off-target effects at homologous sites. We used Cas-OFFinder (http://www.rgenome.net) to find potential off-target sites that differed from on-target sequences by up to 8 nt and that differed by up to 2 nt with a 1- to 5-nt DNA or RNA bulge. Next, we obtained CIGAR string information around ±10 bp from potential cleavage sites in BAM files and derived the most common CIGAR strings. Next, we compared the most common CIGAR strings with wild-type sequences to identify candidate sites with indels. The computer program used in this study is available upon request. We used IGV to validate or invalidate indels at these candidate sites one by one.

Reproducibility. The experiments were not randomized and no experimenters were blinded. *P* values were calculated by the Fisher exact test. No sample-size estimates were performed to ensure adequate power to detect a prespecified effect size. No samples were excluded from analysis.

46. Schmittgen, T.D. & Livak, K.J. Analyzing real-time PCR data by the comparative C_T method. *Nat. Protoc.* **3**, 1101–1108 (2008).
47. Raczy, C. *et al.* Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041–2043 (2013).

© 2015 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.3284