이학박사학위논문

Genome-wide CRISPR/Cas9 off-target profiling via Digenome-seq

2016년 8월

서울대학교 대학원

화학부 생화학 전공

김 대 식

Abstract

Genome-wide CRISPR/Cas9 off-target profiling via Digenome-seq

Daesik Kim

Departments of Chemistry

The Graduate School

Seoul National University

Targeted genome editing using engineered nuclease such as zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeat (CRISPR)/CRISPR associated protein (Cas) systems have been used in cultured cells and whole organisms for functional study and therapeutic study.

Despite broad interest in CRISPR/Cas9 mediated

genomengineering, off-target effects of entire genome have not been established. Therefore, development of methods to profiling genome-wide CRISPR/Cas9 off-target effects is the major challenge in this area.

In this study, I characterize CRISPR/Cas9 off-target effect in clonal cells and bulk populations of cells. First I used Isaac variant calling program to analyze genome-wide indels in clonal cells. Second, I developed nuclease-digested genomes sequencing (digenome-seq) to profiling genome-wide CRISPR/Cas9 off-target effects in bulk populations. Using this methods, I validated off-target sites which indels were induced with frequencies below 0.1% and validated off-target effects can be avoided by replacing with modified sgRNAs. Third, I developed multiplex digenome-seq which can profiling more than ten sgRNA off-target effects in a one time. Based on multiplex digenome-seq result, I made a program for the choice of target sites and the off-target sites predictor respectively.

Keywords : Genome engineering, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), CRISPR-associated protein (Cas), Whole genome sequencing (WGS)

Student number : 2013-30088

Table of Contents

List of Figures

List of Tables

I. Introduction

Programmable endonucleases have been established as flexible tools for genome manipulation in cultured cells and whole organisms (Kim and Kim, 2014). These include zinc finger nucleases (ZFNs) (Bibikova et al., 2003; Kim et al., 2011; Kim et al., 2009; Porteus and Baltimore, 2003; Urnov et al., 2005), transcription activator-like effector nucleases (TALENs) (Kim et al., 2013a; Kim et al., 2013b; Miller et al., 2011), and RNA-guided engineered nucleases (RGENs) (Cho et al., 2013; Cong et al., 2013; Hwang et al., 2013; Jiang et al., 2013; Jinek et al., 2013; Mali et al., 2013b) derived from the type II clustered regularly interspaced repeat (CRISPR)/CRISPR-associated (Cas) system. For the last several years, genome editing via programmable nucleases has been transforming almost every discipline in life science, biotechnology, and medicine. For example, targeted genetic modifications in stem and somatic cells are expected to pave the way for novel gene/cell therapy for the treatment of diverse genetic and acquired diseases (Park et al., 2014; Perez et al., 2008; Wu et al., 2013).

ZFNs consist of the zinc finger protein and DNA cutting restriction enzyme (FokI) (Kim et al. 1996; Bitinaite et al. 1998). The engineered nuclease induces double strand breaks (DSB) through FokI dimerization (Urnov et al. 2010). Like ZFNs, TALENs are also composed of the FokI nuclease, but the DNA binding domain is made up of transcription activator-like effectors (TALEs) derived from the

plant pathogen Xanthomonas spp. bacterium (Boch et al. 2009; Moscou and Bogdanove 2009). TALENs have greater success rates, higher average mutation rates, and lower cytotoxicity than ZFNs (Kim and Kim, 2014).

CRISPR/Cas system is an adaptive prokaryotic adaptive immune system that destroys foreign DNA via three steps. First, fragmented foreign DNA is inserted into CRISPR array between repeat regions. Second, pre-crRNA is expressed in CRISPR locus, and is maturated to CRISPR RNA (crRNA) by RNase III. Third, maturated crRNA interacts with Cas9 protein for target degradation (Jinek et al., 2012). RGENs consist of the Cas9 endonuclease derived from S. pyogenes and single guide RNAs (sgRNAs) that recognize target DNA by Watson-Crick base pairing. This family of Cas9 proteins has two functional nuclease domains, RuvC and NHN. The RuvC domain binds to the non-target DNA strand, and the HNH nuclease domain cleaves the target DNA strand (Nishimasu et al., 2014). This protein recognizes a 5'-NGG-3' protospacer-adjacent motif (PAM) sequence and forms a RNP complex with sgRNA.

These programmable nucleases (ZFN, TALEN, and RGEN) produce site-specific DSBs in the genome, which greatly stimulates targeted mutagenesis and chromosomal rearrangements by homologous recombination (HR) in the presence of donor DNA, or nonhomologous end joining (NHEJ) in the absence of homology templates (Kim and Kim, 2014). Co-injection of a programmable endonuclease and a donor DNA template that contains homology sequences or single-strand

oligodeoxynucleotides (ssODN) produces a single codon change or DNA insertion via HR. Point mutations can be corrected or induced using the HR mechanism. Without donor DNA, DSBs induced small insertion and deletion (indel) or chromosomal rearrangements such as large deletions, duplications and inversions (Brunet et al., 2009; Cho et al., 2014; Lee et al., 2010; Lee et al., 2012).

Unfortunately, recent studies demonstrate that RGENs produce several off-target effects. Recognition of 5'-NAG-3' and 5'-NGA-3' PAM or several mismatches between sgRNA and target DNA, especially in the PAM distal region, can induce mutations at that position (Cho et al., 2014; Cradick et al., 2013; Fu et al., 2013; Hsu et al., 2013; Lin et al., 2014; Pattanayak et al., 2013). To make matters worse RGENs can cleave off-target DNA sequences harboring an extra base (DNA bulge) or lacking a base (RNA bulge) compared to their respective sgRNA sequences (Lin et al., 2014). Off-target DNA cleavages can lead to mutations at unintended genomic loci such as proto-oncogenes and tumor suppressor genes, as well as gross chromosomal rearrangements such as translocations (Brunet et al., 2009; Cho et al., 2014), deletions (Lee et al., 2010), and inversions (Lee et al., 2012). These raise serious concerns about the use of programmable nucleases in research and medicine.

Several methods used to reduce off-target effects are studied as follows: sgRNAs with two extra guanine nucleotides at the 5' end (Cho et al., 2014), truncated sgRNAs (Fu et al., 2014), paired Cas9 nickases (Cho et al., 2014; Mali et al., 2013a; Ran et al., 2013), a catalytically

dead Cas9 (dCas9)-FokI fusion (Guilinger et al., 2014; Tsai et al., 2014), and purified Cas9/sgRNA complex delivery (Kim et al., 2014; Ramakrishna et al., 2014; Zuris et al., 2015). These different methods have been confirmed to reduced off-target effects in the homology sequence with on-target sequence. However off-target effects of entire genome have not been determined.

In the case of ZFN, off-target sites identification methods include systematic evolution of ligands by exponential amplification (SELEX) (Perez et al., 2008), integrase-deficient lentivirus (IDLV) capture in cells (Gabriel et al., 2011), and in vitro selection using a DNA substrate library for nuclease-mediated DNA cleavage (Pattanayak et al., 2011). But none of these methods are comprehensive enough to allow unbiased genome-wide analysis of nuclease specificity. For example, IDLV capture (Gabriel et al., 2011) and in vitro selection (Pattanayak et al., 2011) were independently used by two groups to examine the off-target effects of a CCR5-targeting ZFN, which has been under clinical trials for the treatment of HIV infection. Surprisingly, these two different methods captured entirely different sets of off-target sites in addition to the highly homologous CCR2 site that was already known to be an off-target site, highlighting the limitations of these methods.

Off target sites for RGEN have been found with bioinformatic predictions based on sequence homology (Bae et al., 2014), mismatched guide RNA libraries (Hsu et al., 2013), *in vitro* selection (Pattanayak et al., 2013), reporter assays (Fu et al., 2013), and chromatin

immunoprecipitation coupled with deep sequencing (ChIP-Seq) (Kuscu et al., 2014; Wu et al., 2014). These studies demonstrate that RGENs have off-target effects, however, potential off-target sites have been computationally determined by homology sites. To address this critical issue in the field, it is imperative to develop methods that check the specificity of RGENs and other nucleases on a genomic scale in an unbiased manner.

In this study, I analyze genome-wide off-target effects of RGEN using whole genome sequencing (WGS) in clonal cells and in bulk populations of cells. First, I have subjected gene knockout clonal cells to WGS, and analyzed genome-wide indels using Isaac variant calling program. No off-target indels were found in these clonal cells. Second, I have developed nuclease-digested genomes sequencing (Digenome-seq) that profiles RGEN off-target effects in human population cells using one or several types of RGEN. Our results show that even 'promiscuous' RGENs are highly specific, inducing off-target mutations at only a handful, rather than hundreds or thousands, of sites in the entire genome. Based on Digenome-seq results I have finally developed a computer program to predict RGEN targetable sites which can minimize genome-wide off-target effects of CRISPR-Cas9.

II. Materials and Methods

1. Cas9 and *in vitro* sgRNA

Recombinant Cas9 protein was purchased from ToolGen (South Korea). sgRNAs were synthesized by *in vitro* transcription using T7 RNA polymerase as described previously (Kim et al., 2014). Briefly, sgRNA templates were generated by two complementary oligonucleotides which are quality-checked using MALDI-TOF (Macrogen) using annealing and extension. These templates were cloned by TA cloning (Enzynomics), and DNA template sequence was confirmed by capillary sequencing (Macrogen). 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, and RNase inhibitor, pH 7.9) at 37℃ for 8 hr. Transcribed sgRNAs were pre-incubated with DNaseI to remove template DNA, and purified using PCR purification kits (Macrogen).

2. Cell culture and transfection conditions

HAP1 cells were obtained from Haplogen and cultured in IMDM media supplemented with 10% FBS. HAP1 cells were co-transfected with the Cas9 expression plasmid, sgRNA-encoding plasmid, and the plasmid encoding the blasticidin resistance gene using

Fugene (Promega). Transfected cells were enriched by treatment with 20ug/ml blasticidin. Single cell-derived gene knockout cells were obtained by limiting dilution. K562 cells were maintained in RPMI media supplemented with 10% FBS. K562 cells were electoporated with the Cas9 expression plasmid and sgRNA-encoding plasmid using Nucleofector (Lonza). HeLa cells were cultured in DMEM media supplemented with 10% FBS. HeLa cells ($8x10^4$) were co-transfected with the Cas9 expression plasmid (500ng) and the sgRNA -encoding plasmid (500ng) using lipofectamine 2000 (LifeTechnologies). Genomic DNA was isolated with the DNeasy Tissue kit (Qiagen) according to the manufacturer's instructions after 48hr.

3. *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, Cas9 protein (0.004ug to 40ug), which had been pre-incubated with sgRNA (0.003ug to 30ug) at room temperature for 10min, was mixed with genomic DNA (8ug) in a reaction volume of 400uL (100mM NaCl, 50mM Tris-HCl, 10mM $MgCl_2$, and 100μg/ml BSA) and incubated at 37℃ for 8h. Digested genomic DNA was purified again with DNeasy Tissue kit (Qiagen), after RNase A (50ug/mL) was added to remove sgRNA. Purified digested genomic DNA was mixed with 2x SYBR Green Master Mix and analyzed by real-time quantitative PCR (qPCR). The percentage of target site

cleavage was measured using the $\triangle\triangle C_T$ method (Schmittgen and Livak, 2008).

4. 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 (Kim et al., 2009). 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℃, and size-separated by agarose gel electrophoresis.

5. 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. Indels located 3-bp upstream of the PAM were considered to be the mutations induced by RGENs.

6. Whole genome and digenome sequencing

Genomic DNA (1ug) was fragmented using the Covaris system

(Life Technologies) and polished to generate blunt ends using End Repair Mix. Fragmented DNA was ligated with adapters to produce libraries, which were then subjected to WGS using an Illumina HiSeq X Ten Sequencer at Macrogen. Indels were called by ISSAC software.

7. Analysis of off-target effects at homologous sites

I used Cas-OFFinder (www.rgenome.net) to find potential off-target sites that differed from on-target sequences by up to 8 nucleotides and that differed by up to 2 nucleotides with a 1-nt to 5-nt DNA or RNA bulge. Next, I obtained cigar string information around +/- 10 bp from potential cleavage sites in BAM files and derived the most common cigar strings. Next, I 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. I used IGV to validate or invalidate indels at these candidate sites one by one.

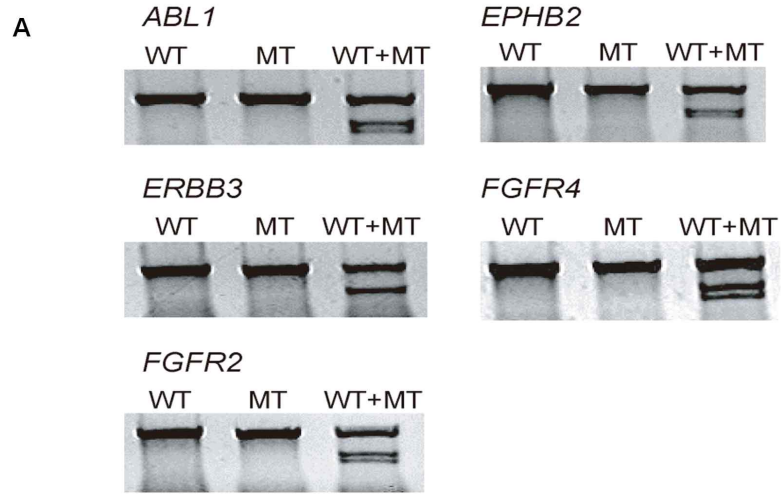
III. Result

A. Off-target analysis of clonal cells using whole genome sequencing (WGS)

1. Generation of mutant human haploid cells.

HAP1 is a human haploid cell which is generated by KBM7, a leukemia cell line (Carette et al., 2009). HAP1 is easy to make homologues Knockout, because this cell has only one copy for its chromosomes. I generated five kinds of KO HAP1 cell lines, each with a single disruption in a kinase gene (ABL1, EPHB2, ERBB3, FGFR2 and FGFR4). RGEN mediated gene KO was confirmed by T7E1 assay and Sanger sequencing (Figure 1A, B).

Figure 1.



B

ABL1	WT	GCTTTGGGGCTGGATAATGGAGCGTGGTGATGA	
	MT	GCTTTGGGGCTGGATAA-------TGGTGATGA	-7
EPHB2	WT	AACTGGCTACGGACCAAGTTTATCCGGCGCCGT	
	MT	AACTGGCAACGG----------TCCGGCGCCGT	-10
ERBB3	WT	CTCTACCATTGCCCAACCTCCGCGTGGTGCGAG	
	MT	CTCTACCATTGCCCAACCTCC----GGTGCGAG	-4
FGFR4	WT	TGCAGAATCTCACCT----TGATTACAGGTGGT	
	MT	TGCAGAATCTCACCTGCAC-----------GT	-12,+4
FGFR2	WT	CCTGTCTCCGCAGGGGGATACGTTTGGGGTCAG	
	MT	CCTGTCTCCGCA-------------GGGGTCAG	-13

Figure 1. Generation of gene KO in HAP1 haploid cells (A) Gene knockout was confirmed in clonal populations of HAP1 haploid cells. WT, wild-type; MT, mutant; WT+MT, a 1:1 mixture of WT and MT PCR amplicons. (B) DNA sequences of wild-type and mutant clones. The PAM is shown in blue. Inserted bases are shown in red.

2. Whole genome sequencing of human haploid cells

Genomic DNA isolated from these mutant HAP1 clones and wild-type HAP1 cells were subjected to WGS via Illumina HiSeq X. Wild-type and ABL1 KO HAP1 cells genomic DNAs were sequenced twice to check the WGS reproducibility. I used Isaac variant calling program (Raczy et al., 2013) to find RGEN-induced off-target indel genome widely (Figure 2). I used several bioinformatics filter to find off-target indel which is induced by RGEN (Table 1 and Figure 2).

First, indels presented in the public database and heterozygous indels were discarded. Second, I compared each KO clone indels with other types of KO or wild-type indels, and overlap indels have removed to get each KO clone-specific indels. Third, I compared RGEN target sites with candidates, and less than 10 nucleotide mismatch with 5'-N(G/A)G-3' PAM have been chosen for next step. Finally I got 9 to 84 indel sites contained a 5'-N(G/A)G-3' PAM sequence and had at least 10 nucleotide matches with respective on-target sequences. Only one candidates validated by Sanger sequencing in ERBB3 KO clone, and this site was not present in the wild-type genome (Figure 3A).

To determine whether validated indel was caused by an RGEN off-target effect, I delivered ERBB3 targeting Cas9 RGEN to HAPI cells. The indel was not detected in validated sites, so I conclude this site was induced by spontaneous mutation (Drake et al., 1998) which is acquired by cell culture (Figure 3B).

The Isaac variant calling program missed one of the five on-target mutation. This mutation consisted of a 12-bp deletion and a 4-bp insertion at the target site (Figure 3C, D). This result show that the limitation of the variant calling algorithm.

Figure 2

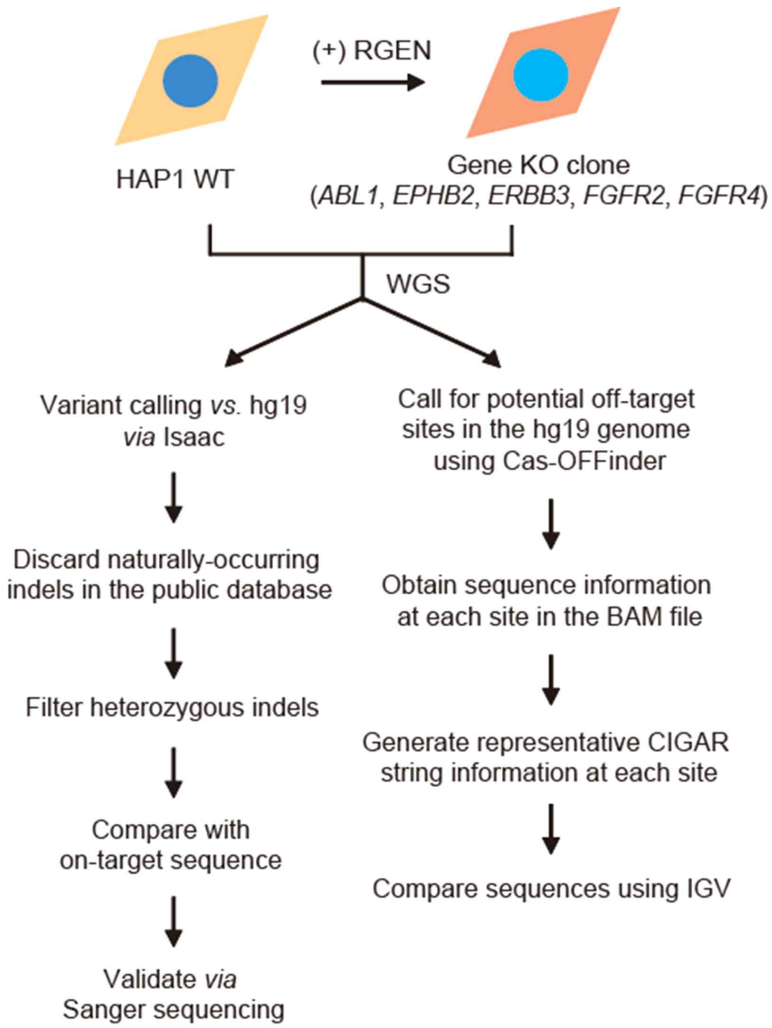


Figure 2. Analysis of off-target effects in gene KO clones via whole genome sequencing (WGS). Schematic workflow of off-target analysis of gene KO clones via WGS.

Table 1. WGS of human haploid cells

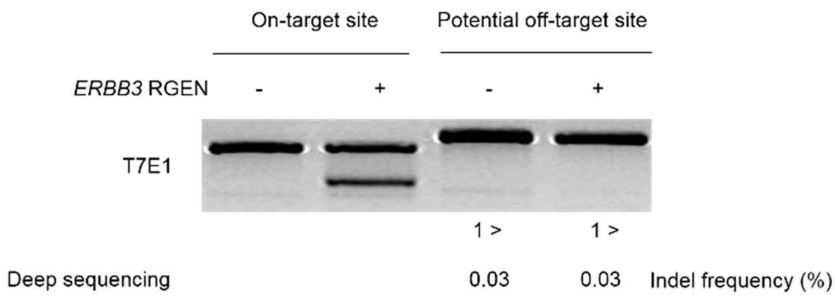
			Knockout clone					
	WT (Experiment 1)	WT (Experiment 2)	*ABL1* (Experiment 1)	*ABL1* (Experiment 2)	*EPHB2*	*ERBB3*	*FGFR2*	*FGFR4*
Raw indel calls	331,325	337,993	322,090	323,620	316,917	319,022	309,976	312,553
Indels not present in the indel database	38,009	40,874	37,376	38,282	36,232	36,642	34,323	34,841
Haploid indels	29,159	30,851	29,707	30,005	28,779	29,142	27,202	27,961
KO clone-specific indels	2,226	2,968	2,910	3,250	2,107	2,196	2,026	2,061
Candidate indels at homologous sites	N/A	N/A	84	73	9	17	17	15
Confirmed by Sanger sequencing	N/A	N/A	0	0	0	1	0	0

Figure 3

A

```
GGTTCCAGTGGGCTGCTTCAGGGAAGGCGGGCCGGGAATTCCTCTCTGTAAGAC   WT HAP1
GGTTCCAGTGG-----------------------------------------GAC   ERBB3 KO clone
```

B



C

```
TGCAGAATCTCACCT----TGATTACAGGTGGT   FGFR4 WT
TGCAGAATCTCACCTGCAC----------GT   FGFR4 MT -12, +4
```

D

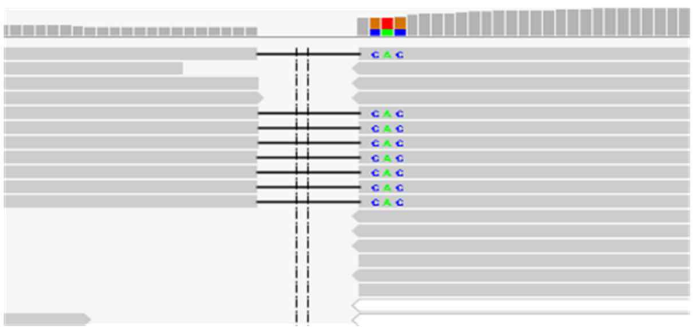


Figure 3. Analysis of off-target effects in gene KO clones via whole genome sequencing (WGS). (A) Off-target candidates with a small deletion in the ERBB3 KO clone which is confirmed by Sanger sequencing. (B) RGEN-mediated mutagenesis at the on-target and potential off-target sites. Mutation frequencies (%) were measured using T7E1 and targeted deep sequencing. (C) The on-target mutant sequence in the FGFR4 KO clone. The PAM sequence is shown in blue and inserted bases are shown in red. (D) Integrative Genomics Viewer (IGV) image at the FGFR4 on-target site.

3. Examining potential off-target sites.

Sometime RGEN induce insertion and deletion together (Figure 3D), which is hard to identified by variant calling algorithms. To solve this problem, I made computer program which generate the consensus sequence of on-/off-target sites using cigar string information (Figure 4). For that, I used Cas-OFFinder to import the list of potential off-target sites that differed from on-target sites by up to 8 nucleotides or that differed by 2 nucleotides with a DNA or RNA bulge (Figure 5A). I made consensus sequence of these potential off-target sites, and this program only identified on-target mutation (Figure 5B). These result shows that it is difficult to find off-target mutation using WGS analysis of clonal cells.

Figure 4

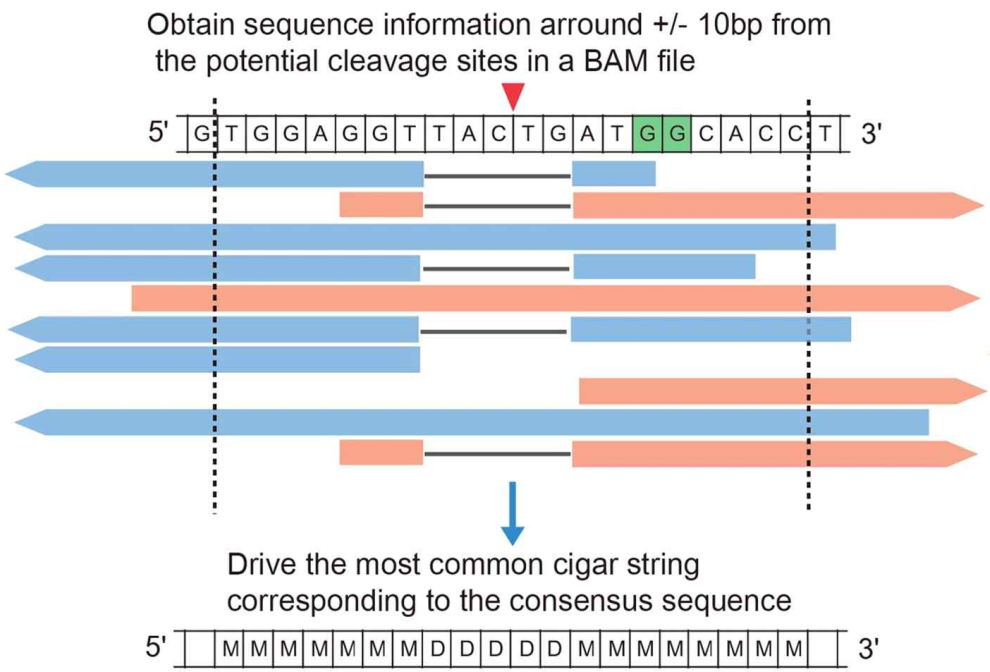


Figure 4. Schematic of consensus sequence generation.

Figure 5

A

	Number of sites with no bulge	Number of sites with a bulge
ABL1	766,196	14,506
EPHB2	381,542	2,524
ERBB3	363,617	2,424
FGFR2	436,084	2,037
FGFR4	1,303,358	8,395

B

On-target sites

ABL1	WT	MMMMMMMMMMMMMMMMMMM	*FGFR2*	WT	MMMMMMMMMMMMMMMMMMM
ABL1	MT	MMMMDDDDDDDMMMMMMMMM	*FGFR2*	MT	MDDDDDDDDDDDDDSSSMMM
EPHB2	WT	MMMMMMMMMMMMMMMMMMM	*FGFR4*	WT	MMMMMMMMMMMMMMMMMMM
EPHB2	MT	MDDDDDDDDDDSSSSSSSMM	*FGFR4*	MT	MMMMMMDDDDDDDDSSSSSS
ERBB3	WT	MMMMMMMMMMMMMMMMMMM			
ERBB3	MT	MMMMMMMMMMDDDDMMMMMM			

Figure 5. Analysis of off-target effects in gene KO clones via consensus sequence generation. (A) The number of potential off-target sites that differ from on-target sites by up to 8 nucleotides or by 2 nucleotides with a DNA or RNA bulge of up to 5 nucleotides in length. (B) On-target mutations in five KO clones identified by consensus sequence comparison.

B. Digenome-seq for genome-wide RGEN off-target profiling.

1. Genomic DNA digestion using RGENs *in vitro*.

I next asked whether I could profile RGEN off-target mutation in a bulk population of cells. To answer this question, I subject RGEN mediate digested genomes to WGS. For genomic DNA digestion, I mixed genomic DNA with pre-incubated Cas9 and HBB gene-specific sgRNA complex at variable concentrations that ranged from 0.03 nM to 300 nM. I next used quantitative PCR to examine DNA cleavage at these sites (Figure 6). In the HBB on-target and off-target 1 (OT1) sites, DNA was almost completely digested at a low RGEN concentration (0.03nM Cas9) (Figure 7A). However, the OT3 site was completely digested only at high RGEN concentration (Figure 7A). The other two sites, OT7 and OT12 were not digested even high RGEN concentration (Figure 7A). I next transfected this HBB targeting RGEN to HAP1 cells and mutation was detected by deep sequencing. HBB gene-specific RGEN induced high indel frequency at both on-target and OT1 sites (Figure 7B). OT3 site also detected mutation with a frequency of 4.3% (Figure 7B). The mutation of other two potential off-target sites, OT7 and OT12, were not detected using deep sequencing (detection limit, ~0.1%) (Figure 7B).

Figure 6

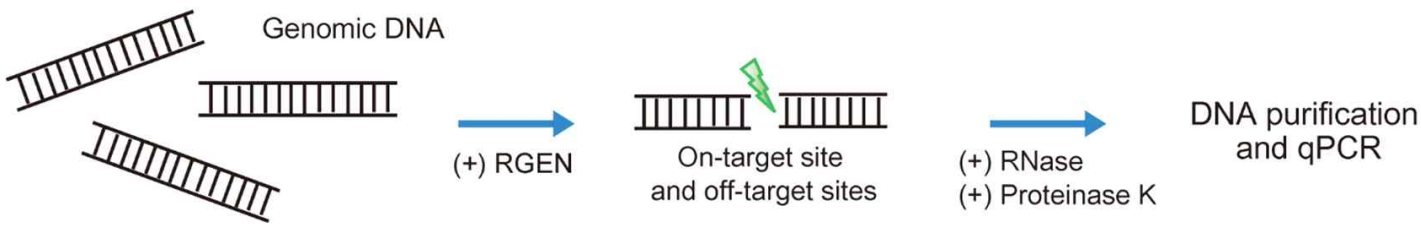
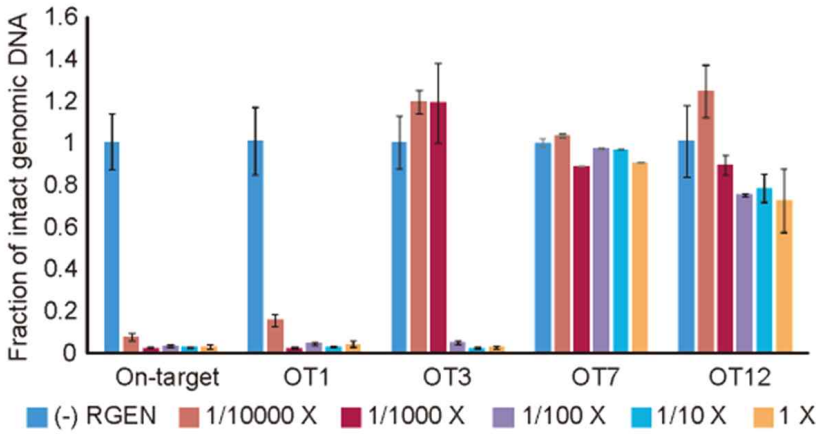


Figure 6. Schematic flow of RGEN-mediated *in vitro* cleavage of genomic DNA.

Figure 7

A



	Target sequence
On-target	CTTGCCCCACAGGGCAGTAACGG
OT1	TCAGCCCCACAGGGCAGTAAGGG
OT3	GCTGCCCCACAGGGCAGCAAAGG
OT7	CCTCTCCCACAGGGCAGTAAAGG
OT12	CCTGTCCCACAGGGCAGGAAGGG

B

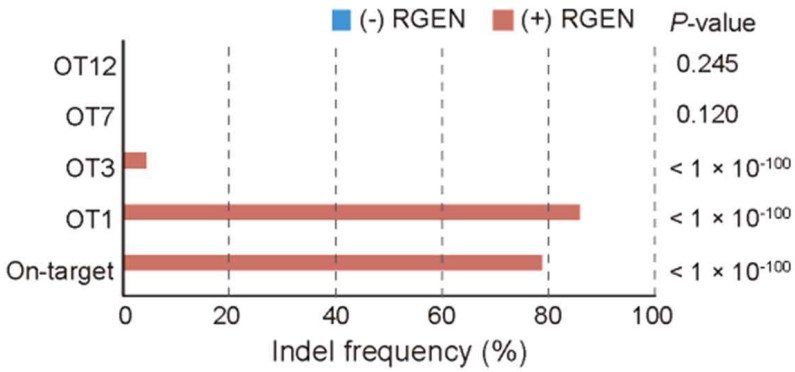


Figure 7. RGEN-mediated genomic DNA digestion *in vitro*. (A) Fraction of intact genomic DNA not cleaved by the HBB-targeting RGEN at on-target and four potential off-target sites. For the 1X reaction, Cas9 protein (40ug, 300nM) and sgRNA (30ug, 900nM) were incubated with 8ug of HAP1 genomic DNA in a volume of 400 uL for 8 h. Both Cas9 and sgRNA were serially diluted by 10-fold to 10,000-fold. The fraction of uncleaved DNA was measured by qPCR. (Bottom) DNA sequences of the on-target and the four potential off-target sites. Mismatched nucleotides are shown in red and the PAM sequence is shown in blue. (B) Measuring RGEN-driven mutation frequencies with the targeted deep sequencing at the on-target and potential off-target sites.

2. Nuclease-digested genomes sequencing (Digenome-seq) : Straight alignment vs. staggered alignment

Genomic DNA was purified from mock- and RGEN-transfected HAP1 cells before and after *in vitro* RGEN digestion at 300nM Cas9 protein. These four different genomic DNAs were subjected to WGS to observe difference between non-digested and digested DNA (Figure 8). After aligning sequence reads into the reference genome, I observed sequence alignments pattern at the on-target and the four different homologous sites using Integrative Genomics Viewer (IGV). Digested DNA has unusual patterns of straight alignments at the on-target, OT1, and OT3 sites (Figure 9 and Figure 10A, B). OT 7 and OT12 sites, which are invalidated off-target sites, showed both straight and staggered alignments (Figure 10C, D). In contrast, I cannot detect unusual sequence alignment in non-digested DNA. I next compared the digested DNA from intact genome with RGEN-transfected genome. I could examine both straight and staggered alignments at on-target, OT1, and OT3 sites (Figure 9 and Figure 10A, B). RGEN-mediated mutant sequence was not cleaved by *in vitro* RGEN treatments, so these mutant sequences induce staggered alignments.

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 bona fide off-target sites.

Figure 8

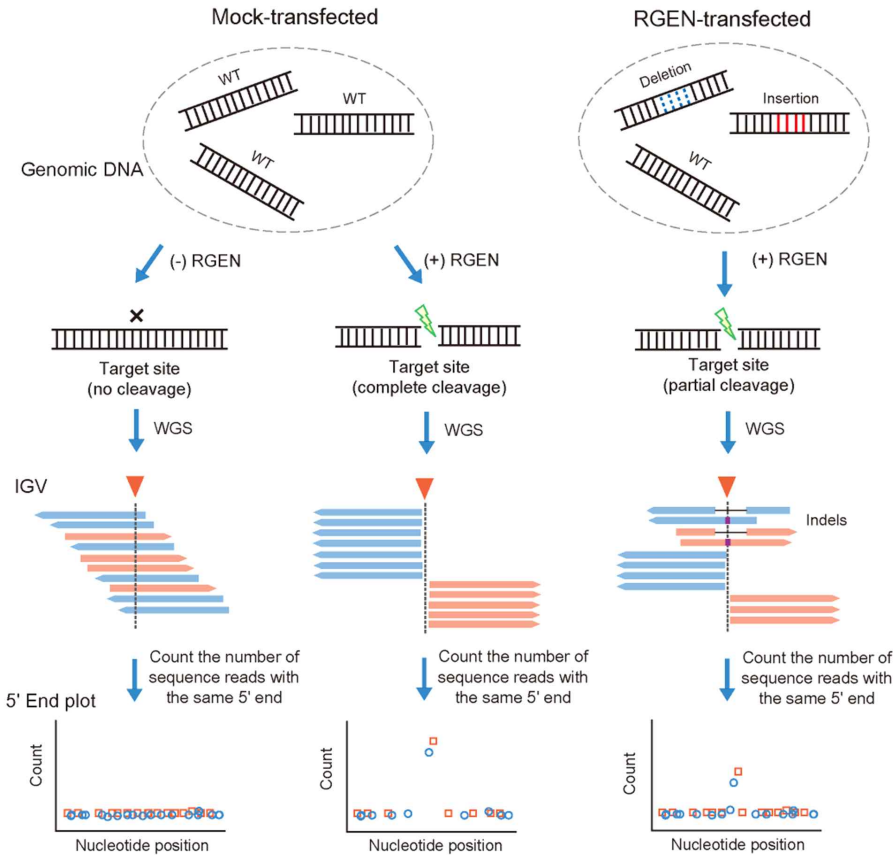


Figure 8. RGEN-induced digenome sequencing to profiling off-target sites. Schematic overview of nuclease-digested whole genome sequencing for the identification of off-target sites. Genomic DNA isolated from mock-transfected or RGEN-transfected cells is digested by the RGEN, and subjected to WGS. Sequence reads are aligned to the reference genome (hg19) and visualized using the IGV program. Forward and reverse sequence reads are shown in pink and blue, respectively. Red triangles and vertical lines indicate cleavage positions.

Figure 9

Figure 9. Representative IGV images obtained using the HBB-specific RGEN at the on-target site.

Figure 10

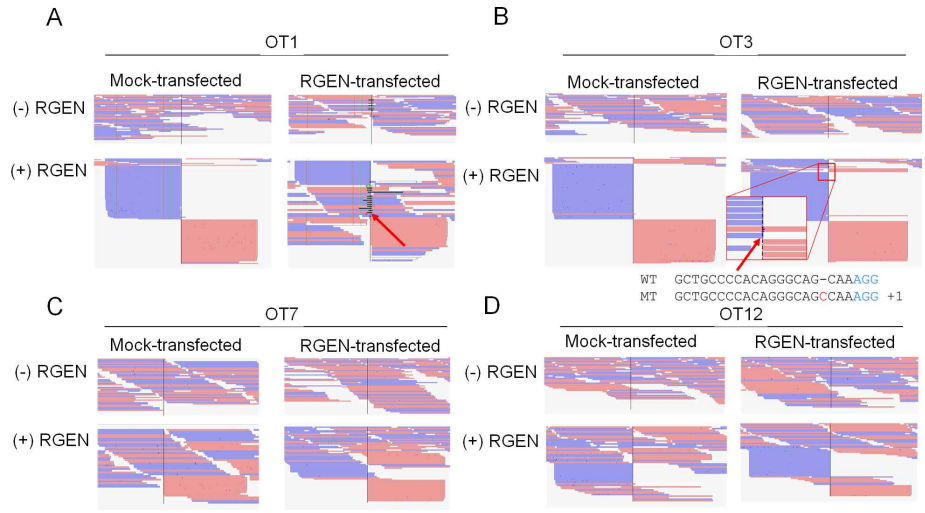


Figure 10. RGEN-induced Digenome sequencing to capture off-target sites. (A-D) Representative IGV images obtained using the HBB-specific RGEN at the potential off-target sites OT1 (A), OT3 (B), OT7 (C), and OT12 (D).

3. 5' End plot at single nucleotide resolution

To identify potential RGEN off-target sites on a genomic scale, I need to develop a computer program that searches for straight alignments of sequence reads. First, I 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 (Figure 11A, B). Because both Watson and Crick strands were sequenced, I assumed that almost an 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 (Figure 11B). The intact genome that had not undergone RGEN treatment did not detected such double-peak patterns at these sites (Figure 11B).

Next, I 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 in mock-transfected digenome and two different concentration (3nM Cas9, 300nM Cas9) RGEN-transfected digenome (Figure 11A, B).

Figure 11

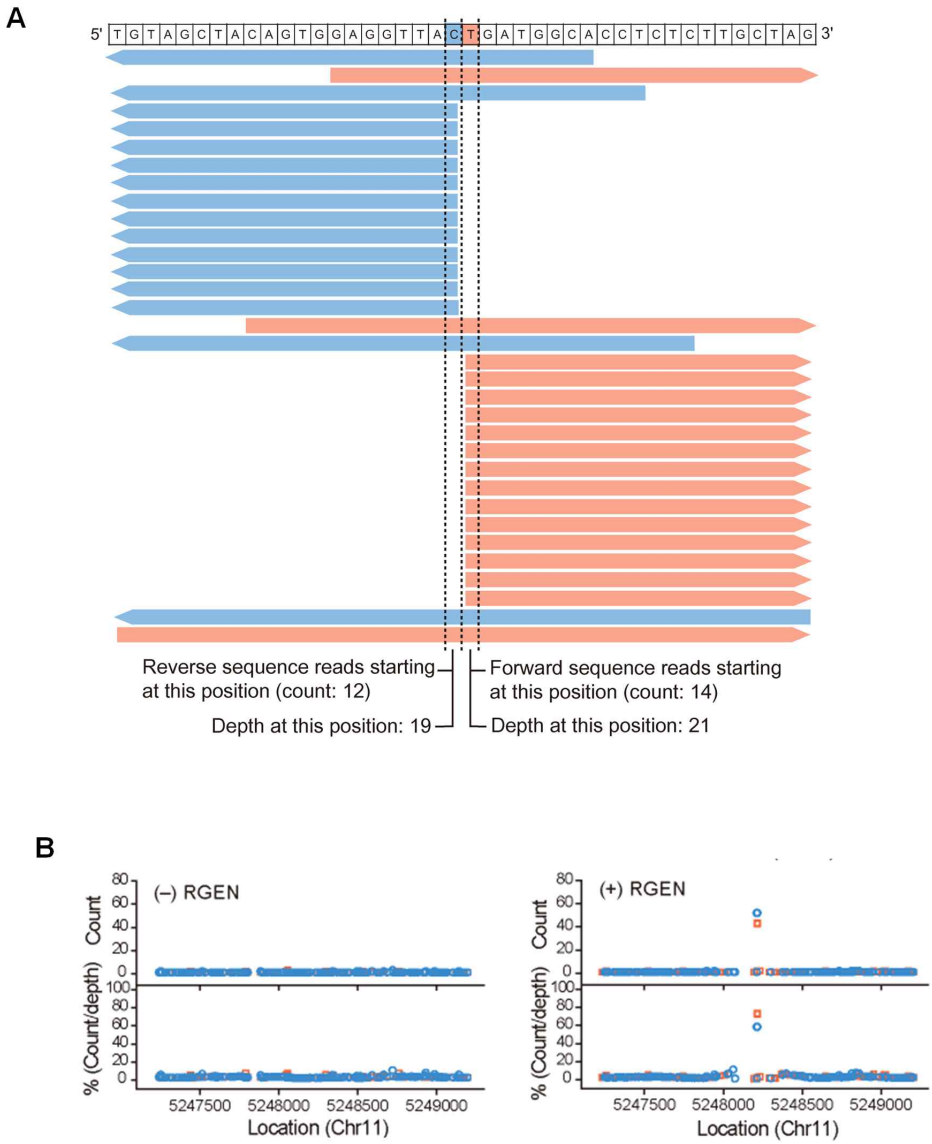

A
5' T G T A G C T A C A G T G G A G G T T A C T G A T G G C A C C T C T C T T G C T A G 3'
Reverse sequence reads starting at this position (count: 12)
Depth at this position: 19
Forward sequence reads starting at this position (count: 14)
Depth at this position: 21
B
(–) RGEN
(+) RGEN
Count
% (Count/depth)
80
60
40
20
0
100
5247500
5248000
5248500
5249000
Location (Chr11)

Figure 11. RGEN-induced Digenome sequencing to capture off-target sites. (A) 5’ End plots showing the absolute and relative number of sequence reads with the same 5’ end across nucleotide positions in OT1 and OT3 region. (B) 5’ End plots showing the absolute and relative number of sequence reads with the same 5’ end across nucleotide positions at on-target site.

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, which showed double-peak patterns in a 5' end plot and straight alignments in an IGV image (Figure 12A). Among these sites, one and two sites in the digenomes treated with 3 nM and 300 nM RGEN, respectively, were false positives that resulted from naturally-occurring indels (Figure 13A-C). Such patterns were observed at 125 sites, including the three validated on- and off-target sites in the RGEN-transfected digenome (Figure 12A). Importantly, the two invalidated OT7 and OT12 sites did not show double-peak patterns in 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. 74 (97%) of 76 candidate sites found in the mock-transfected digenome (300 nM) were also identified in the RGEN-transfected digenome (Figure 12A). I 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 nucleotides, all but the one at the 5' end were conserved (Figure 12B). Furthermore, the sequence logo or 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 two nucleotides (Figure 12B, C). I also found that the fewer mismatches there were in homologous sites, the more likely they were to be captured by Digenome-Seq. Thus, 7 out of 15 (47%) and 14 out of 142 (10%) homologous sites that differed by 3 and 4 nucleotides, respectively, from the on-target site were captured, but only 15 out of 1,191 sites (1.2%) and one out of 7,896 sites (0.013%) that differed by 5 and 6 nucleotides, respectively, were captured (Figure 12D). 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.

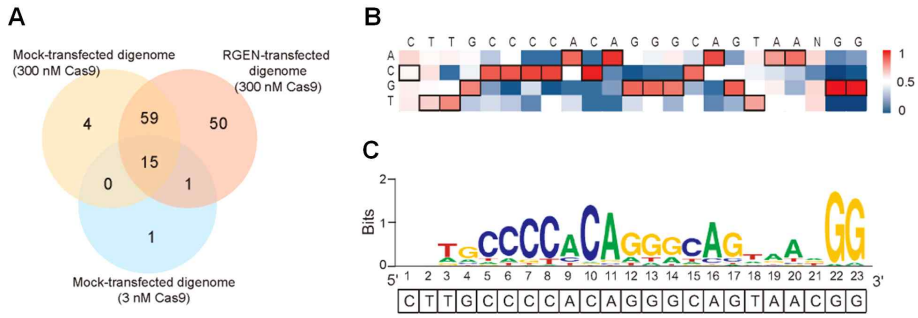
Table 2.

HBB Mock-transfected digenome												
		Forward strand				Reverse strand						
	chr.	position	count	depth	%(count/depth)	position	count	depth	%(count/depth)	gene		DNAsequence
On-target	chr11	5248215	43	48	89.6	5248214	52	57	91.2	Exonic	*HBB*	CTTGCCCCACAGGGCAGTAACGG
HBB_1	chr1	38230668	31	38	81.6	38230667	25	32	78.1	Exonic	*EPHA10*	CTCTGTCTCGCGCTGCTTTTGGG
HBB_2	chr1	177593980	34	55	61.8	177593979	55	76	72.4	Intergenic		TCTACCCCACATGGCAGTAATGG
HBB_3	chr1	191839022	12	29	41.4	191839021	10	25	40	Intergenic		CCATAGCACTCTTTAAAAAAAGC
HBB_4	chr2	91869721	15	74	20.3	91869720	15	74	20.3	Intergenic		CTTACCTCACAGGGCAGTGAGAG
HBB_5	chr2	112686732	63	85	74.1	112686731	10	32	31.2	Exonic	*MERTK*	GGTCCCGGGAATAGCGGGTAAGG
HBB_6	chr2	188075775	14	36	38.9	188075774	20	42	47.6	Intergenic		ATGCAAGTCCAATATCCAGT GGG
HBB_7	chr2	211832279	29	56	51.8	211832278	11	36	30.6	Intergenic		TAAGCCCCACACAGCAGTTAAGG
HBB_8	chr2	240591539	30	37	81.1	240591538	25	32	78.1	Intergenic		ACAGCCCCACAGGGCACTAGAGG
HBB_9	chr3	3662556	49	54	90.7	3662555	42	47	89.4	Intergenic		AAAGCCCCACAGGGTAGTAGAGG
HBB_10	chr3	19957634	38	54	70.4	19957633	38	54	70.4	Intronic	*EFHB*	GCTACCCCACAGGGCATTAGGGG
HBB_11	chr3	104567551	13	30	43.3	104567550	16	33	48.5	Intergenic		ATGAAAGTCCAATATCCAGT GGG
HBB_12	chr4	45763604	29	36	80.6	45763603	45	52	86.5	Intergenic		GCTGCCCCACATGACAGAAATGG
HBB_13	chr4	48091817	26	37	70.3	48091816	49	60	81.7	Exonic	*TXK*	ACTCGTCTCCGATATCCAGT TGG
HBB_14	chr4	55979545	52	81	64.2	55979544	30	59	50.8	Exonic	*KDR*	GGTGTAACCCGGAGTGACCAAGG
HBB_15	chr4	67338877	14	37	37.8	67338876	20	43	46.5	Intergenic		TTTGACCCACAGGGCAGTAATGG
HBB_16	chr4	125564266	14	32	43.8	125564265	10	28	35.7	Intergenic		CCCTCCCCACAGGGCAGTGAGAG
HBB_17	chr4	148531374	38	43	88.4	148531373	32	37	86.5	Intergenic		GTTACCTCACAGAGCAGAAAGGG
HBB_18	chr4	151226685	20	55	36.4	151226684	19	55	34.5	Intronic	*LRBA*	TCTGCCCCACAAGACTGTAAAGG
HBB_19	chr4	165593737	27	56	48.2	165593736	26	56	46.4	Intronic	*MIR5684*	TATGCTCCACAGGGTAGTAATGA
HBB_20	chr5	14347051	46	62	74.2	14347050	48	63	76.2	Intronic	*TRIO*	CATACCCCACAGGTCAGT-AAGG
HBB_21	chr5	26107853	13	43	30.2	26107852	13	43	30.2	Intergenic		AATACCCCACAGGGAAGTGATGG
HBB_22	chr5	131423385	38	55	69.1	131423384	25	42	59.5	Intergenic		TCTGCCCCACAGGCCAGGAAGGG
HBB_23	chr6	23709579	29	52	55.8	23709578	14	37	37.8	Intergenic		GAAGCCCTACAGGGCAGCAATGG
HBB_24	chr6	50041372	38	49	77.6	50041371	60	71	84.5	Intergenic		TCTGCCCCACATGGCAGTAATGA
HBB_25	chr6	80093919	34	51	66.7	80093918	36	53	67.9	Intergenic		TGAGTTCTCCAATATCCAGT TGG
HBB_26	chr6	85738203	51	58	87.9	85738202	29	35	82.9	Intergenic		ACTGCCCCACAGGGAAGTAATAG
HBB_27	chr6	156371508	14	34	41.2	156371507	13	33	39.4	Intergenic		CATGCTCCACAGAGCAGCAAAGG
HBB_28	chr8	41296595	37	45	82.2	41296594	40	47	85.1	Intergenic		TCAGCCCCACAGGTCAGCAATGG
HBB_29	chr8	134024458	22	50	44	134024457	16	44	36.4	Intergenic		ATTACCCCACAGGGCAGCCAAGG
HBB_30	chr9	78341070	19	36	52.8	78341069	44	62	71	Intergenic		TGTTACCCCCAGGGAAGTAT AGG
OT1	chr9	104595883	57	64	89.1	104595882	63	69	91.3	Intergenic		TCAGCCCCACAGGGCAGTAAGGG
HBB_32	chr9	134609673	26	38	68.4	134609672	28	40	70	Intronic	*RAPGEF1*	TTTGCCCCTCAGGGCAGCTAAGG
HBB_33	chr9	134994964	21	28	75	134994963	29	36	80.6	Intergenic		CCTGCCCCACAGGGCAATTATGG
HBB_34	chr10	71843328	33	38	86.8	71843327	46	52	88.5	Intronic	*H2AFY2*	CATGGCCAGGAAGAGAAGGCTGG
HBB_35	chr10	72286450	21	32	65.6	72286449	25	36	69.4	Intronic	*PALD1*	CAAGCCCCACAGGGCAGACAGGG
HBB_36	chr10	73555691	25	53	47.2	73555690	18	46	39.1	Exonic	*CDH23*	CAGGCCCCACAGGACAGGAAGGG
HBB_37	chr10	73563394	13	31	41.9	73563393	17	34	50	Intronic	*CDH23*	AGTGCCACACAGGGCAGTAT AGG
HBB_38	chr10	111569275	12	29	41.4	111569274	22	39	56.4	Intergenic		ATGCAAGTCCAATATCCAGT GGG
HBB_39	chr11	3125346	24	43	55.8	3125345	18	36	50	Intronic	*OSBPL5*	CTGGCCCCACAGGGCAGGTAGGG

HBB_40	chr11	30888494	10	37	27	30888493	11	38	28.9	Intronic	*DCDC5*	ACTTCCCCACAGGGCAGAAGTGG
HBB_41	chr11	59611432	28	34	82.4	59611431	58	64	90.6	Exonic	*GIF*	CGGCCAGATTCATGGCAATCAGG
HBB_42	chr11	76387498	22	28	78.6	76387497	36	42	85.7	Intergenic		GCTGCCCCTCAGGGAAGTAT GGG
HBB_43	chr11	104908796	11	37	29.7	104908795	12	38	31.6	Intergenic		ATGCAAGTCCAATATCCAGT AGG
HBB_44	chr11	125807920	11	42	26.2	125807919	15	46	32.6	Intergenic		ATAGCCCCATAGGGCAGAAT AGG
HBB_45	chr12	27234755	51	75	68	27234754	19	43	44.2	Intronic	*C12orf71*	GATGCCTCACAGGACAGGAAGGG
HBB_46	chr12	40327469	41	49	83.7	40327468	32	40	80	Intronic	*SLC2A13*	GCTATGGTTCCTGAACGGCCTGG
HBB_47	chr12	54421218	17	49	34.7	54421217	10	42	23.8	Intronic	*HOXC4*	TCCAGCCAGAAAGAGAAGGCTGG
HBB_48	chr12	66309907	11	42	26.2	66309906	18	49	36.7	Intronic	*HMGA2*	ATTGCCCCACGGGGCAGTGACGG
OT3	chr12	124803834	53	54	98.1	124803833	56	58	96.6	Intergenic		GCTGCCCCACAGGGCAGCAAAGG
HBB_50	chr13	44886376	44	58	75.9	44886375	48	62	77.4	Intergenic		GGAGCCCCACAGGGCAGAGAGGG
HBB_51	chr14	36889538	47	50	94	36889537	55	58	94.8	Intergenic		GTTATCCCACAGGACAGTGAGGG
HBB_52	chr14	49319039	15	49	30.6	49319038	19	53	35.8	Intergenic		ATTACCCCACAGGACAGAAATAG
HBB_53	chr14	59445901	27	50	54	59445900	21	44	47.7	Intergenic		TCTTCCCCAATATCCAGT-- AGG
HBB_54	chr14	94585327	51	52	98.1	94585326	28	29	96.6	Intergenic		ATGGCCCCACAAGGCAGAAATGG
HBB_55	chr15	29983547	42	56	75	29983546	37	51	72.5	Intergenic		CCAGCCCCACAGGGCAGTAAAGG
HBB_56	chr15	46598129	47	48	97.9	46598128	54	55	98.2	Intergenic		GTTGCCCCTCAGGACAGTACAGG
HBB_57	chr15	88488821	13	48	27.1	88488820	18	53	34	Intronic	*NTRK3*	CCTGCCCCACAGGGCAGCCAAGG
HBB_58	chr15	99709337	19	37	51.4	99709336	17	33	51.5	Intronic	*TTC23*	TGTGCCCCACAGGG-AGTGAAGG
HBB_59	chr16	49082904	27	30	90	49082903	42	46	91.3	Intergenic		GCAGCCCCACAGGTCAGTGAGGG
HBB_60	chr17	8370253	24	26	92.3	8370252	49	51	96.1	ExoniC	*NDEL1*	TTGCTCCCACAGGGCAGTAACGG
HBB_61	chr18	745994	58	59	98.3	745993	45	46	97.8	ExoniC	*YES1*	AAAATACCTCGTTGATTTCCAGG
HBB_62	chr18	6663844	23	39	59	6663843	23	39	59	Intergenic		GTTGCCCCACTGGGGAGAAAAGG
HBB_63	chr18	42330301	23	43	53.5	42330300	14	34	41.2	Intronic	*SETBP1*	ATAGCCTCACAGGGCAGAGAGGG
HBB_64	chr19	8560462	21	49	42.9	8560461	11	37	29.7	Intronic	*PRAM1*	AAATCCCCACAGGGCAGT-AAGG
HBB_65	chr19	29880768	40	52	76.9	29880767	26	38	68.4	Intronic	*LOC284395*	TGTGCCCCACAGG-CAGTAATGG
HBB_66	chr19	34262013	18	32	56.2	34262012	13	27	48.1	Intronic	*CHST8*	CTTGCTCCACAGGGCAGGTATGG
HBB_67	chr19	37539042	28	54	51.9	37539041	12	36	33.3	Intergenic		CTTGCACCACAGAGCACTAAGGG
HBB_68	chr19	50010010	14	41	34.1	50010009	15	42	35.7	Intergenic		ATTGCCCCCCAGGTCAGTAGGGG
HBB_69	chr20	17385713	12	35	34.3	17385712	15	38	39.5	Intronic	*PCSK2*	GTTGCCCC---ACGCAGTAT GGG
HBB_70	chr20	39992928	33	52	63.5	39992927	17	36	47.2	Intronic	*EMILIN3*	AGTGGCCCCCAGGGCAGTGAGGG
HBB_71	chr20	58136220	10	37	27	58136219	23	50	46	Intergenic		TTTACCCCACAGGGCATTTAAGG
HBB_72	chr22	17230623	44	48	91.7	17230622	52	56	92.9	Intergenic		TGTGCCCCACAGAGCACTAAGGG
HBB_73	chr22	26043758	16	31	51.6	26043757	11	26	42.3	Intronic	*ADRBK2*	AAAATACCTCATTAATTTCCAGG
HBB_74	chr22	35537395	31	37	83.8	35537394	32	38	84.2	Intergenic		AGTGCCCCACAGGGGAGAAATGG
HBB_75	chrX	75006257	22	33	66.7	75006256	30	41	73.2	Intergenic		GTGGCCCCACAGGGCAGGAATGG

Table2. Digenome-captured HBB off-target candidate sites.

Figure 12



D

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.

Figure 12. Off-target sites of the HBB RGEN captured by Digestome-Seq. (A) Venn diagram showing the number of Digenome-captured sites using the HBB RGEN in mock-transfected or RGEN-transfected cells. (B) Heatmap comparing digestome-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 digestome-captured sites. (D) Summary of Digesome-Seq and targeted deep sequencing. N.D., not determined.

Figure 13

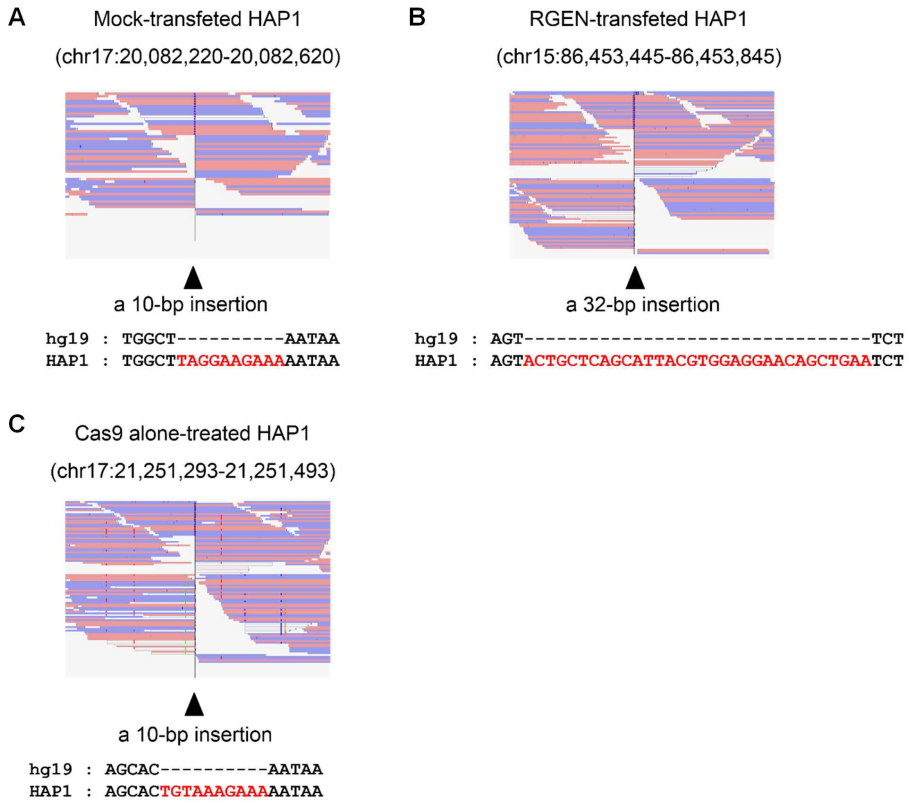


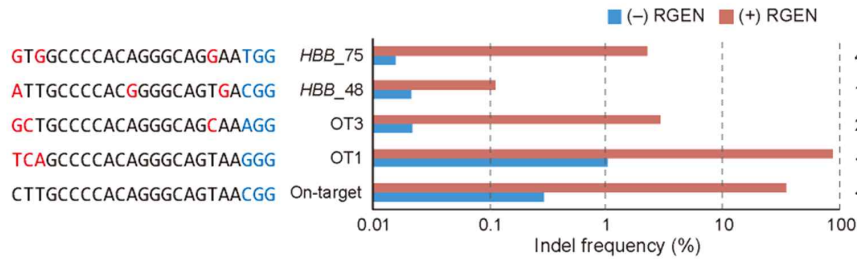
Figure 13. False-positive positions captured in the intact genome sequences. (A-C) Representative IGV images around false-positive sites that resulted from naturally-occurring indels in HAP1 cells.

4. Deep sequencing to confirm off-target effects at candidate sites

I performed deep sequencing to validate or invalidate off-target effects at the 74 common sites identified in the two independent digenomes (Figure 12D). In addition, I also tested the other 8 sites that differed from the on-target site by three nucleotides but were not captured by Digenome-Seq. No off-target indels were detected at these 8 sites with a frequency of at least 0.1% and greater than that of negative control (Fisher exact test, $p < 0.01$) (Figure 12D). Indels were observed at 5 sites including already-validated on-target, OT1, and OT3 sites, among the 74 sites, with frequencies ranging from 0.11% to 87% (Figure 14A). 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 (Figure 14B, C). These two sites differed from the on-target site by three nucleotides. These results show that Digenome-Seq is a sensitive and reproducible method to profile genome-wide nuclease off-target effects in an unbiased manner.

Figure 14

A



B

HBB_48

```
GCTATTGCCCCACGGGGCAG-TGACGGTAC   WT
GCTATTGCCCCACGGGGCAG-----GGTAC   -4
GCTATTG----------------ACGGTAC   -15
GCTATTGCCCCACGG------------TAC   -11
GCTATTGCCCCACGGGGCAGTTGACGGTAC   +1
GCTATTGCCCCACGGGGCAGATGACGGTAC   +1
```

C

HBB_75

```
GTTGTGGCCCCACAGGGCAG-GAATGGCAGCG   WT
GTTGTGGCCCCACAGGGCAG----------CG   -9
GTTGTGGCCCCACAGGGCAGGGAATGGCAGCG   +1
GTTGTGGCCCCACAGGGCAG--AATGGCAGCG   -1
GTTGTGGC--------------------AGCG   -19
GTTGTGG---------------AATGGCAGCG   -14
```

Figure 14. Off-target sites of the HBB RGEN validated by targeted deep sequencing. (A) 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). (B-C) Examples of deep sequencing result of new validated HBB off-target region. Inserted nucleotides are shown in red and the PAM sequence is shown in blue.

5. Digenome sequencing with another 'promiscuous' RGEN

I performed Digenome-Seq with another RGEN that had been shown to induce on-target mutations at a VEGF-A locus and off-target mutations at four homologous sites. A total of 81 sites, including the on-target and four validated off-target sites, were captured that showed double-peak patterns (Figure 15A). 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 (Figure 15B). I 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 (Figure 15B, C). I 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 by 3 or fewer nucleotides from the on-target site but were not captured by Digenome-Seq (Figure 15D). This RGEN was highly active in HAP1 cells, producing indels at the target site with a frequency of 87% and at the four previously-validated off-target sites with frequencies that ranged from 0.32% to 79%. In addition, four new off-target sites were identified at which indels were induced with frequencies that ranged from 0.065% to 6.4% (Figure 16A-E). The indel frequency at each of these sites obtained using the RGEN was significantly greater than that obtained using an empty vector control

(Fisher exact test, $p < 0.01$). These off-target sites contained one to six 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 6 sites (0.043%) were captured by digenome sequencing and, among them, only one site was validated by deep sequencing (Figure 15D). To the best of our knowledge, an RGEN off-target site with 6-nt mismatches had never previously been identified. None of these validated off-target sites contained a DNA or RNA bulge, although 40 out of 81 sites captured by Digesome-Seq contained a missing or extra nucleotide compared to the 20-nt target sequence. At all the other sites, including those not captured by Digesome-Seq, indel frequencies obtained using the RGEN were below 0.05% or were smaller than or not statistically different from those obtained using an empty vector control.

Table 3.

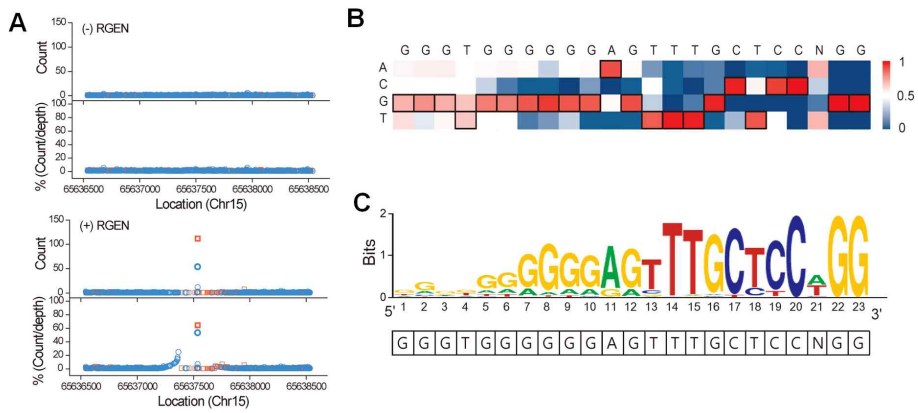
VEGF-A Mock-transfected digenome												
		Forward strand				Reverse strand						DNAsequence
	chr.	position	count	depth	%(count/depth)	position	count	depth	%(count/depth)	gene		GGGTGGGGGGAGTTTGCTCCTGG
VEGFA_1	chr01	42740063	14	59	23.7	42740062	19	60	31.7	Intronic	*FOXJ3*	AGTTGAGGGGAGTTTA-TCCTGG
VEGFA_2	chr01	48314379	14	44	31.8	48314378	22	57	38.6	Intronic	*TRABD2B*	GGGCAGGGAAAGTTTG-TCCAGG
VEGFA_3	chr01	99347651	35	66	53	99347650	20	50	40	Intergenic		GGGGAGGGGAAGTTTGCTCCTGG
VEGFA_4	chr01	102122651	11	46	23.9	102122650	12	46	26.1	Intergenic		TGCTGAGGGGAATTTGC-CCAGG
VEGFA_5	chr01	233157354	22	58	37.9	233157353	35	71	49.3	Intronic	*PCNXL2*	GGAGGAGGGGAGTCTGCTCCAGG
VEGFA_6	chr02	66561098	16	29	55.2	66561097	47	68	69.1	Intergenic		GGGTGG-GGCAGTTTG-TCCGGG
VEGFA_7	chr02	118827653	12	52	23.1	118827652	15	50	30	Intergenic		AGATGAGGGGAGTTAG-CCCTGG
VEGFA_8	chr02	157256493	18	53	34	157256492	11	43	25.6	Intergenic		GGGGCAGGGGA-CTTGCTCCAGG
VEGFA_9	chr02	205137461	49	89	55.1	205137460	46	77	59.7	Intergenic		GGCTAGAGGGAGTTTG-CCCTGG
VEGFA_10	chr02	209437600	13	56	23.2	209437599	48	94	51.1	Intergenic		AGGGAGGGAGAATTTGCTCCTGG
VEGFA_11	chr03	8753103	23	60	38.3	8753102	46	81	56.8	Intergenic		TTTTGGAGGGAATTTG-CTCCGG
VEGFA_12	chr03	38316671	26	64	40.6	38316670	17	50	34	Intergenic		AACCTGGGGGGGCTTGCTCCTGG
VEGFA_13	chr03	55562940	28	59	47.5	55562939	18	50	36	Intronic	*SLC22A13*	TATGGGGATGGGGTTGCTCCCGG
VEGFA_14	chr03	97310666	28	54	51.9	97310665	63	96	65.6	Intronic	*EPHA6*	GTGTGGGAAGAGTTTG-TCCTGG
VEGFA_15	chr03	128284321	32	69	46.4	128284320	28	65	43.1	Intergenic		AGGTGGTGGGAGCTTGTTCCTGG
VEGFA_16	chr04	8453803	23	69	33.3	8453802	27	74	36.5	Intronic	*TRMT44*	GAGTGGGTGGAGTTTGCTACAGG
VEGFA_17	chr04	21369324	28	59	47.5	21369323	63	111	56.8	Intronic	*KCNIP4*	TCCATGGGGGAGTTTG-CTCTGG
VEGFA_18	chr04	23116018	16	51	31.4	23116017	17	53	32.1	Intergenic		CAGGGGGAGAGATTTGCTCCAGG
VEGFA_19	chr04	54967137	52	104	50	54967136	15	60	25	Intronic	*GSX2*	GGAGGTGGAAAGTTTGCTCCAGG
VEGFA_20	chr04	65046699	15	47	31.9	65046698	29	60	48.3	Intergenic		GGGTAAGGGAAGTTTG-CTCTGG
VEGFA_21	chr04	65941717	23	51	45.1	65941716	17	43	39.5	Intergenic		GGCTGGTGGGAGTTT-CTCCAGG
VEGFA_22	chr04	157417951	25	66	37.9	157417950	26	70	37.1	Intergenic		TGATGGGGAAAGTTTGC-TCAGG
VEGFA_23	chr04	190326147	10	35	28.6	190326146	10	38	26.3	Intergenic		ACAGTCGCCCCTTTGTTTTGGCC
VEGFA_24	chr05	156390	16	50	32	156389	13	51	25.5	Intronic	*PLEKHG4B*	TGCTCGGGGGAGTTTGCACCAGG
VEGFA_25	chr05	7067159	23	44	52.3	7067158	47	70	67.1	Intergenic		GAGGGTGGGGAGTTTACTCCTGG
VEGFA_26	chr05	32945275	25	55	45.5	32945274	53	89	59.6	Intergenic		GCGTGGGGGGTGTTTGCTCCCGG
VEGFA_27	chr05	56172079	23	68	33.8	56172078	12	56	21.4	Intronic	*MAP3K1*	GGTGGGGGTGGGTTTGCTCCTGG
VEGFA_28	chr05	57030872	22	54	40.7	57030871	26	57	45.6	Intergenic		TCTG-AGGGGAGTTTG-CTCTGG
VEGFA_29	chr05	139263024	19	59	32.2	139263023	36	86	41.9	Intergenic		TTGGGGGGGCAGTTTGCTCCTGG
VEGFA_30	chr05	140642769	17	53	32.1	140642768	22	58	37.9	Intergenic		CAAGGTGGGAGGATTGCTCCAGG
VEGFA_31	chr05	146236270	11	55	20	146236269	26	69	37.7	Intronic	*PPP2R2B*	TCTTTGGGAGAGTTTGCTCCAGG
VEGFA_32	chr06	42229599	51	94	54.3	42229598	41	81	50.6	Intronic	*TRERF1*	GGGGCAAGGGAGTTTG-CTCAGG
On-target	chr06	43737297	28	53	52.8	43737296	18	37	48.6	Exonic	*VEGFA*	GGGTGGGGGGAGTTTGCTCCTGG
VEGFA_34	chr06	50485682	23	64	35.9	50485681	42	88	47.7	Intergenic		ATGTGTGGGGAATTTGCTCCAGG
VEGFA_35	chr06	91365255	40	72	55.6	91365254	16	43	37.2	Intergenic		CCCGGGGGGAAGCTTGCTCCAGG
VEGFA_36	chr06	103251063	14	50	28	103251062	25	60	41.7	Intergenic		TCCATGGGG-GATTTGCTCCAGG
VEGFA_37	chr06	115347691	13	39	33.3	115347690	10	27	37	Intergenic		AACCACAGCAGTGCAGGACATCA
VEGFA_38	chr07	17819097	15	47	31.9	17819096	39	76	51.3	Intergenic		ACAACTGGGGAGTTTGCTCCTGG
VEGFA_39	chr07	31001913	15	46	32.6	31001912	10	41	24.4	Intergenic		TGGGGTGGTGAGTTG-CTCCTGG
VEGFA_40	chr07	32333724	34	56	60.7	32333723	33	61	54.1	Intronic	*PDE1C*	GAAGGGAGGAGGTTTGCTCCTGG

VEGFA_41	chr07	73254136	11	38	28.9	73254135	21	53	39.6	Intronic	*WBSCR27*	GGA-GGGTGGAGTTG-CTCCTGG
VEGFA_42	chr08	74709100	55	100	55	74709099	16	55	29.1	Intronic	*UBE2W*	GG--GGGGTGAGTTTG-TCCTGG
VEGFA_43	chr08	128303251	20	57	35.1	128303250	15	47	31.9	Intergenic		TCTTGGGGAGAGTTTGC-CCAGG
VEGFA_44	chr09	9028824	18	55	32.7	9028823	38	77	49.4	Intronic	*PTPRD*	TGACTGGGGAAGTTTGC-CCAGG
VEGFA_45	chr09	91304357	11	39	28.2	91304356	13	30	43.3	Intergenic		GGATAGTTCCATTATGACTGCCC
VEGFA_46	chr09	93925190	17	65	26.2	93925189	13	59	22	Intergenic		GGGGGTGGGGAGCATGCTCCAGG
VEGFA_47	chr09	122768852	21	62	33.9	122768851	17	58	29.3	Intergenic		TCTGGAGGA-AGTTTGC-CCAGG
VEGFA_48	chr09	123154231	11	43	25.6	123154230	15	49	30.6	Intronic	*CDK5RAP2*	AGGGTGAGGGG-CTTGCTCCAGG
VEGFA_49	chr10	25067928	49	76	64.5	25067927	17	40	42.5	Intergenic		GGGGGGAAGGAGTTTC-TCCTGG
VEGFA_50	chr10	81363268	23	52	44.2	81363267	24	46	52.2	Intergenic		CACTGAGGGGAGTTTGC-CCAGG
VEGFA_51	chr10	99376528	26	73	35.6	99376527	22	71	31	Exonic	*MORN4*	TATGAGGGGGAGTTTGC-CCAGG
VEGFA_52	chr10	132171413	20	52	38.5	132171412	35	61	57.4	Intergenic		CCCTGGGGA-AGTTTGT-CCAGG
VEGFA_53	chr11	17395396	62	87	71.3	17395395	20	30	66.7	Exonic	*NCR3LG1*	GGGGAGGCGGAGTTTG-TCCTGG
VEGFA_54	chr11	57835243	18	49	36.7	57835242	20	55	36.4	Intronic	*OR9Q1*	GAGGTGGGGTGATTTGCTCCAGG
VEGFA_55	chr11	72475898	36	62	58.1	72475897	14	39	35.9	Intronic	*STARD10*	TTCCAGGGGGAGTTTC-TCCGGG
VEGFA_56	chr11	117481208	45	75	60	117481207	11	37	29.7	Intronic	*DSCAML1*	GGGCAAGGGGAGGTTGCTCCTGG
VEGFA_57	chr11	126678277	10	47	21.3	126678276	11	45	24.4	Intronic	*KIRREL3*	GGGGAGGGGGAGTTAG-CCCTGG
VEGFA_58	chr12	1988077	25	49	51	1988076	39	72	54.2	Intronic	*CACNA2D4*	CGGGGGAGGGAGTTTGCTCCTGG
VEGFA_59	chr12	5619395	40	58	69	5619394	15	23	65.2	Intergenic		CGGAGGGGTGAGTTTG-TCCCGG
VEGFA_60	chr12	26641302	20	57	35.1	26641301	20	54	37	Intronic	*ITPR2*	AGTTTGGGGGAGTTTGCCCCAGG
VEGFA_61	chr12	51888011	11	35	31.4	51888010	49	82	59.8	Intronic	*SLC4A8*	AATAGTG--GGAGTTGCTCCTGG
VEGFA_62	chr12	107832636	24	60	40	107832635	35	79	44.3	Intronic	*BTBD11*	TCTTGGGGGGAAGTTGCTCCAGG
VEGFA_63	chr12	131690199	12	53	22.6	131690198	45	93	48.4	Intronic	*LOC116437*	GGGAGGGTGGAGTTTGCTCCTGG
VEGFA_64	chr13	31251013	28	59	47.5	31251012	34	62	54.8	Intergenic		TGTAGAGGGAGTTTTGCTCCCGG
VEGFA_65	chr14	30157649	10	42	23.8	30157648	32	68	47.1	Intronic	*MIR548AI*	TGGGGGCGGGAGTTT-CTCCTGG
VEGFA_66	chr14	61906429	17	51	33.3	61906428	31	71	43.7	Intronic	*PRKCH*	TT---GGGGGAGTTTG-CTCAGG
VEGFA_67	chr15	65637537	112	174	64.4	65637536	54	101	53.5	Intronic	*IGDCC3*	GGATGGAGGGAGTTTGCTCCTGG
VEGFA_68	chr15	67385016	42	103	40.8	67385015	62	137	45.3	Intronic	*SMAD3*	TGGTGGAGGGAGTTTG-TCCAGG
VEGFA_69	chr15	78857354	22	92	23.9	78857353	23	93	24.7	Intergenic		AGTGGTGGGGGACTTGCTCCAGG
VEGFA_70	chr15	84047385	32	107	29.9	84047384	89	179	49.7	Intergenic		GGAGTCAGGGAATTTGCTCCTGG
VEGFA_71	chr15	86453645	12	53	22.6	86453644	16	60	26.7	Intergenic		CAGAGTTCTTTTGGCATCCCAGT
VEGFA_72	chr16	8763213	36	86	41.9	8763212	11	47	23.4	Intergenic		AAGTAAGGGAAGTTTGCTCCTGG
VEGFA_73	chr17	3830416	23	63	36.5	3830415	14	61	23	Intronic	*ATP2A3*	GAGTGAGGGGGAGTTTCTCCAGG
VEGFA_74	chr17	21251093	14	56	25	21251092	14	47	29.8	Intergenic		AGCACAATAAACCAGCGAAATGG
VEGFA_75	chr17	32986325	44	73	60.3	32986324	37	72	51.4	Intergenic		GGGGGTGGGGACTTTGCTCCAGG
VEGFA_76	chr17	39796328	36	62	58.1	39796327	32	58	55.2	Exonic	*KRT42P*	TAGTGGAGGGAGCTTGCTCCTGG
VEGFA_77	chr17	66834243	28	40	70	66834242	43	56	76.8	Intergenic		GGAAGGGGGGAGTTTG-TCCTGG
VEGFA_78	chr17	80625576	41	77	53.2	80625575	40	83	48.2	Intronic	*RAB40B*	TGACGGGGGGAGTTTG-CTCTGG
VEGFA_79	chr18	366714	38	73	52.1	366713	24	54	44.4	Intronic	*COLEC12*	GGGGGCAGGGAGATTGCTCCTGG
VEGFA_80	chr19	1177850	13	61	21.3	1177849	19	68	27.9	Intergenic		AAGTGGGGGCAGTTTGC-CCAGG
VEGFA_81	chr19	8150194	21	76	27.6	8150193	16	69	23.2	Intronic	*FBN3*	GAAACGGGGGAGTTTG-CTCAGG
VEGFA_82	chr20	60265710	18	45	40	60265709	25	52	48.1	Intronic	*CDH4*	GGACGGGGGGAGTTT-CTCCAGG
VEGFA_83	chr22	18454623	12	49	24.5	18454622	23	61	37.7	Intronic	*MICAL3*	GGAAAGGAGGAGCTTGCTCCAGG

VEGFA_84	chr22	37215276	42	70	60	37215275	46	75	61.3	Intronic	*PVALB*	GGGTGGGGGGAGTTTGCCCCAGG
VEGFA_85	chrX	19185601	30	55	54.5	19185600	12	32	37.5	Intergenic		GGGAGGGGAGAGTTTGTTCCAGG
VEGFA_86	chrX	128497221	25	44	56.8	128497220	16	31	51.6	Intergenic		AGGGGTAGGGAGATTGCTCCTGG

Table3. Digenome-captured VEGFA off-target candidate sites.

Figure 15



D

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	1	1	1	0	0
2	4	3	3	1	0
3	32	5	2	27	0
4	283	12	1	271	N.D.
5	2,176	7	0	2,169	N.D.
6	13,892	6	1	13,886	N.D.

Figure 15. Off-target sites of the VEGF-A RGEN captured by Digenome-seq. (A) 5' End plots at the one of VEGF-A off-target site. (B) Heatmap 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.

Figure 16

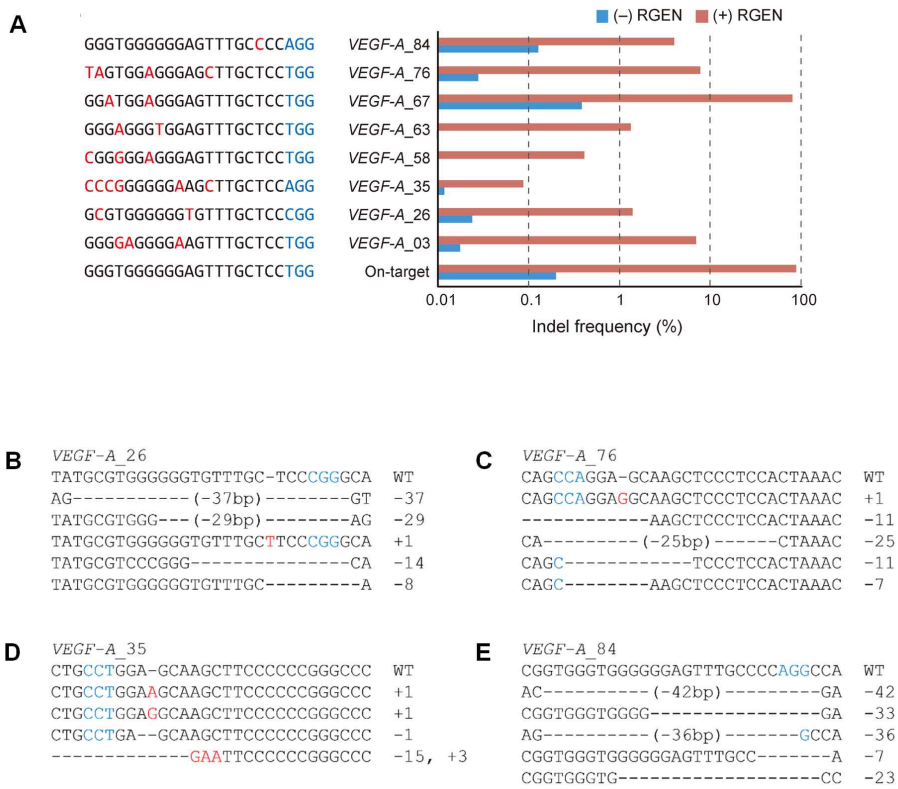


Figure 16. Validation of off-target sites captured by Digenome-seq using VEGFA targeting RGEN. (A) Off-target sites validated by targeted deep sequencing. Blue and red bars represent indel frequencies obtained using mock-transfected HAP1 cells and the VEGF-A RGEN-transfected HAP1 cells, respectively. (Left) DNA sequences of on-target and off-target sites. Mismatched bases are shown in red. The PAM is shown in blue. (B-E) New validated Off-target sites were detected by targeted deep sequencing. Inserted nucleotides are shown in red and the PAM sequence is shown in blue.

6. Avoiding RGEN off-target effects via modified sgRNAs

sgRNAs with two extra guanine nucleotides at the 5' terminus (termed ggX20 sgRNAs) can efficiently discriminate on-target sites from homologous sites that differ by two or more nucleotides, reducing off-target effects by orders of magnitude without sacrificing on-target effects. I replaced the two 'promiscuous' gX19 (HBB) and GX19 (VEGF-A) sgRNAs with respective ggX20 sgRNAs ('g' and 'G' represent a mismatched guanine and matched guanine, respectively.) (Figure 17A) and measured on-target and off-target mutation frequencies in HAP1 and K562 cells at the sites identified by Digesome-Seq and validated by targeted deep sequencing. Strikingly, indels were barely detectable above noise levels (i.e. deep sequencing error) at the several validated off-target sites in the two genes when I used the ggX20 sgRNAs (Figure 17B-E). In contrast, these sgRNAs were almost equally active, compared to the gX19 and GX19 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. Based on the specificity ratio of on- to off-target indel frequencies, the two modified sgRNAs showed up to 660-fold greater specificity than the conventional sgRNAs (Table 4).

Figure 17

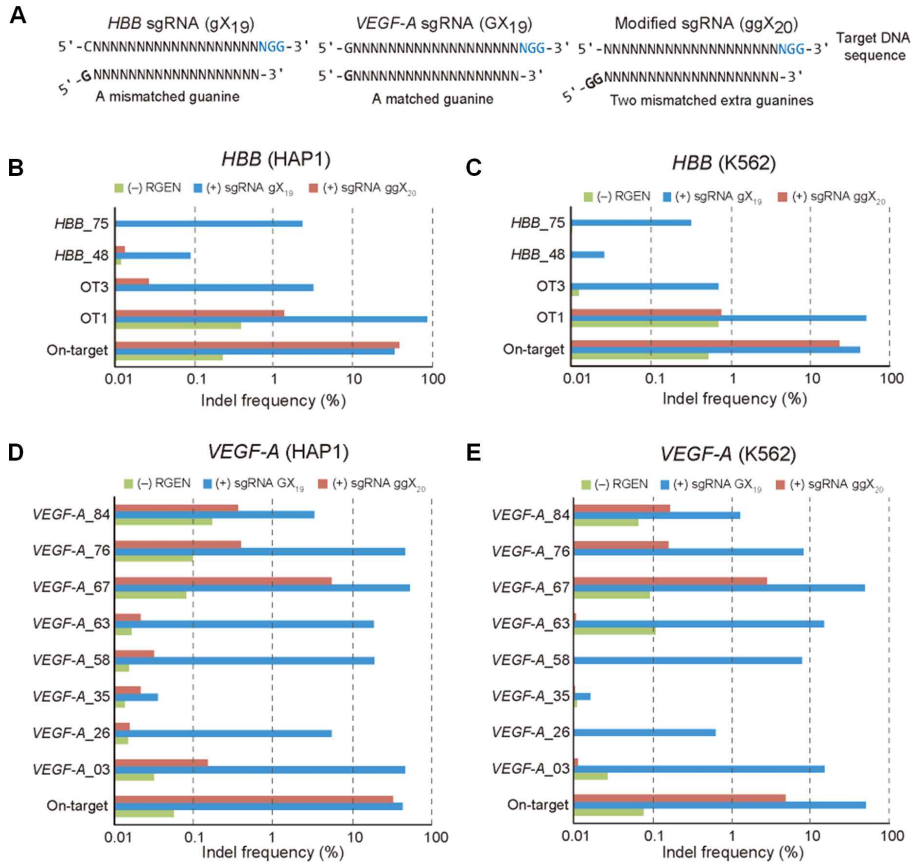

A
HBB sgRNA (gX19)
VEGF-A sgRNA (GX19)
Modified sgRNA (ggX20)
Target DNA sequence
A mismatched guanine
A matched guanine
Two mismatched extra guanines
B
HBB (HAP1)
C
HBB (K562)
D
VEGF-A (HAP1)
E
VEGF-A (K562)
(–) RGEN
(+) sgRNA gX19
(+) sgRNA ggX20
HBB_75
HBB_48
OT3
OT1
On-target
VEGF-A_84
VEGF-A_76
VEGF-A_67
VEGF-A_63
VEGF-A_58
VEGF-A_35
VEGF-A_26
VEGF-A_03
0.01
0.1
1
10
100
Indel frequency (%)

Figure 17. Comparison of conventional sgRNAs with modified sgRNAs that include two extra guanine nucleotides. (A) Schematic of gX19 sgRNA, GX19 sgRNA, and ggX20 sgRNA. (B and C) Indel frequencies at the HBB on-target and the four validated off-target sites in HAP1 cells (B) and K562 cells (C) were measured via targeted deep sequencing. N.A., not applicable when indel frequencies with RGENs were lower than those with negative control. P value was calculated by Fisher exact test. (D and E) Indel frequencies at the VEGF-A on-target and the eight validated off-target sites in HAP1 cells (D) and K562 cells (E) were measured via targeted deep sequencing.

Table 4.

	Specificity ration (on/off)					
	HAP1			K562		
	gX19	ggX20	ggX20/gX19	gX19	ggX20	ggX20/gX19
OT1	0.39	28	72	0.84	30	36
OT3	11	1400	130	59	2700	45
HBB_48	380	2900	7.6	1600	4900	3
HBB_75	14	4100	290	130	2300	18

	Specificity ration (on/off)					
	HAP1			K562		
	GX19	ggX20	ggX20/GX19	GX19	ggX20	ggX20/GX19
VEGF-A_03	0.94	210	230	3.3	430	130
VEGF-A_26	7.9	2100	270	79	590	7.5
VEGF-A_35	1200	1500	1.3	3000	460	0.15
VEGF-A_58	2.3	1000	460	6.3	580	91
VEGF-A_63	2.3	1500	660	3.4	440	130
VEGF-A_67	0.81	5.9	7.3	1	1.7	1.7
VEGF-A_76	0.94	84	89	6.2	30	4.8
VEGF-A_84	13	90	6.9	39	29	0.74

Table 4. Comparison specificity ratio of conventional sgRNAs with modified sgRNAs. Specificity ratio calculated by dividing indel frequencies at the on-target site with those at an off-target site in HBB and VEGF-A.

C. Multiplex Digenome-seq for genome-wide target specificities of RGEN

1. Improving Digenome-seq

First, I 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. I found that these sites could be identified by assuming that Cas9 can produce one or two-nucleotide overhangs in addition to blunt ends. I assigned a DNA cleavage score to each nt position, based on patterns of alignments of sequence reads (Figure 18). This 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 (Figure 19): 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.

I 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 (nt) compared to the on-target site (Lin et al., 2014), which were captured with those

transcribed using an oligonucleotide duplex (Figure 20A). 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, as shown by sequence logos obtained computationally by comparing DNA sequences around cleavage sites with each other (Figure 20B). 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.

Figure 18

Score at position i =

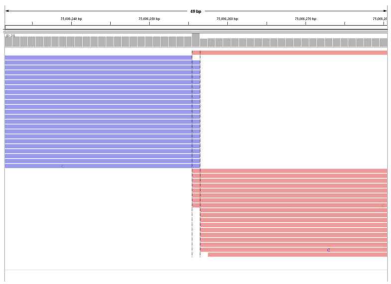
$$\sum_{a=1}^{5} \frac{(F_i - 1)}{D_i} X \frac{(R_{i-4+a} - 1)}{D_{i-4+a}} \times (F_i + R_{i-4+a} - 2) + \sum_{a=1}^{5} \frac{(R_{i-1} - 1)}{D_{i-1}} X \frac{(F_{i-3+a} - 1)}{D_{i-3+a}} \times (R_{i-1} + F_{i-3+a} - 2)$$

F_i : Number of forward sequence reads starting at position i

R_i : Number of reverse sequence reads starting at position i

D_i : Sequencing depth at position i

Example :



		75006253	75006254	75006255	75006256	75006257	75006258	75006259
Count	Reverse	0	0	1	22	0	0	0
	Forward	0	0	0	9	9	1	0
Depth		23	23	23	31	18	19	19

Score at position 75006256

= (22-1)/31 * [(0-1)/23 * (22+0-2) + (9-1)/31 * (22+9-2) + (9-1)/18 * (22+9-2) + (1-1)/19 * (22+1-2) + (0-1)/19 * (22+0-2)]

+ (9-1)/18 * [(0-1)/23 * (9+0-2) + (1-1)/23 * (9+1-2) + (22-1)/31 * (9+22-2) + (0-1)/18 * (9+0-2) + (0-1)/19 * (9+0-2)]

= 20.66

Figure 18. *in vitro* DNA cleavage scoring system for Digenome-seq analysis. I assigned a DNA cleavage score to each nucleotide position across the human genome using the following equation and set a cutoff value of 2.5 to identify *in vitro* DNA cleavage sites.

Figure 19.

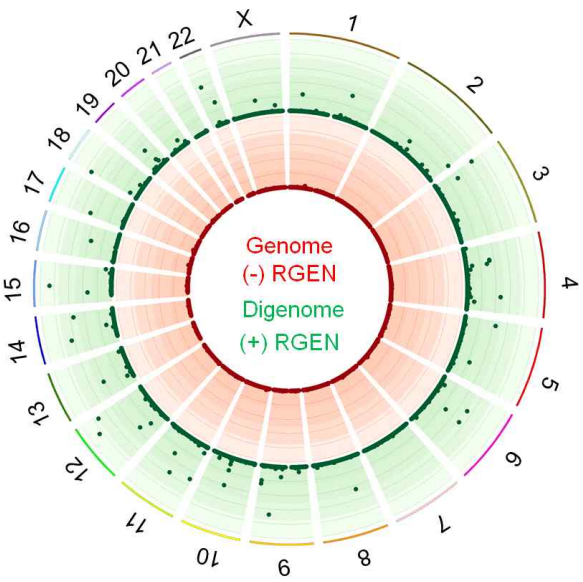


Figure 19. Scoring system of Digenome-seq analysis. Genome-wide Circos plots of *in vitro* DNA cleavage scores. Human genomic DNA (red) or RGEN-digested genomic DNA (green) was subjected to whole genome sequencing.

Figure 20.

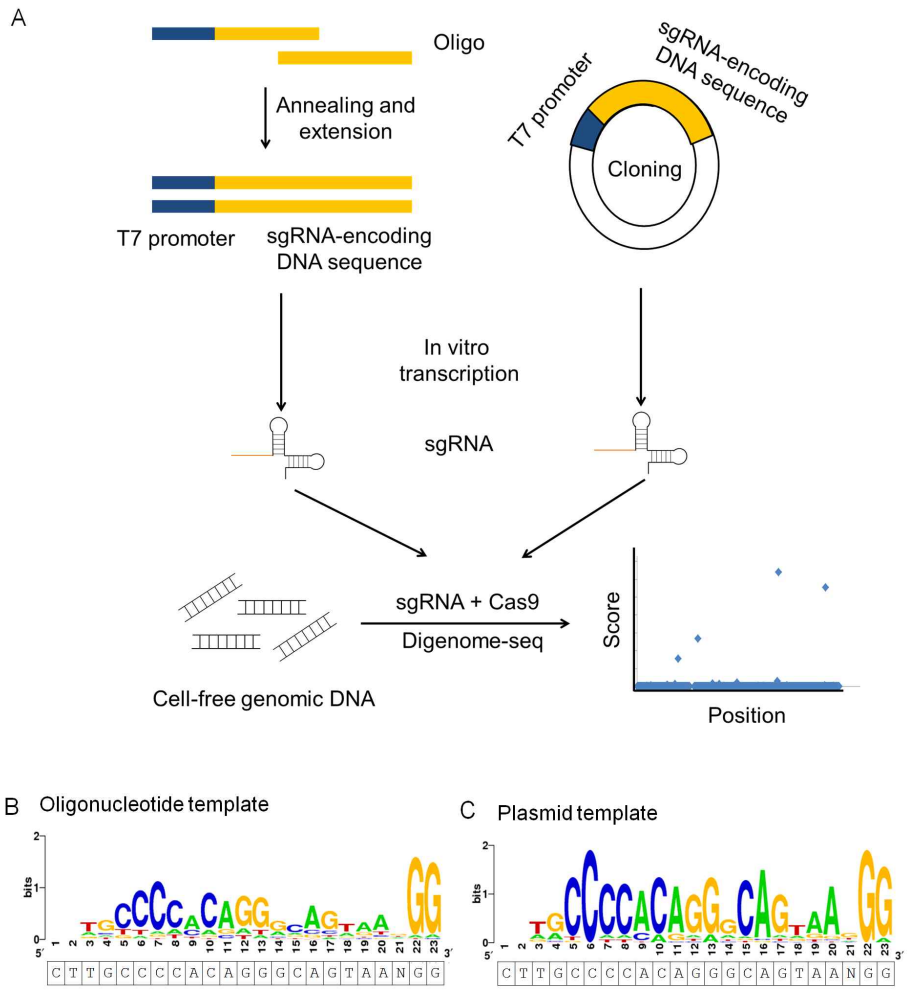

A
Oligo
Annealing and extension
T7 promoter
sgRNA-encoding DNA sequence
T7 promoter
sgRNA-encoding DNA sequence
Cloning
In vitro transcription
sgRNA
sgRNA + Cas9
Digenome-seq
Cell-free genomic DNA
Score
Position
B Oligonucleotide template
C Plasmid template
bits
C T T G C C C C A C A G G G C A G T A A N G G

Figure 20. Comparison of Digenome-seq using sgRNA transcribed from an oligonucleotide duplex or a plasmid. (A) Schematic overview of Digenome-seq using sgRNA transcribed from an oligonucleotide duplex or a plasmid. (B-C) Sequence logos obtained using sgRNA transcribed from an oligonucleotide duplex (B) or a plasmid (C).

2. Multiplex Digenome-seq

Unlike other methods, Digenome-seq can be multiplexed without increasing sequencing depth proportionally to the number of nucleases. I 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. I digested human genomic DNA with a mixture of the Cas9 protein, the 10 sgRNAs, and one additional sgRNA targeted to the HBB gene, which I had analyzed in our previous study (Kim et al., 2015), and carried out two independent WGS analyses (Figure 21). Genome-wide *in vitro* cleavage sites were identified computationally using the scoring system. A total of 964 sites were found in the human genome. All of these sites were then classified computationally according to the edit distance from the on-target sites (Figure 21).

Figure 21.

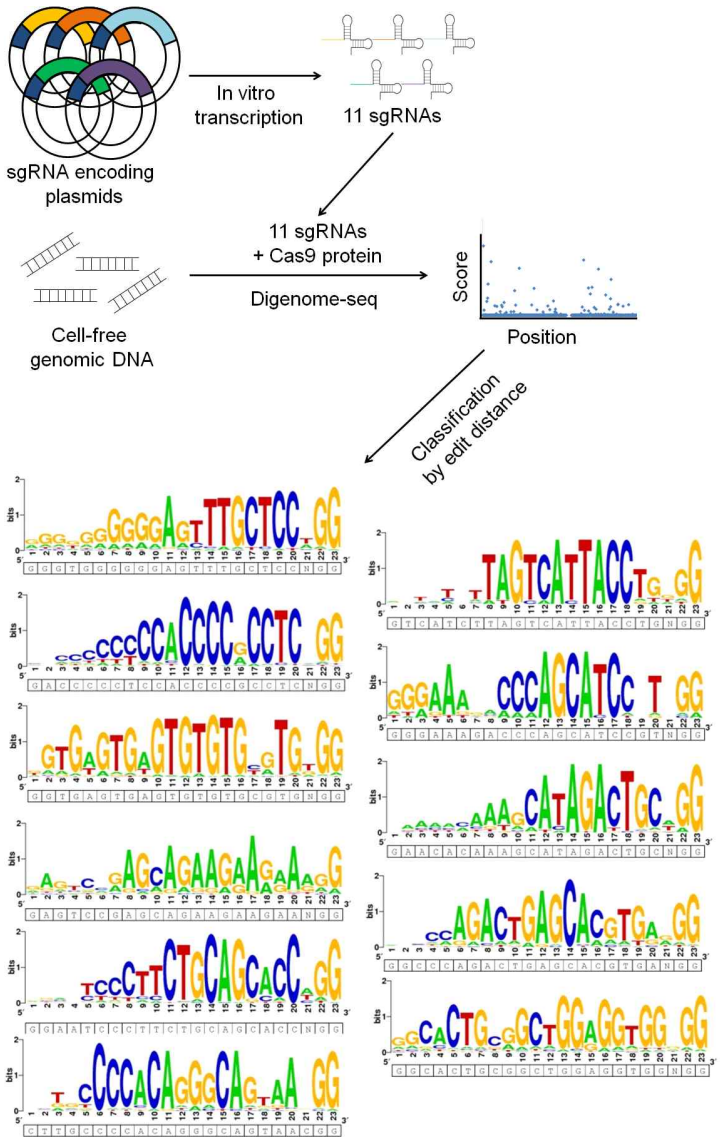


Figure 21. Multiplex Digenome-seq. Schematic overview of multiplex Digenome-seq.

First, I checked whether an sgRNA in a pool can cleave its on-target and off-target sites. 17 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) (Figure 22A and B). Note that more sites are captured at higher concentration of sgRNAs. Importantly, all the four off-target sites as well as the on-target site that had been validated using targeted deep sequencing in our previous study were identified by multiplex Digenome-seq. This result suggests that each sgRNA in a pool of up to 11 sgRNs can guide Cas9 to most of its on-target and off-target sites independently from each other, supporting the basis of multiplexing.

Figure 22.

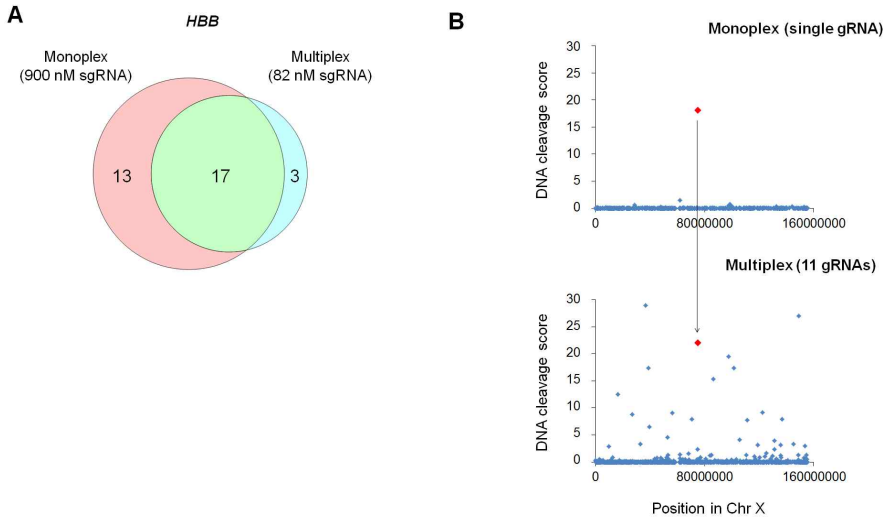


Figure 22. Multiplex Digenome-seq. (A) A Venn diagram showing the number of *in vitro* cleavage sites captured by monoplex and muliplex Digenom-seq analyses. (B) *in vitro* DNA cleavage scores across Chromosome X obtained by monoplex or multiplex digenome seq.

3. *In vitro* cleavage sites

The 11 sgRNAs showed a wide spectrum of genome-wide specificities : Then umber of cleavage sites per sgRNA in the human genome ranged from 13 to 302 (Figure 23). As expected, all of the 11 on-target sites and most of the sites with 1 or 2 mismatches, identified in the human genome using Cas-OFFinder (Bae et al., 2014), were captured by Digenome-seq (Figure 24A). However, sites with more than 3 mismatches were rarely captured. The fraction of Digenome-captured sites decreased exponentially as the number of mismatches increased from 3 to 6 (Figure 24A). I also found that sites with 2 or more mismatches in the seed region were much less likely to be cleaved *in vitro* than those with 0 or 1 mismatch in the seed ($P < 0.01$, Student's t-test) (Figure 24B).

Interestingly, I found a strong correlation (R^2=0.93) between the number of Digenome-captured sites and the number of homologous sites with 6 or fewer nt mismatches in the human genome (Figure 24C). The 6 sgRNAs with fewer than 13,000 such homologous sites in the human genome were much more specific ($P < 0.01$, Student's t-test), cleaving 46 or fewer sites *in vitro* (28 sites/sgRNA, on average), than the other 5 sgRNAs with more than 16,000 such sites, cleaving 63 or more sites *in vitro* (161 sites/sgRNA, on average) (Figure 24C and D). 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 (Figure

25A) (Tsai et al., 2015). The VEGFA 2 site was an outlier, at least partially causing the poor correlation. I noted, however, that the 5 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.

Figure 23.

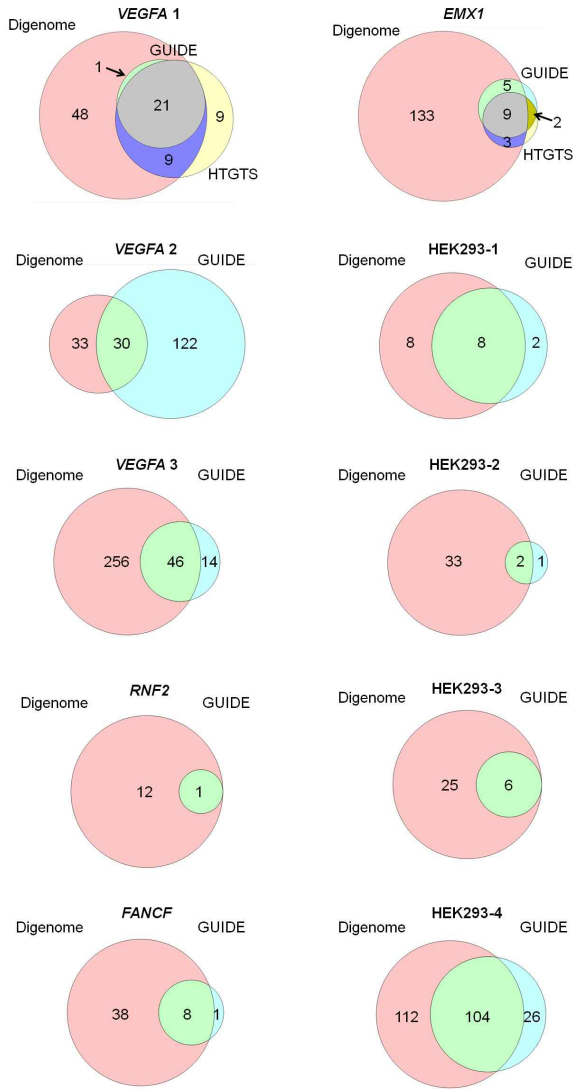


Figure 23. Comparison of the Digenome-seq, GUIDE-seq, and HTGTS. Venn diagrams showing the number of sites captured by Digenome-seq, GUIDE-seq, and HTGTS.

Figure 24.

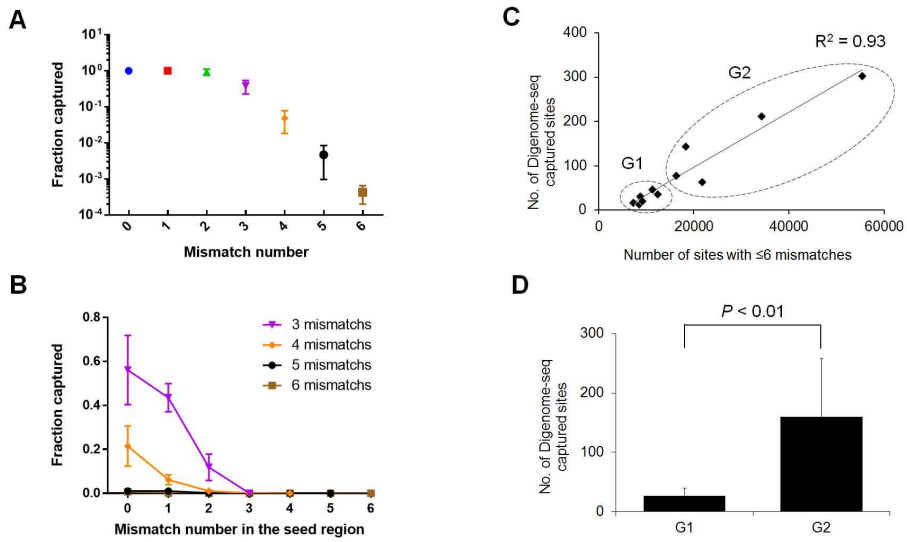


Figure 24. Analysis of multiplex Digenome-captured sites. (A-B) Fractions of sites captured by Digenome-seq according to the total mismatch number (A) and the mismatch number in the seed region (B). (C-D) Scatterplot of the number of sites with 6 or fewer mismatches in the human genome vs. the number of Digenome-captured sites (top). 11 RGEN target sites were divided into two groups, G1 and G2 (those with fewer than 13,000 and 16,000 sites, respectively, harboring 6 or fewer mismatches in the human genome) (bottom). Error bars represent SEM. The P value was calculated by Student's t-test.

Figure 25.

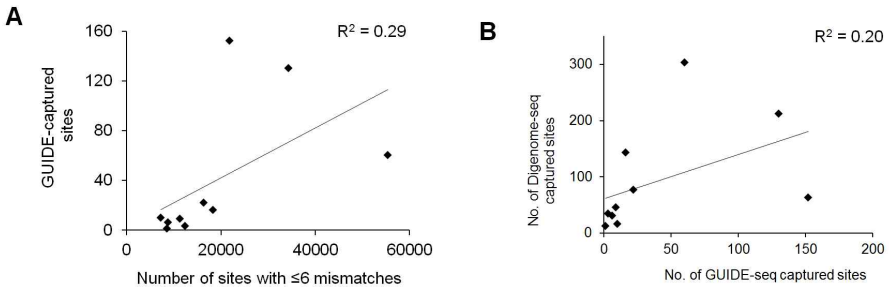


Figure 25. Comparison of the Digenome-seq and GUIDE-seq. (A) Poor correlation between the number of GUIDE-seq positive sites and the number of homologous sites with 6 or fewer mismatches in the human genome. (B) Scatterplot of the number of GUIDE-captured sites vs. the number of Digenome-captured sites.

4. Digenome-seq vs. other methods

On average, multiplex Digenome-seq successfully identified 80.8% of sites captured previously by GUIDE-seq (Figure 23). For example, all of the GUIDE-captured sites using the three sgRNAs specific to the VEGFA 1, 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 (Figure 23). 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, I observed a poor correlation (R^2 = 0.20) between the number of Digenome-positive sites and that of GUIDE-positive sites (Figure 25B). It is likely that many additional sites those are cleaved *in vitro* and, thereby, captured by Digenome-seq are not accessible in cells, owing to chromatin.

Digenome-seq yielded more candidate off-target sites than GUIDE-seq for 9 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 VEGFA 1 and EMX1 sites had also been analyzed by HTGTS (Figure 23). Most of sites (31 out of 40 sites for VEGFA 1 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 9 and 2 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 (Figure 26) or the low concentration (82 nM) of the sgRNA. These problems could be alleviated by performing WGS at a higher sequencing depth or using a higher concentration of the sgRNA in a monoplex analysis, respectively.

Figure 26.

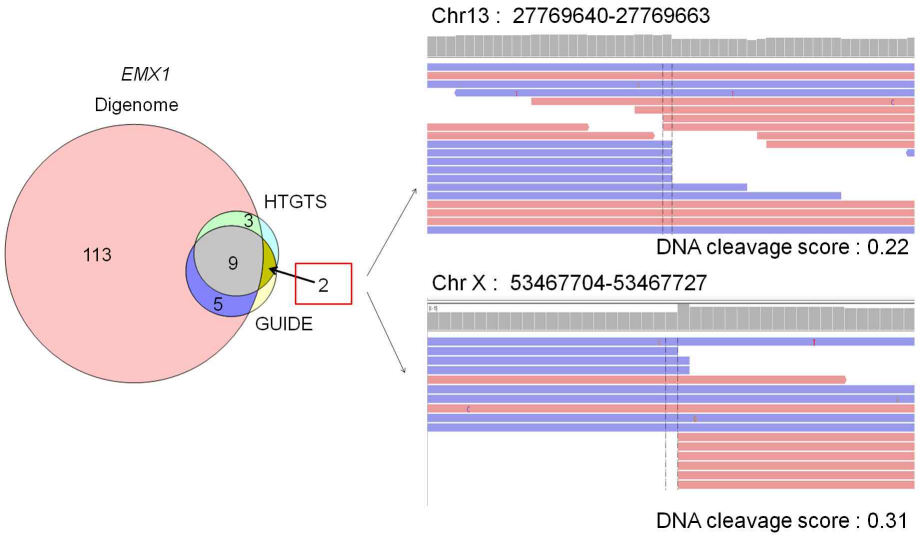


Figure 26. Two EMX1 off-target sites captured by HTGTS and GUIDE-seq but missed by Digenome-seq. The two numbers below the IGV images indicate DNA cleavage scores calculated using the equation in Fig. 15, which are smaller than the cutoff value of 2.5.

The VEGFA 2-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.

I 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 (Figure 27). 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, although it may still be useful for profiling the specificities of dCas9-based transcription factors and epigenome regulators (Tsai et al., 2015).

Figure 27.

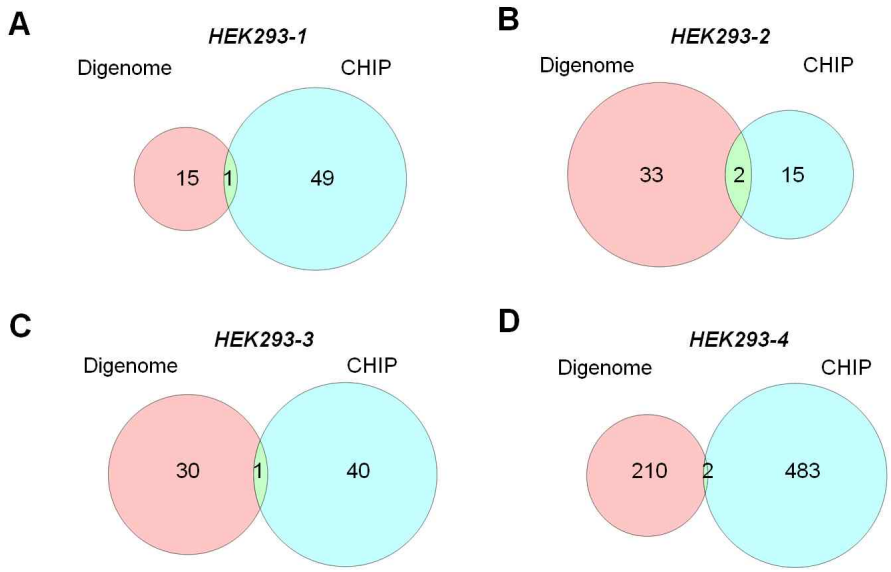


Figure 27. Comparison of the Digenome-seq and CHIP-seq. Venn diagrams showing the number of sites captured by Digenome-seq and CHIP-seq in HEK293-1(A), -2(B), -3(C), and -4(D) site.

Next, I compared Digenome-seq with direct in situ breaks labeling, enrichment on streptavidin and next-generation sequencing (BLESS) (Ran et al., 2015). Majority of BLESS captured sites were not identified by Digenome-seq and sequence logo ,which is captured in only BLESS, showing any similarity with on-target sequence (Figure 28). This result suggests BLESS had high background positive candidate sites.

Figure 28.

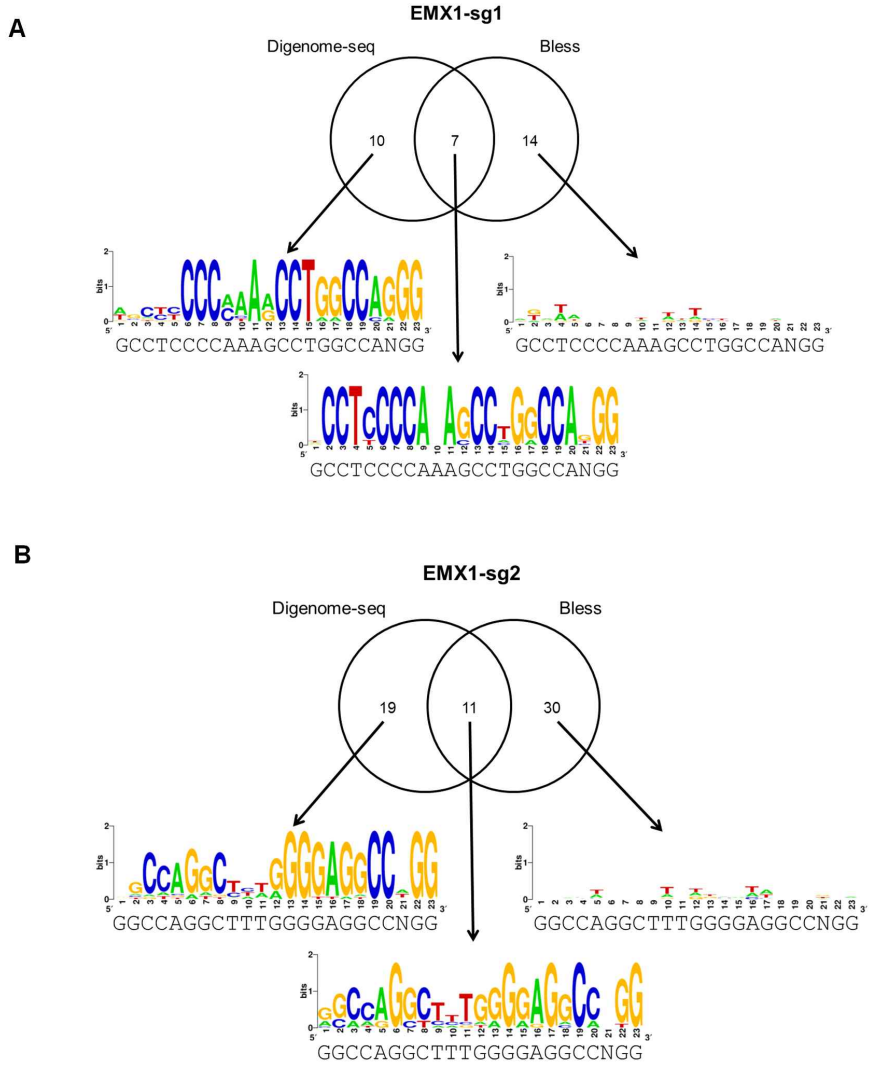


Figure 28. Comparison of the Digenome-seq and Bless. (A-B) Venn diagrams showing the number of sites captured by Digenome-seq and Bless. Sequence logos obtained using Digenome-seq only captured sites, commonly identified by the two methods, and Bless only captured sites.

5. Validation of off-target sites in cells

I 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 5). I 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.

Table 5.

		Digenome only	Digenome and GUIDE	GUIDE only
VEGFA1	Total captured sites	57	22	0
	Number of NGS-tested sites	15	22	0
	Number of validated sites	6	20	0
VEGFA2	Total captured sites	33	30	122
	Number of NGS-tested sites	8	22	14
	Number of validated sites	0	22	10
VEGFA3	Total captured sites	256	46	14
	Number of NGS-tested sites	18	27	9
	Number of validated sites	4	22	5
EMX1	Total captured sites	129	14	2
	Number of NGS-tested sites	16	12	2
	Number of validated sites	3	9	2
FANCF	Total captured sites	38	8	1
	Number of NGS-tested sites	8	8	1
	Number of validated sites	1	8	0
RNF2	Total captured sites	12	1	0
	Number of NGS-tested sites	12	1	0
	Number of validated sites	2	1	0
HEK1	Total captured sites	8	8	2
	Number of NGS-tested sites	3	8	2
	Number of validated sites	1	7	2
HEK2	Total captured sites	33	2	1
	Number of NGS-tested sites	16	2	1
	Number of validated sites	1	2	0
HEK3	Total captured sites	25	6	0
	Number of NGS-tested sites	14	6	0
	Number of validated sites	2	6	0
HEK4	Total captured sites	112	104	26
	Number of NGS-tested sites	17	24	16
	Number of validated sites	1	19	4
Total	Total captured sites	703	241	168
	Number of NGS-tested sites	127	132	45
	Number of validated sites	21	116	23

Table 5. Validation of off-target sites in human cells using next-generation sequencing (NGS). 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.

Indel frequencies at most of these validated sites were below 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 (Figure 29). It still is possible that indels could be induced at NGS-invalidated sites with frequencies below noise levels (0.001% to 4%, depending on the site).

To reduce off-target effects, I replaced sgRNAs with versions containing two extra guanines at the 5' terminus (termed ggX_{20} sgRNAs) (Cho et al., 2014) (Figure 30A). These modified sgRNAs were more specific than their respective GX_{19} sgRNAs by up to 598 fold (Figure 30B-D and Figure 31). It is of note that off-target indels were not detected above noise levels with the RNF2-specific ggX_{20} sgRNA (Figure 31C).

Figure 29.

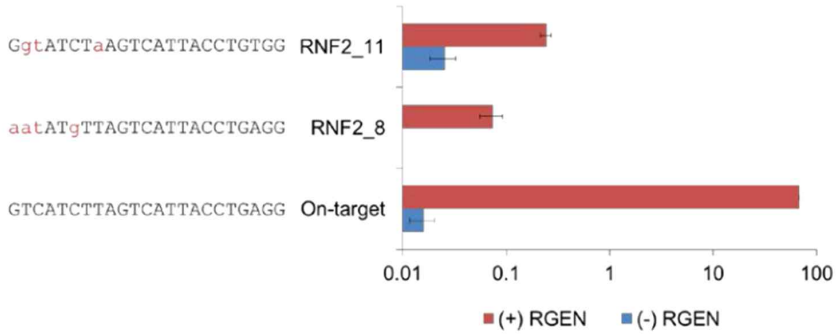


Figure 29. Indel frequencies (log scale) at on-target and off-target sites determined in HeLa cells transfected with the RNF2-specific sgRNA.

Figure 30.

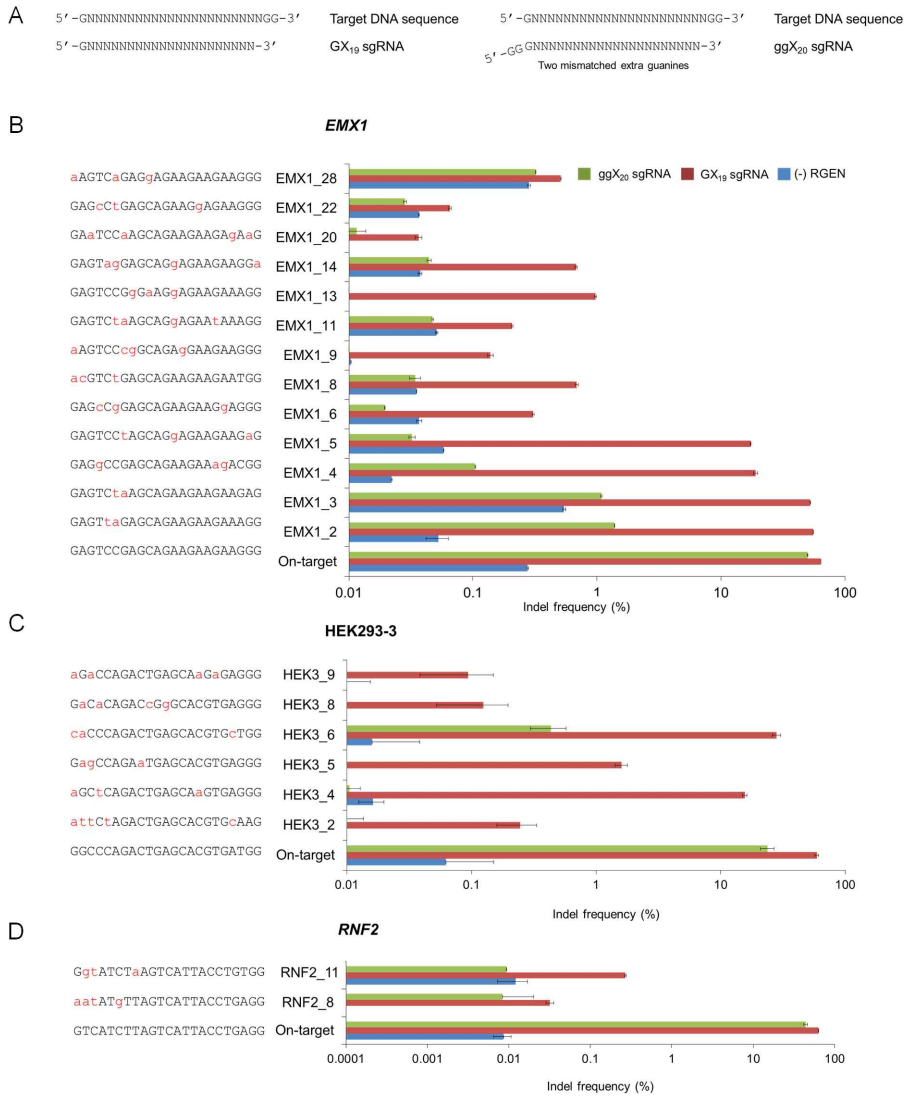


Figure 30. Indel frequencies determined using targeted deep sequencing at off-target sites. (A) Schematic of conventional sgRNAs (gX_{19} sgRNA) and modified sgRNAs (ggX_{20} sgRNA). Indel frequencies at NGS-validated on- and off-target sites for the EMX1 (B), HEK293-3 (C), and RNF2 sgRNAs (D).

Figure 31.

A *EMX1*

	Specificity ratio (on/off)		
Target	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

B HEK293-3

	Specificity ratio (on/off)		
Target	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

C *RNF2*

	Specificity ratio (on/off)		
Target	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 31. Comparison specificity ratio of conventional sgRNAs with modified sgRNAs. Specificity ratio calculated by dividing indel frequencies at the on-target site with those at an off-target site in EMX1, HEK293-3 and RNF2.

D. Generation of RGEN targetable sites prediction program

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 6 mismatches in the human genome and that has no homologous sites with up to 2 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 (Table 6). In addition, I present an off-target score (Table 7) that accounts for numbers of potential off-target sites in the genome, fractions of these sites captured by Digenome-seq (Figure 24B), and median indel frequencies at these sites (Figure 33). 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 will soon be available at our website (www.rgenome.net/digenome).

Table 6.

Gene	Exon	No. of PAM (NGG) -containing sites	No. of sites with no homologous sites harboring up to 2 mismatch in the human genome & No. of sites with fewer than 13,000 homologous sites harboring up to 6 mismatches
VEGFA	Exon1	235	79
	Exon2	8	0
	Exon3	26	18
	Exon4	6	0
	Exon5	1	0
	Exon6	14	5
	Exon7	8	4
	Exon8	252	34
	Total	550	140
EMX1	Exon1	238	73
	Exon2	29	8
	Exon3	245	37
	Total	512	118
FANCF	Exon1	373	90
	Total	373	90
RNF2	Exon1	50	12
	Exon2	4	0
	Exon3	8	0
	Exon4	14	0
	Exon5	21	0
	Exon6	10	0
	Exon7	173	8
	Total	280	20
Total		1715	368

Table 6. Number of targetable sites which are desirable to minimize off-target effects.

Figure 32.

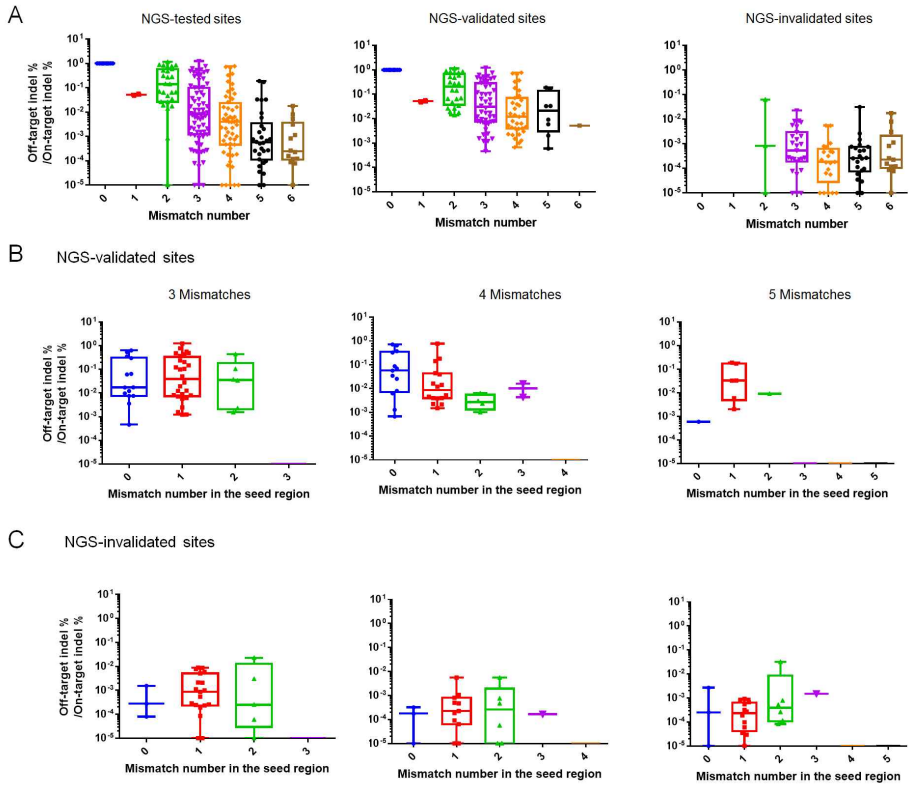


Figure 32. Analysis of NGS-validated and -invalidated off-target sites. Plots of relative indel frequencies (log scale) at off-target sites harboring the number of mismatches indicated in the entire 20-nt sequence (A) or the 10-nt seed sequence (B and C). NGS-tested sites (A) were divided into two groups, 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.

Table 7.

No. of mismatches	No. of mismatches in the seed region	No. of sites[a]	Fraction captured[b]	Median indel frequency[c]	No. of sites x Fraction captured x Median indel frequency
0	-	0	1.0	1.0	0.0
1 or 2	-	1	1.0	0.15	0.15
3	0	7	0.56	0.030	0.12
	1	7	0.44	0.0077	0.024
	2	4	0.12	0.0030	0.0014
	3	0	0.0020	0.00010	0.0
4	0	68	0.22	0.030	0.45
	1	73	0.062	0.0039	0.018
	2	115	0.010	0.00088	0.0010
	3	16	0.0013	0.00088	0.000018
	4	4	0.0	0.0	0.0
5	0	136	0.010	0.00067	0.00091
	1	674	0.010	0.00067	0.0045
	2	888	0.0015	0.00067	0.00089
	3	521	0.00025	0.00067	0.000087
	4	91	0.0	0.0	0.0
	5	3	0.0	0.0	0.0
6	0	426	0.0067	0.00026	0.00074
	1	2641	0.0017	0.00026	0.0012
	2	5673	0.000047	0.00026	0.000069
	3	4954	0.000047	0.00026	0.000061
	4	1846	0.0	0.0	0.0
	5	197	0.0	0.0	0.0
	6	10	0.0	0.0	0.0
				Off-target score:	0.77

Table 7. Calculation of an off-target score assigned to the EMX1 target sequence (5'-GAGTCCGAGCAGAAGAAGAANGG-3') in the human genome.

Figure 33.

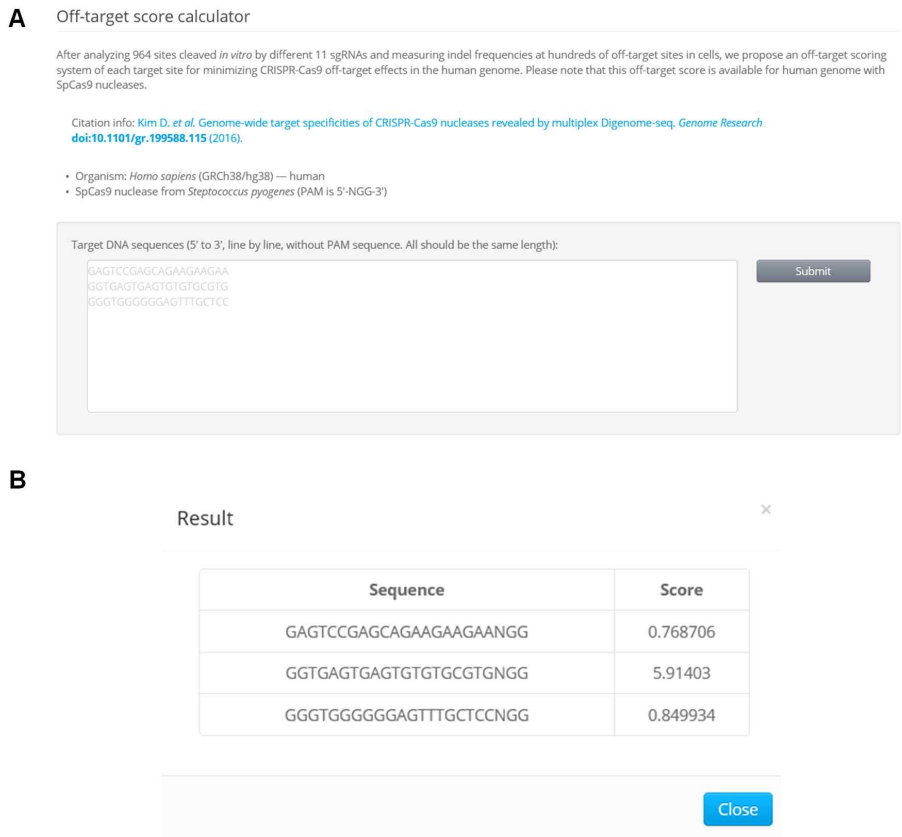


Figure 33. Off-target score calculator. (A) Overview of Off-target score calculator. (B) Example of off-target score calculator result.

E. Generation of RGEN potential off-target sites prediction program based on Digenome-seq.

1. *In vitro* cleavage of genomic DNA

To make a potential off-target sites prediction program, I did digenome-seq using 100-types of sgRNA which targeting different gene (Figure 34A). First, I cloning the plasmid DNA which can express sgRNA by T7 promoter. Second each sgRNAs were expressed by *in vitro* transcription. Third cell-free genomic DNA was mixed with 100-types of sgRNA/Cas9 protein complex and subjected to WGS.

After alignment of sequencing data to hg19 reference genome, I checked straight alignment using IGV at on-/off-target sites (Figure 34B).

Figure 34.

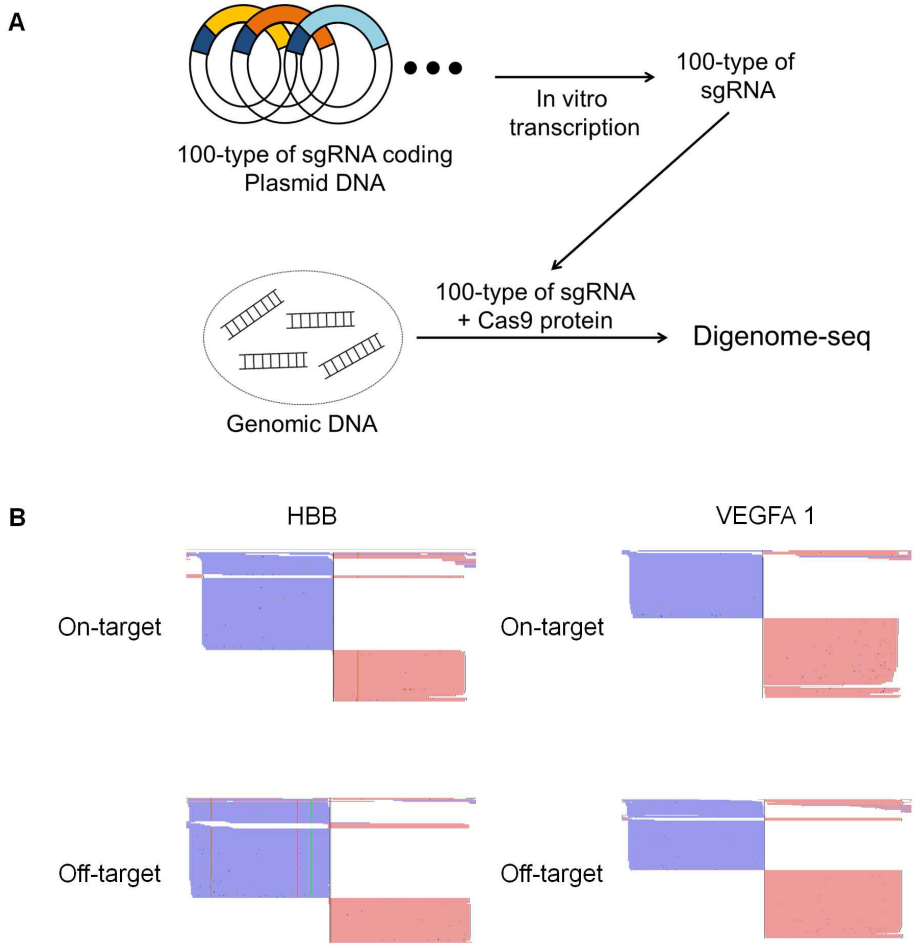


Figure 34. Multiplex Digenome-seq using 100-type sgRNA. (A) Overview of Multiplex Digenome-seq via 100-type sgRNA and Cas9 protein. (B) Representative IGV images obtained using the Multiplex Digenome-seq at HBB and VEGFA1 on-/off-target site.

2. Generation of RGEN potential off-target sites prediction program

To make software program, I checked the cleavage of potential off-target sites which had one to six mismatch with on-target sites and I divided these sites to cleavage sites and non-cleavage sites. I made off-target prediction program using machine learning by finding the difference character between cleavage site and non-cleavage sites (Figure 35). Using these computer program, It is possible to predict potential off-target sites (Table 8).

Figure 35.

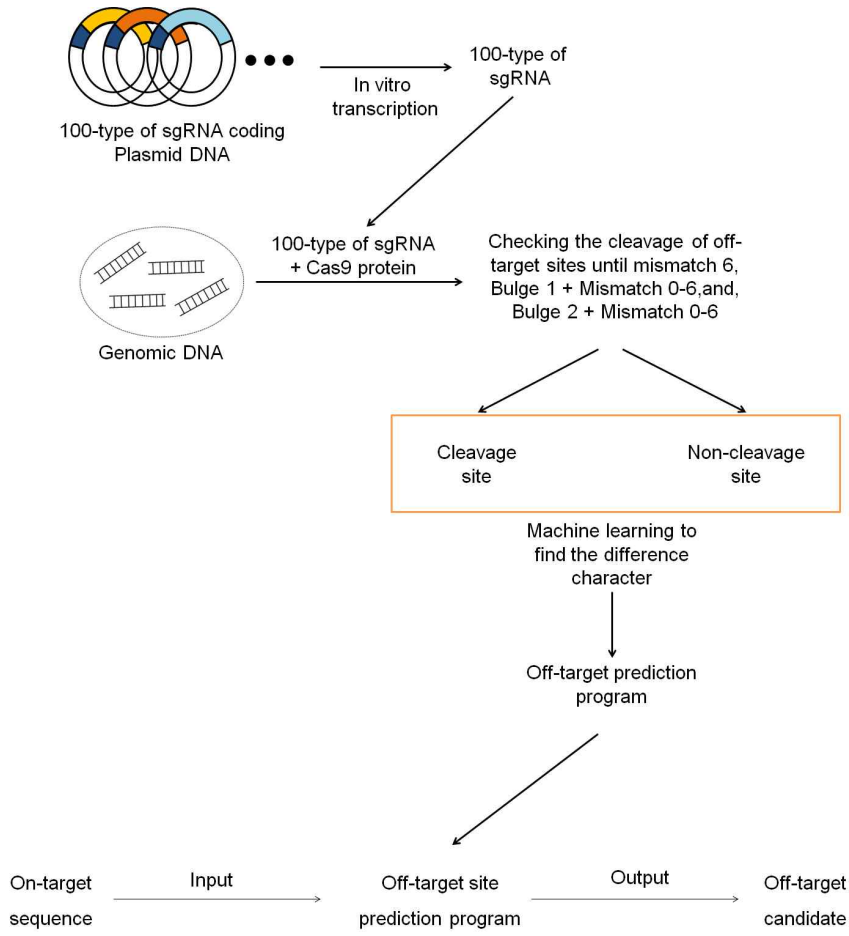


Figure 35. Making of RGEN potential off-target sites prediction program. Overview of generation of RGEN potential off-target sites prediction program.

Table 8.

	bulge_type	sgRNA	DNA	chroms	Position	mismatch	bulge	bed_start	bed_end	log_ibs1000
Rank 1	X	GAGTCCGAGCAGAAGAAGAANRG	GAGTCCGAGCAGAAGAAGAAGGG	chr2	73160981	0	0	73160981	73161004	0.976812398
Rank 2	X	GAGTCCGAGCAGAAGAAGAANRG	GAGTtaGAGCAGAAGAAGAAAGG	chr5	45359062	2	0	45359060	45359083	0.688734292
Rank 3	X	GAGTCCGAGCAGAAGAAGAANRG	GcGaCaGAGCAGAAGAAGAAGGG	chr2	21489994	3	0	21489994	21490017	0.216598949
Rank 4	X	GAGTCCGAGCAGAAGAAGAANRG	tAtaCgGAGCAGAAGAAGAATGG	chr1	35818887	4	0	35818885	35818908	0.205753281
Rank 5	X	GAGTCCGAGCAGAAGAAGAANRG	aAtaCaGAGCAGAAGAAGAATGG	chr1	184236226	4	0	184236226	184236249	0.178576271
Rank 6	X	GAGTCCGAGCAGAAGAAGAANRG	GttcaaGAGCAGAAGAAGAATGG	chr16	54831350	5	0	54831350	54831373	0.144072218
Rank 7	X	GAGTCCGAGCAGAAGAAGAANRG	GAGTaaGAGaAGAAGAAGAAGGG	chr4	87256687	3	0	87256685	87256708	0.08635215
Rank 8	X	GAGTCCGAGCAGAAGAAGAANRG	aAGTCaGAGgAGAAGAAGAAGGG	chr15	61646860	3	0	61646860	61646883	0.08178587
Rank 9	X	GAGTCCGAGCAGAAGAAGAANRG	GAGaCtGAGaAGAAGAAGAAAGG	chr17	8640213	3	0	8640213	8640236	0.058285312
Rank 10	X	GAGTCCGAGCAGAAGAAGAANRG	ttcTCaGAGCAaAAGAAGAATGG	chr2	205473546	5	0	205473546	205473569	0.052926109
Rank 11	X	GAGTCCGAGCAGAAGAAGAANRG	cAaaCgGAGCAGAAGAAGAAAGG	chr17	72740376	4	0	72740376	72740399	0.050444139
Rank 12	X	GAGTCCGAGCAGAAGAAGAANRG	tAaTCCaAtCAGAAGAAGAAGGG	chr10	5401770	4	0	5401770	5401793	0.049991096
Rank 13	X	GAGTCCGAGCAGAAGAAGAANRG	GtGaCaGAGCAaAAGAAGAAAGG	chr22	34716270	4	0	34716268	34716291	0.032400627
Rank 14	X	GAGTCCGAGCAGAAGAAGAANRG	aAGcCCaAGgAGAAGAAGAAGG	chr19	46265337	4	0	46265337	46265360	0.025892542
Rank 15	X	GAGTCCGAGCAGAAGAAGAANRG	GAGTCtaAGCAGAAGAAGAAGAG	chr15	44109746	2	0	44109746	44109769	0.021419144
Rank 16	DNA	GAGTCCGAGCAGAAGAAGA-ANRG	GAaTCCaAGCAGAAGAAGAGAAGG	chr11	62365266	2	1	62365265	62365289	0.018983301
Rank 17	X	GAGTCCGAGCAGAAGAAGAANRG	agGTtaGAGCAaAAGAAGAAAGG	chr10	58498666	5	0	58498666	58498689	0.018055351
Rank 18	X	GAGTCCGAGCAGAAGAAGAANRG	GAGTagGAGgAGAAGAAGAAAGG	chr4	31076550	3	0	31076548	31076571	0.015539706
Rank 19	X	GAGTCCGAGCAGAAGAAGAANRG	GAcTCCtAGCAaAAGAAGAATGG	chr12	119985926	3	0	119985924	119985947	0.014818936
Rank 20	X	GAGTCCGAGCAGAAGAAGAANRG	tAtggCaAGCAGAAGAAGAAAGG	chr12	33507673	5	0	33507673	33507696	0.012843922
Rank 21	X	GAGTCCGAGCAGAAGAAGAANRG	aAGTCtGAGCAcAAGAAGAATGG	chr5	9227145	3	0	9227145	9227168	0.011697691
Rank 22	X	GAGTCCGAGCAGAAGAAGAANRG	GtcaCaGAGCAGAAGAAaAATGG	chr13	24335189	5	0	24335189	24335212	0.010553065
Rank 23	RNA	GAGTCCGAGCAGAAGAAGAANRG	c-GTCtGAGCAGAAGAAGAATGG	chr6	9118795	2	1	9118792	9118814	0.00986482
Rank 24	X	GAGTCCGAGCAGAAGAAGAANRG	attcCaGAaCAGAAGAAGAATGG	chr12	12312581	6	0	12312581	12312604	0.009609559
Rank 25	X	GAGTCCGAGCAGAAGAAGAANRG	tAGcCaGAaCAGAAGAAaAACGG	chr4	138405998	5	0	138405998	138406021	0.009014651
Rank 26	X	GAGTCCGAGCAGAAGAAGAANRG	GAGgagGAGCAGAAGAAaAAAGG	chr14	100929519	4	0	100929517	100929540	0.008721652
Rank 27	X	GAGTCCGAGCAGAAGAAGAANRG	GttTtgaAGCAGAAGAAGAAAGG	chr1	227565632	5	0	227565630	227565653	0.008084588

Table 8. Off-target sites predictor. Example of off-target predictor result.

IV. Discussion

The genome-wide target specificity of CRISPR-Cas9 system is a broad interest in the genome-editing field. Prior studies have documented the CRISPR-Cas9 off-target effects via bioinformatics prediction based on sequence homology, mismatched guide RNA libraries, *in vitro* selection, reporter assays, and ChIP-seq. However, these studies have either been computationally determined by homology sites or have focused on CRISPR-Cas9 binding affinity. In this study, I have used Digenome-seq for genome-wide target cleavage specificity of CRISPR-Cas9 system in an unbiased manner.

Several studies have been suggested to 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 to genome-wide screening of engineered nucleases (Gabriel et al., 2011; Kuscu et al., 2014; Wu et al., 2014), but these methods rely on DNA binding rather than DNA cleavage. Unfortunately, most DNA binding sites are not coupled with DNA cleavage sites (Figure 27). Integrase-defective lentiviral vector (IDLV) capture and *in vitro* selection are two different methods that detect nuclease cleavage sites rather than binding sites (Fu et al., 2014; Wang et al., 2015). However IDLV capture loss several known off-target sites and *in vitro* selection uses biased DNA substrate library that consists of $>10^{12}$ variants. Veryr

ecently, two groupsr eported methods,t ermed HTGTS (High-throughput genome-wide translocation sequencing) (Frock et al., 2015), GUIDE-seq (Tsai et al., 2015), and Bless (Direct in situ breaks labeling, enrichment on streptavidin and next-generation sequencing) (Ran et al., 2015), of capturing double-strand breaks (DSBs) induced by Cas9 in cells. However these methods had high background levels because of many DSB which was arisen in cellular DNA.

I found that the number of off-target candidates are comfortable depend on sgRNA in the range of 13 to 316. These finding is consistent with other genome-wide off-target profiling such as GUIDE-seq (Tsai et al., 2015). Therefore it is important to choose target sites with few off-target effects. Based on 11 kinds of digenome-seq and NGS validation result, I made program to predict off-target effect of target sites (Figure 31). It is believed that this program will eventually be capable of finding less off-target RGEN for therapeutic use.

To minimize or avoid off-target effects is a crucial issue for therapeutic use of RGEN. Several methods such as dimeric Cas9 systems (paired Cas9 nickases (Cho et al., 2014; Mali et al., 2013a; Ran et al., 2013) 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) have been published to reduce RGEN off-target effects. Recently two dependent group published that Cas9 protein modification can help to enhance RGEN specificity. In

this study, I used modified guide RNAs to ggX_{20} sgRNA (additional two extra guanine satthe 5' terminus) to reduce off-target effects. I found that sgRNA modification could increase RGEN specificity up to 598 fold. These sgRNA or Cas9 modifications will help to make specific use of RGEN.

In the early RGEN off-target research, serval paper published that SpCas9 have high off-target mutagenesis effects in cells(Cradick et al., 2013;Fu et al., 2013;Hsu et al., 2013;Lin et al., 2014;Pattanayak et al., 2013). These papers finds potential off-target sites using in silico prediction based on similarity to the intended target site and mentioned that SpCas9 off-target cleavage sites have been detected up to five mismatches relative to the intended target sequence. However these papers choose 'promiscuous' RGEN, which had many similar sequences in the genome and these 'promiscuous' RGEN also had less that ten validated off-target in the whole genome with low mutagenesis level (Figure 15, 16). These validated off-target sites can be reduced by sgRNA modification. These result mentioned SpCas9 had somehow off-target sites and these off-target sites can be reduced by sgRNA or Cas9 modification.

In conclusion, Digenome-seq enables genome-wide off-target profiling of CRISPR-Cas9 system and this method also allow genome-wide profiling of other artificial nuclease such as ZFN, TALEN, Meganuclease, and Cpf1(Zetsche et al., 2015). Before use in a gene or cell therapy application, off-target effect should be monitored by Digesome-Seq to avoid oncogenic mutations and unwanted mutations

in essential genes. This method will also provide a litmus test in the development of next generation genome editing tools.

V. Reference

Bae, S., Park, J., and Kim, J.S. (2014). Cas-OFFinder: A fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics.

Bibikova, M., Beumer, K., Trautman, J.K., and Carroll, D. (2003). Enhancing gene targeting with designed zinc finger nucleases. Science 300, 764.

Bitinaite, J., Wah, D.A., Aggarwal, A.K., Schildkraut, I. (1998). FokI dimerization is required for DNA cleavage. Proceedings of the National Academy of Sciences of the United States of America 95, 10570-10575.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstad,t A., Bonas, U. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. Science 326, 1509-1512.

Brunet, E., Simsek, D., Tomishima, M., DeKelver, R., Choi, V.M., Gregory, P., Urnov, F., Weinstock, D.M., and Jasin, M. (2009). Chromosomal translocations induced at specified loci in human stem cells. Proceedings of the National Academy of Sciences 106, 10620-10625.

Carette, J.E., Guimaraes, C.P., Varadarajan, M., Park, A.S., Wuethrich, I., Godarova, A., Kotecki, M., Cochran, B.H., Spooner, E., Ploegh, H.L., et al. (2009). Haploid genetic screens in human cells

identify host factors used by pathogens. Science 326, 1231-1235.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nature biotechnology 31, 230-232.

Cho, S.W., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S., and Kim, J.S. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome research 24, 132-141.

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.

Cradick, T.J., Fine, E.J., Antico, C.J., and Bao, G. (2013). CRISPR/Cas9 systems targeting beta-globin and CCR5 genes have substantial off-target activity. Nucleic acids research 41, 9584-9592.

Drake, J.W., Charlesworth, B., Charlesworth, D., and Crow, J.F. (1998). Rates of spontaneous mutation. Genetics 148, 1667-1686.

Frock, R.L., Hu, J., Meyers, R.M., Ho, Y.J., Kii, E., and Alt, F.W. (2015). Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nature biotechnology 33, 179-186.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nature biotechnology 31, 822-826.

Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014). Improving CRISPR-Cas nuclease specificity using truncated

guide RNAs. Nature biotechnology.

Gabriel, R., Lombardo, A., Arens, A., Miller, J.C., Genovese, P., Kaeppel, C., Nowrouzi, A., Bartholomae, C.C., Wang, J., Friedman, G., et al. (2011). An unbiased genome-wide analysis of zinc-finger nuclease specificity. Nature biotechnology 29, 816-823.

Guilinger, J.P., Thompson, D.B., and Liu, D.R. (2014). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nature biotechnology 32, 577-582.

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. Nature biotechnology 31, 827-832.

Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.R., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nature biotechnology 31, 227-229.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nature biotechnology 31, 233-239.

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.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. eLife 2, e00471.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H.R., Hwang, J., Kim, J.I., and Kim, J.S. (2015). Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nature methods 12, 237-243, 231 p following 243.

Kim, H., and Kim, J.S. (2014). A guide to genome engineering with programmable nucleases. Nature reviews Genetics 15, 321-334.

Kim, H., Um, E., Cho, S.R., Jung, C., and Kim, J.S. (2011). Surrogate reporters for enrichment of cells with nuclease-induced mutations. Nature methods 8, 941-943.

Kim, H.J., Lee, H.J., Kim, H., Cho, S.W., and Kim, J.S. (2009). Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly. Genome research 19, 1279-1288.

Kim, S., Kim, D., Cho, S.W., Kim, J., and Kim, J.S. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. Genome research 24, 1012-1019.

Kim, Y., Kweon, J., Kim, A., Chon, J.K., Yoo, J.Y., Kim, H.J., Kim, S., Lee, C., Jeong, E., Chung, E., et al. (2013a). A library of TAL effector nucleases spanning the human genome. Nature biotechnology 31, 251-258.

Kim, Y.G., Cha, J., Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proceedings of the National Academy of Sciences of the United States of America 93, 1156-1160.

Kim, Y.K., Wee, G., Park, J., Kim, J., Baek, D., Kim, J.S., and Kim, V.N. (2013b). TALEN-based knockout library for human microRNAs. Nature structural and molecular biology 20, 1458-1464.

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. Nature biotechnology 32, 677-683.

Lee, H.J., Kweon, J., Kim, E., Kim, S., and Kim, J.S. (2012). Targeted chromosomal duplications and inversions in the human genome using zinc finger nucleases. Genome research 22, 539-548.

Lin, Y., Cradick, T.J., Brown, M.T., Deshmukh, H., Ranjan, P., Sarode, N., Wile, B.M., Vertino, P.M., Stewart, F.J., and Bao, G. (2014). CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic acids research 42, 7473-7485.

Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. (2013a). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nature biotechnology 31, 833-838.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013b). RNA-guided human genome engineering via Cas9. Science 339, 823-826.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2011). A TALE nuclease architecture for efficient genome editing. Nature

biotechnology 29, 143-148.

Moscou, M.J., Bogdanove, A.J. (2009). A simple cipher governs DNA recognition by TAL effectors. Science 326, 1501.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. Cell.

Pattanayak, V., Lin, S., Guilinger, J.P., Ma, E., Doudna, J.A., and Liu, D.R. (2013). High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nature biotechnology 31, 839-843.

Porteus, M.H., and Baltimore, D. (2003). Chimeric nucleases stimulate gene targeting in human cells. Science 300, 763.

Raczy, C., Petrovski, R., Saunders, C.T., Chorny, I., Kruglyak, S., Margulies, E.H., Chuang, H.Y., Kallberg, M., Kumar, S.A., Liao, A., et al. (2013). Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. Bioinformatics 29, 2041-2043.

Ramakrishna, S., Kwaku Dad, A.B., Beloor, J., Gopalappa, R., Lee, S.K., and Kim, H. (2014). Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. Genome research 24, 1020-1027.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015). In vivo genome editing using Staphylococcus aureus Cas9. Nature 520, 186-191.

Ran, F.A., Hsu, P.D., Lin, C.Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., 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.

Schmittgen, T.D., and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C(T) method. Nature protocols 3, 1101-1108.

Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J., and Joung, J.K. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nature biotechnology 32, 569-576.

Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P., et al. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nature biotechnology 33, 187-197.

Urnov, F.D., Miller, J.C., Lee, Y.L., Beausejour, C.M., Rock, J.M., Augustus, S., Jamieson, A.C., Porteus, M.H., Gregory, P.D., and Holmes, M.C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature 435, 646-651.

Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., Gregory P.D. (2010). Genome editing with engineered zinc finger nucleases. Nature reviews Genetics 11(9): 636-646.

Wang, X., Wang, Y., Wu, X., Wang, J., Wang, Y., Qiu, Z., Chang, T., Huang, H., Lin, R.J., and Yee, J.K. (2015). Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using

integrase-defective lentiviral vectors. Nature biotechnology 33, 175-178.

Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nature biotechnology 32, 670-676.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163, 759-771.

Zuris, J.A., Thompson, D.B., Shu, Y., Guilinger, J.P., Bessen, J.L., Hu, J.H., Maeder, M.L., Joung, J.K., Chen, Z.Y., and Liu, D.R. (2015). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing *in vitro* and in vivo. Nature biotechnology 33, 73-80.

국문 초록

크리스퍼 유전자가위는 인간 세포를 비롯한 다양한 동식물의 유전체 교정에 사용되어 왔지만, 인간 DNA 전체에 대한 정확성을 측정할 수 있는 방법이 없었다. 이를 해결하기 위해 인간 DNA를 크리스퍼 유전자 가위로 처리한 후 전유전체 시퀀싱(Whole genome sequencing)을 통해 표적 염기서열(On-target)과 비표적 염기서열(Off-target)을 찾는 방법인 절단 유전체 시퀀싱(Digenome-seq)을 개발하였다. 절단 유전체 시퀀싱을 이용하여 찾아낸 비표적 염기서열이 실제 세포내에서도 작동하는 것을 확인 하였으며, 크리스퍼 유전자 가위를 구성하는 가이드 RNA의 말단에 구아닌 염기를 추가함으로서 비표적위치에는 작동하지 않고 인간 유전체에서 단 한군데에만 작용하는 정교한 유전자가위를 만드는데 성공하였다. 이 실험결과를 통해 크리스퍼 유전자가위가 기존에 알려진 것과는 달리 매우 정교하다는 사실을 보고하였다.

새롭게 만든 절단 유전체 시퀀싱 방법을 응용하여 한번에 10개 이상의 크리스퍼 유전자 가위의 오프타깃을 예측할 수 있는 Multiplex Digenome-seq 방법을 개발하였다. 또한, Multiplex Digenome-seq의 결과를 이용하여 비표적위치가 적은 크리스퍼 유전자 가위를 예측할 수 있는 프로그램과 크리스퍼 유전자 가위의 비표적 위치를 예측할 수 있는 프로그램을 만들었다. 이러한 기술을 이용하면 치료 목적으로 크리스퍼 유전자 가위를 사용하게 될 때, 정교한 유전자 가위를 찾거나 특정 크리스퍼 유전자 가위의 비표적 위치를 예측하여 부작용을 없애는 데에 큰 도움이 될 것이다.

학　　번 : **2013-30088**