Genome-wide prediction and profiling of Off-target cleavage by CRISPR-Cas9

Soobin Jung

Department of Medical Science

The Graduate School, Yonsei University

Genome-wide prediction and profiling of Off-target cleavage by CRISPR-Cas9

Directed by Professor Hyongbum Kim

The Doctoral Dissertation
submitted to the Department of Medical Science.
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Medical Science

Soobin Jung

June 2018

ACKNOWLEDGEMENT

길다면 길고, 짧다면 짧은 박사학위과정을 진행하면서 저의 학위 논문이 잘 마무리 될 수 있게 많은 분들의 도움이 있었습니다. 이 글을 통해 감사의 인사를 드리고자 합니다.

먼저 저의 연구에 새로운 방향을 제시해 주시고 마지막까지 큰 그림을 그릴 수 있게 늘 아낌없는 지도를 해주신 김형범 교수님 감사드립니다. 연구자로서 교수님 덕분에 좋은 연구를 할 수 있었고, 좋은 연구자로서 가져야 할 자세를 배우게 되었습니다. 또한 바쁘신 중에도 면밀하게 논문의 심사를 맡아주신 조성래 교수님, 배상수 교수님, 김현석 교수님, 김진석 교수님 감사드립니다.

또한 제가 무사히 박사학위를 마칠 수 있게 묵묵히 뒤에서 아낌없는 지원을 해 주신 부모님께 가장 큰 감사의 인사를 드립니다. 10년이라는 서울생활동안 맘 고생을 많이 하시면서도 먼 타지에 있는 딸래미를 위해 끝없는 믿음과 희생, 그리고 든든한 버팀목이 되어주신 엄마, 아빠, 언니를 위해 모든 것을 다 양보한 하나밖에 없는 내 여동생 원빈이, 그리고 저 대신 부산에서 엄마에게 많은 도움과 힘이 되어준 이모에게 이 박사학위 취득의 영광을 돌리고 싶습니다. 사랑합니다.

대학원 생활을 하면서 저에게 없어선 안 될 또다른 존재들은 바로 친구들과 실험실 동료들이었습니다. 같은 분야에 종사하면서 공감하는 모든 기쁨과 슬픔을 항상 함께 공유해준 내 대학 친구들, 늘 든든한 쉼터가 되어준 부산 친구들, 친형제자매같이 진심으로 대해준 대학원 언니, 오빠들, 그리고 거의 대부분의 시간을 함께 부대끼며 많은 정이 든 우리 실험실 동료들 모두 너무 감사하고, 각자의 위치에서 가장 행복한 사람들이 되길 바랍니다.

마지막으로, 동갑이지만 오빠 같은 든든함으로 늘 편하게 기댈 수

있게 해준 남자친구야, 고맙고 앞으로 더 행복하게 지내자!

지면으로 미처 언급하지 못했지만, 저를 아끼고 격려해 주셨던 모든 분들께 진심으로 감사하다는 말씀을 전합니다. 더욱 정진하며 바른 모습으로 성장해 대한민국의 생명과학자로서 꼭 필요한 존재가 되도록 노력하겠습니다.

2018년 6월

정수빈 올림

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Genome-wide prediction and profiling of Off-target cleavage by CRISPR-Cas9

Soobin Jung

Department of Medical Science

The Graduate School, Yonsei University

(Directed by Professor Hyongbum Kim)

ABSTRACT

Although CRISPR RNA-guided nucleases (RGENs) are widely used in genome-editing studies, their off-target cleavage activity on a genome-wide scale has not been yet distinguished. Here, we developed a method to detect unintended DNA double-stranded breaks (DSBs) introduced by CRISPR-Cas9, and by extension, predict genome-wide off-target sites of specific guide RNA sequences through a deep learning-based method. This prediction tool implemented experimental *in vivo* data and thus, should be more accurate than previously reported computational prediction methods. Our effort represents the most exhaustive genome-wide survey of Cas9 off-target effects in evaluating its safety before clinical applications.

Key words: CRISPR-Cas9, Genome editing

Genome-wide prediction and profiling of Off-target cleavage by CRISPR-Cas9

Soobin Jung

Department of Medical Science

The Graduate School, Yonsei University

(Directed by Professor Hyongbum Kim)

I. INTRODUCTION

CRISPR-Cas (clustered, regularly interspaced, short palindromic repeats (CRISPR)–associated (Cas))-derived RNA-guided engineered nucleases (RGENs) derived from prokaryotic adaptive immune system are now widely used for genome editing in many biological research as well clinical applications. In the type Ⅱ CRISPR systems, CRISPR regions are transcribed as *pre-CRISPR* RNA (pre-crRNA) and processed to give rise to target-specific crRNA. Invariable target-independent *trans-activating* crRNA (tracrRNA) is also transcribed from the locus and contributes to the processing of pre-crRNA[1]. In genome editing research, a single-guide RNA engineered as a fusion complex of crRNA and tracrRNA is synthesized with CRISPR-associated protein 9 (Cas9) to form an active DNA endonuclease. After Cas9 recognizes a 5′ -NGG- 3′ protospacer adjacent motif (PAM) sequence and sgRNA recognizes a 20bp sequence identical to the guide sequence of target DNA, RGEN induces DNA double-stranded breaks (DSBs). The cleaved target DNA is repaired by nonhomologous end-joining (NHEJ) in the absence of homology templates or homology directed repair (HDR) in the present of donor DNA. NHEJ can induce variable length insertion/deletion alterations (indels)[2] at the double stranded break region. Although sgRNA recognizes a specific target DNA sequence, RGEN can often introduce unintended cleavages at non-target DNA sequence. RGEN can tolerate mismatches up to several nucleotides[3-6]. For this reason, RGEN cuts non-target DNA like the target DNA. This phenomenon is called “off-target effect”. Off-target DNA cleavages can lead to genomic alterations at the unintended genomic loci, inactivate essential genes, or activate oncogenes. This problem is a major challenge when applying CRISPR-Cas9 to clinical therapeutics. Accordingly, the therapeutic use of RGENs in humans has demanded a cautious approach such as reducing the risk of off-target effects.

Previous research groups have reported methods to improve RGEN specificity and reduce off-target effects by using sgRNAs with two extra guanine nucleotides at the 5′ end[7], truncated sgRNAs[8], paired Cas9 nickases[7,9,10], a catalytically dead Cas9 (dCas9)-FokI fusion[11,12] and delivery of purified Cas9 protein[13-15]. Although these approaches have been shown to reduce the frequency of off-target alterations, these RGEN variants could not completely remove off-target effects from the entire genome as it is very difficult to completely resolve the off-target effects.

Many methods have been developed to identify RGEN off-targets across the whole genome using both *in vivo* and *in vitro* systems. *In vivo* techniques such as GUIDE-seq[16], IDLV[19], HTGTS[20], and BLESS[21-23] in addition to *in vitro* methods such as Digenome-seq[17,18] and CIRCLE-seq[36] can be used to quantify off-target effects. However, some methods could not identify off-target sites on a genome-wide scale. In order to better assess the specificity of genome-wide methods, it will be necessary to develop a more highly sensitive method that can detect even low-frequency alterations. It is the most critical issue that sensitive, unbiased, and genome-wide method for profiling off-target effects needs to be addressed when evaluating RGEN application for therapeutic purposes.

Therefore, we developed our own method that identified off-targets comprehensively. Previously, we constructed a paired library system for guide RNA-target sequences and demonstrated that the system can represent endogenous target activities corresponding to synthetic target sequence[37]. The guide RNA sequence and target sequence are on the same strand, which allows the guide RNA to easily bind to the target sequence to achieve a more rapid and precise detection of target activity levels. We applied this concept to profiling off-target effects. When compared to the previous methods, we found that our method was able to yield more comprehensive and sensitive results.

Although various techniques have been developed to profile off-target sites, scaling these assays to a genome-wide level has been difficult for most research groups due to the high costs and laborious experiments[24]. Therefore, several groups have developed computer-based off-target prediction models. These models can learn the statistical regularities of guide RNA–target sequence pairs, which enable an inexpensive and rapid *in silico* screening of off-target effects across the genome for guide RNAs that were previously left unexamined.

There are three main approaches for modelling off-target prediction programs[25]. First approach is to *Search and Filter* on a genome-wide scale for potential targets of a specific guide RNA[25]. Second approach is *Scoring* the potential off-target activity that is expected for a certain guide RNA–target pair[25]. The last approach is to *Aggregate* the scores into a single off-target potential with which to assess the guide RNA[25]. Many tools have been reported utilizing the first approach of search and filter, including Cas-OFFinder[26], CRISPOR[27], CHOP-CHOP[28], E-CRISPR[29], CRISPR-DO[30], CROP-IT[31], and COSMID[32]. Researchers are able to detect potential off-targets of interesting guide RNA sequences by adjusting the parameters of the search such as the number of mismatches. The second and third approaches, scoring and aggregation, have been implemented less frequently compared to the first approach. Accordingly, the common tools that are currently available are Massachusetts Institute of Technology (MIT) web server[33], CFD[24], and CCTOP[34]. The CHOP-CHOP tool indicates the number of potential off-targets without scoring. MIT, CFD, and CCTOP are rule-based tools that require the researchers to set the rule 'artificially'. These tools do not undergo an extensive training step, which can cause the results to be biased and inaccurate. The recently reported machine learning-based tool[25] called Elevation develops a two-layer regression model (Boosted RT + L1 Regression). In the first layer, the Elevation tool scores a single guide RNA-target

pair by learning to predict the off-target activity for a single mismatch of the guide RNA-target pairs and combines the predictions with multiple mismatches[25]. This method has shown to produce the best results among the currently available off-target prediction tools. However, the results were found to not correlate well with genome-wide off-targets[25].

For this reason, we developed Deep-Learning based off-target prediction model. There are some key differences between machine learning and deep learning methods. Machine learning uses types of automated algorithms, which learn to predict future decisions and model functions, using the available input data. Deep learning interprets the data features and its relationships using neural networks in which the relevant information is passed through several stages of data processing. Importantly, machine learning needs thousands of data points for development, while deep learning requires millions of data points. The main difference among these two models is the management by users or itself. Our model is an end-to-end deep learning framework based on a convolutional neural network (CNN) for off-target score prediction[35]. We expect that our model will outperform the previous prediction models. Furthermore, this will be the first step in determining off-target scores comparable to the on-target scores of guide RNAs.

II. MATERIALS AND METHODS

1. Oligonucleotides

For high-throughput experiments, a total of 11,000 spCas9 on and off target sequences (1,100 for experiments 10 guides, independently) were screened and designed from Cas-OFFinder[26] (http://www.rgenome.net/cas-offinder) . We designed each oligonucleotide to contain the 20-nt guide-RNA-encoding sequence, scaffold region, 18-nt barcode, and 34-nt target sequence in a total length of 170 nucleotides. The oligonucleotides were independently synthesized by CustomArray (Bothell, WA) and Cellemics (Seoul, South Korea).

2. Plasmid library preparation

The oligonucleotide pool was PCR-amplified with Phusion Polymerase (NEB, Ipswich, MA). The Lenti-gRNA-Puro plasmid(Addgene;84752) was linearized with BsmBI enzyme (NEB). Products were gel-purified using a MEGAquick-spin total fragment DNA purification kit (iNtRON Biotechnology, Seongnam, South Korea) and assembled with an NEBuilder HiFi DNA assembly kit (NEB). Next, the assembled product was transformed into electrocompetent cells (Lucigen, Middleton, WI) via a MicroPulser (Bio-Rad, Hercules CA). Transformed cells were seeded onto Luria-Bertani (LB) agar plates supplemented with 50 μg/ml carbenicillin and incubated for 16 hr at 37 ℃. Before harvest, the library coverage was calculated as (total number of colonies/total number of gRNA-target pairs in sample). The resulting library coverage ranged from 30× to 36×. Total colonies were harvested and plasmids were extracted using a Plasmid Midiprep kit (Qiagen, Hilden, Germany).

3. Lentivirus production

For lentivirus production, transfer plasmids (containing the gene of interest), psPAX2, and pMD2.G were mixed at a ratio of 4:3:1, and a total of 18 μg of the plasmid mixture was delivered to 80–90% confluent HEK293T cells (ATCC) using Lipofectamine LTX (Invitrogen, Carlsbad, CA). After 12 h of transfection, cells were refreshed with 10 ml of growth medium. The supernatant- (or media)- containing virus was collected at 48 h after the initial transfection. Two batches of virus-containing media were combined and centrifuged at 2,100g at 4 ℃ for 5 min. Next, the supernatantswere filtered through a Millex-HV 0.45 μm low-protein-binding membrane (Millipore, Darmstadt, Germany) and stored at -80 ℃ until use. To estimate the efficiency of spCas9 virus production, samples of these batches of frozen virus-containing media were thawed, and virus production efficiency wasmeasured with a Lenti-X p24 Rapid Titer Kit (Clontech, Mountain View, CA) according to the manufacturer's instructions.

4. Cell library generation

Six cell libraries were constructed independently as described below. HEK293T cells (6.0×10^5 cells per each pool of 1k oligonucleotide library) were transduced with the above-mentioned lentiviral vector containing gRNA-encoding and target sequences in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% FBS (FBS, Gibco, Waltham MA). After 1 d, transduced cells were treated with 2 μg/ml puromycin for the following 3 days to remove untransduced cells. To preserve the libraries, at least 6.0×10^5 cells of each were maintained throughout the study. Cas9 delivery into the cell libraries. A total of 6.0×10^5 cells were seeded and transduced with spCas9-encoding lentiviral vectors in DMEM supplemented with 10% FBS. After 1 d of transduction, the culture medium was exchanged with DMEM supplemented with 10% FBS and 20 μg/ml blasticidin S (InvivoGen); cultures were then incubated for 7 days to remove untransduced cells. All steps were repeated

equivalently for each cell library.

5. T7E1 assay

To verify the on-target activity of 10 guides, we first cloned guide sequence to pGX19 vector. And the guide expression vector and spCas9 vectors were co-transfected into HEK293 T cell. After 48 hrs, genomic DNA was isolated using the Wizard Genomic DNA purification Kit (Promega, Madison, WI) according to the manufacturer's instructions. The region including the nuclease target site was nested PCR-amplified using appropriate primers. PCR was performed in two different reactions. First PCR reaction condition [Template: 300ng genomic DNA; Denaturation at 95°C for 30 secs, Annealing at 60°C for 30 secs, Extension at 72°C (extension time is one minute per kb); 25 PCR cycles]. Second PCR reaction condition [Template: First PCR amplicon sample: distilled water, diluted to 1:200 ratio; Denaturation at 95°C for 30 secs, Annealing at 60°C for 30 secs, Extension at 72°C; 35 PCR cycles]. The amplicons were denatured by heating and annealed to allow formation of heteroduplex DNA, which was treated with 5 units of T7 endonuclease 1 (New England Biolabs) for 20 min at 37°C followed by analysis using 2% agarose gel electrophoresis. Mutation frequencies were calculated based on the band intensities using Image J software and the following equation (Guschin et al, 2010): mutation frequency (%) = $100 \times (1 - (1 - \text{fraction cleaved})^{1/2})$, where the fraction cleaved is the total relative density of the cleavage bands divided by the sum of the relative density of the cleavage bands and uncut bands.

6. Targeted deep sequencing

Genomic DNA was extracted from the cell library using a Wizard Genomic DNA purification kit (Promega, Fitchburg, WI). Target sequences integrated by lentivirus were PCR-amplified using Phusion polymerase (NEB). A total of 12 μg of genomic DNA per cell library was used as the first PCR template to achieve >100× coverage over the library (assuming 10 μg of genomic DNA for 1.0×10^6 293T cells). For each cell library, we performed 12 separate 50-μl reactions, in which the initial genomic DNA amount was 1 μg per 50-μl reaction, and then combined all of the resulting products. For the cells transduced with spCas9- and crRNA-encoding lentivirus, 100 ng of genomic DNA was used for PCR amplification of integrated target sites. The PCR products from the first reaction were then purified with a MEGAquick-spin Total Fragment DNA Purification Kit (iNtRON Biotechnology) and 20 ng of purified products was annealed with both Illumina adaptor and barcode sequences for the second PCR. The resulting products were isolated, purified, mixed, and analyzed using MiSeq or HiSeq (Illumina, San Diego, CA).

7. Analysis of indel frequencies

Deep-sequencing data were sorted and analyzed by custom Python scripts. Based on the 18-base barcode sequences, each crRNA-target pair was sorted. Insertions or deletions located around the expected cleavage site (i.e., a 9-bp region centered on the middle of the cleavage site) were considered to be Cas9-induced mutations. To exclude the background indel frequencies that originated from oligonucleotide synthesis and the target site amplification procedure, the bona fide indel frequency induced by Cas9 and crRNA activity was calculated by subtracting the background indel frequency in the cell library in the absence of Cas9 delivery from the observed indel frequency. To increase the accuracy of analysis, deep-sequencing data from

high-throughput experiments were filtered; different filtering conditions were used for each high-throughput experiment, depending on the required sizes of the data sets, required accuracy of each component of the data sets, the total number of reads, and the background indel frequency in the library.

8. Convolutional neural network

A convolutional neural network (CNN) is a type of feed-forward artificial neural network. The key aspect of CNNs is that they can learn hierarchical spatial representations, rather than relying on laborious manual feature engineering. The architectural components of a CNN include three types of layers: convolution layers, pooling layers, and fully connected layers. In the convolution layers, weight vectors called filters are multiplied across the subregions of all the data. They enable CNNs to discover locally correlated patterns regardless of their locations in the data. The pooling layers perform the maximum or average subsampling of non-overlapping subregions, providing invariance to local transitions. The fully connected layers aggregate local features into more highly abstract features by computing weighted sums and applying nonlinear functions. Designed to analyze spatial information, CNNs have made major advances in various tasks such as image recognition and natural language processing. The amount of data required for proper CNN model development can vary considerably depending on the objectives of each task, the data complexity, and other factors; nevertheless, a rough rule of thumb is that 5,000 labeled examples per category would generally be sufficient for acceptable performance. In bioinformatics, CNNs are also showing great promise for genomic sequence analysis. Traditional approaches in genomic sequence analysis often incorporate hard-coded position weight matrices (PWMs) to identify regulatory motifs. On the other hand, an initial convolution layer in a CNN corresponds to motif detectors where PWMs are not hard-coded but solely learned from data. Prior studies

have demonstrated that CNNs can outperform state-ofthe- art methods in diverse applications, including predictions of transcription factor binding affinity and DNA sequence accessibility.

9. Deep-Learning model design

Our off-target prediction model is a deep-learning framework for spCas9 off-target indel frequency prediction. This model receives a 34-bp target sequence as input, and it produces a regression score that highly correlates with spCas9 activity. The model can automatically learn informative representations of target sequences relevant to spCas9 activity profiles. The model proceeds in four stages. (1) The one-hot encoding input layer converts the sequence into numerical representations for downstream processing. It encodes the nucleotide in each position as a four-dimensional binary vector, in which each element represents the type of nucleotide: A, C, G, and T. The encoding layer then concatenates the binary vectors into a 4-by-34 dimensional binary matrix representing the whole 34-bp target sequence. (2) The convolution layer performs one-dimensional convolution operations with 80 filters of length 5. The filters slide along only one axis (i.e., sequence length) of the one-hot encoded matrix containing the 4-nt channels. This process is equivalent to scanning learned PWMs across the target sequence in conventional techniques. The convolution layer then applies the rectified linear unit (ReLU) nonlinear function [f(x) = max (0,x)]] to the convolution outputs. The pooling layer computes the average in each of the non-overlapping windows of size 2, providing invariance to local shifts and reducing the number of parameters. (3) The modeluses three fully connected layers with 80, 40, and 40 units, respectively. Each unit in the fully connected layers performs linear transformations of the outputs of the previous layer, and applies the rectified linear unit nonlinear function. Multiple nonlinear layers enable the model to learn hierarchical representations of data with increasing levels of abstraction. (4) The last

stage, the regression output layer, performs a prediction of spCas9 off-target activity.

10. Training of Deep-Learning model

Model selection and pre-training. First, we split CD33 data sets[24] by random sampling. To demonstrate the reliability of the model selection process, we conducted nested cross-validation with CD33 data set. In eachfold of the outer ten fold cross-validation, we randomly constructed training data sets with different sizes to evaluate the performance improvements associated with training data sets of different sizes. Each training data set was used for the following model selection in the inner fivefold cross-validation and for training of the selected model. Of note is that the validation data set was fixed for all of the training data sets of different sizes within the same fold of the outer crossvalidation. In each fold of the inner cross-validation, the respective training and validation data sets were used to train and validate 180 model candidates with different hyperparameter configurations of the number of filters, filter lengths, the number of fully connected layers, and the number of units in each fully connected layer. Among the 180 model candidates with different hyperparameter configurations, we selected the model that showed the minimum average validation loss as the final model for Deep-Learning model.

III. Result

1. Limitation of current off-target profiling methods

We first investigated GUIDE-seq which is one of the reported off-target profiling methods[15]. GUIDE-seq is a way to directly analyze the RGEN-induced double stranded break (DSB) region, since double-stranded oligodeoxynucleotide (dsODN) is embedded in the RGEN-induced DSB region in the genome. Therefore, this method is the most sensitive and accurate method of the all off-target profiling methods and has been cited in many groups. Therefore, we focused on GUIDE-seq. We investigated the number of off-targets of 10 target sequences which are used in GUIDE-seq based on the potential off-targets. There is a web-site which shows all potential off-targets per one target sequence[26]. We compared the number of off-targets detected by GUIDE-seq to Cas-OFFinder. Although the off-targets found in Cas-OFFinder are not real site in genome, GUIDE-seq detected very small number of off-targets of potential off-targets (Figure 1, Table 1) . GUIDE-seq finds the most off-targets among the currently reported methods. Therefore, this indicates that the current methods have still limitation to represent all the genome-wide off-targets.

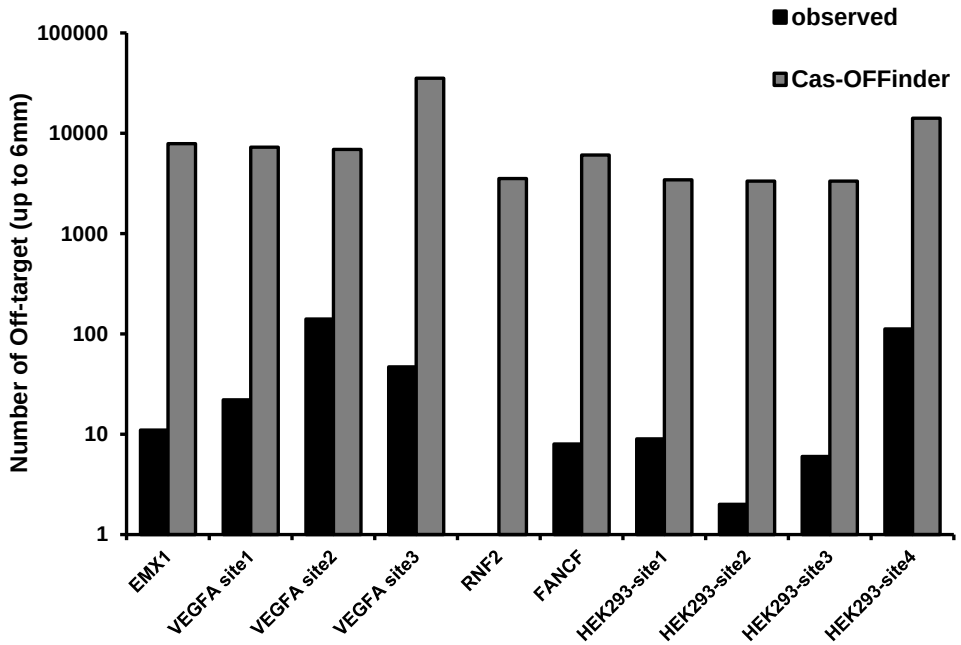


Figure 1. Relative rate of the number observed off-targets in GUIDE-seq to the potential off-targets in Cas-OFFinder. The number of off-targets of 10 guides used in GUIDE-Seq indicates as black bar, and the number of potential off-targets in Cas-OFFinder indicates as grey bar. (http://www.rgenome.net/cas-offinder/) up to 6 mismatches, no bulge.

Table 1. Comparison of the number of potential off-targets with observed off-targets in GUIDE-seq and Cas-OFFinder

	observed	Cas-OFFinder
EMX1	**11**	**7877**
VEGFA site1	**22**	**7267**
VEGFA site2	**141**	**6914**
VEGFA site3	**47**	**35471**
RNF2	**0**	**3541**
FANCF	**8**	**6082**
HEK293-site1	**9**	**3437**
HEK293-site2	**2**	**3336**
HEK293-site3	**6**	**3345**
HEK293-site4	**112**	**14134**

2. Development of a high-throughput guide RNA target paired library

A. Choosing potential off-targets

We previously developed a library that has guide RNA-target pairs and we applied this concept to profiling off-targets of CRISPR-Cas9. First, we searched potential off-targets of 10 target sequences which had been used in GUIDE-Seq in Cas-OFFinder. We set the searching conditions with NGG PAM for spCas9 and up to six mismatches and no bulge. GUIDE-seq detected off-targets up to 6 mismatches, therefore we set the highest mismatch number as 6 (Table 2). 'Bulge' sites have a skipped position at the sgRNA-protospacer interface. We also tried to consider the bulge. However, when we analyzed the results, our prediction tool could not classify the off-targets by mismatch and the off-targets by bulge. For this reason, we excepted this condition to avoid the overlap with off-targets casued by mismatches. And the chance of the off-target by the bulge was rare, therefore we decided to not consider the bulge and we could focused on the off-targets by Mismatches. We chose about 1,000 to 1,200 potential off-targets of each target sequence in ascending order of the mismatch number. If the number of potential off-targets was more than 1,000 before 6 mismatches, we chose potential off-targets as fair as possible. We set the rule that we included all off-targets which had been detected by GUIDE-seq.

Table 2. Potential off-target of 10 guides used in GUIDE-Seq

target site	target sequence	Potential Off-target		
		1~4mismatch	5mismatch	6mismatch
EMX1	GAGTCCGAGCAGAAGAAGAAGGG	87	966	6824
VEGFA site1	GGGTGGGGGGAGTTTGCTCCAGG	151	942	6174
VEGFA site2	GACCCCCTCCACCCCGCCTCCGG	102	1596	5216
VEGFA site3	GGTGAGTGAGTGTGTGCGTGTGG	5239	8220	22012
RNF2	GTCATCTTAGTCATTACCTGAGG	36	394	3111
FANCF	GGAATCCCTTCTGCAGCACCTGG	143	805	5134
HEK293 site1	GGGAAAGACCCAGCATCCGTGGG	50	399	2988
HEK293 site2	GAACACAAAGCATAGACTGCGGG	38	329	2969
HEK293 site3	GGCCCAGACTGAGCACGTGATGG	33	343	2969
HEK293 site4	GGCACTGCGGCTGGAGGTGGGGG	346	2086	11702

B. Generation of guide RNA and on/off target paired library

We designed a library of each target sequence. We developed a library that has guide RNA and on/off target pairs (Figure 2). We synthesized a pool of error-free oligonucleotides that contain a guide RNA sequence and the corresponding on- and off- target sequence on the same strand. The error-free oligo pools were PCR-amplified, and cloned into a lentiviral plasmid using Gibson assembly. After making at least 30 folds coverage plasmid libraries, we made cell libraries from HEK293T cells by lentiviral delivery. Then we also treated Cas9 lentivirus to the cell libraries, which led to guide-RNA-directed cleavage and indel formation in the target sequence integrated in the genome. When we previously tested our library with other nuclease, AsCpf1, it was highly correlated (R^2=0.7) with endogenous targets corresponding synthetic target sequence[37] (Figure 3). Therefore, the synthetic on- and off- target sequences are able to represent the endogenous sequences.

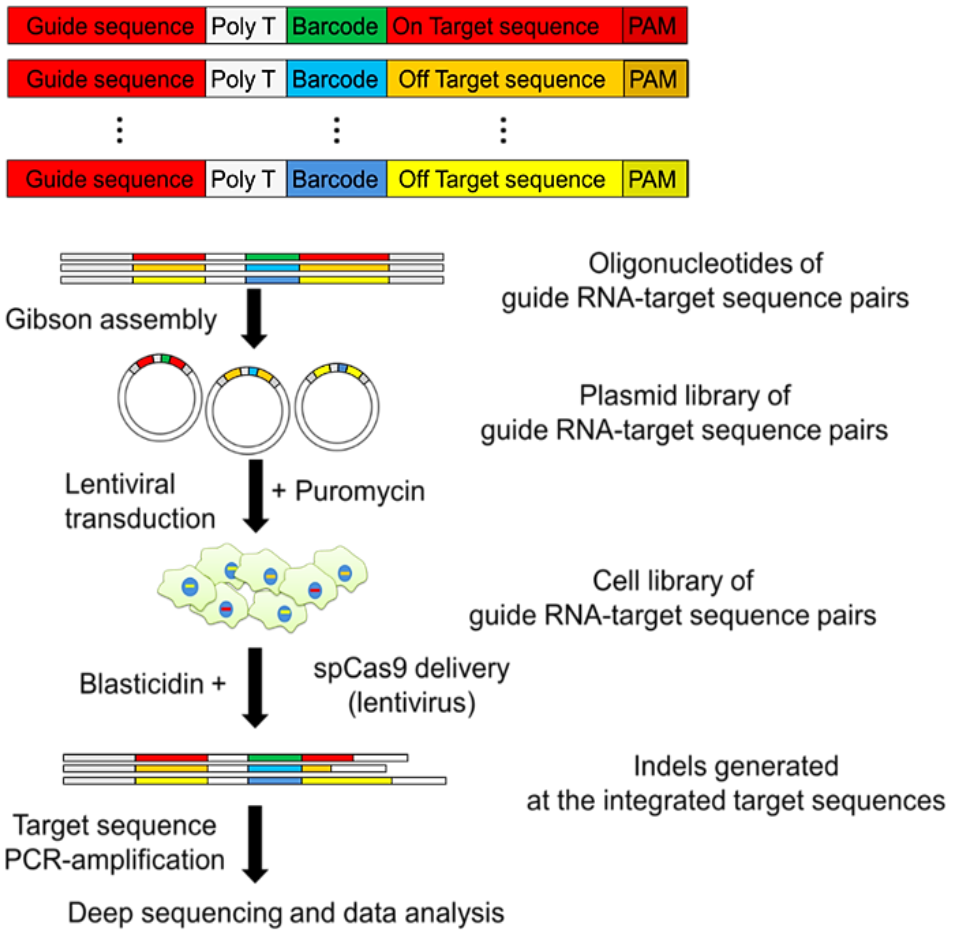


Figure 2. Overview of profiling of cas9 off-target. Construction of cas9 off-target library at each guide sequnce. Each library includes about 1000 potential off-targets in order of mismatch numbers.

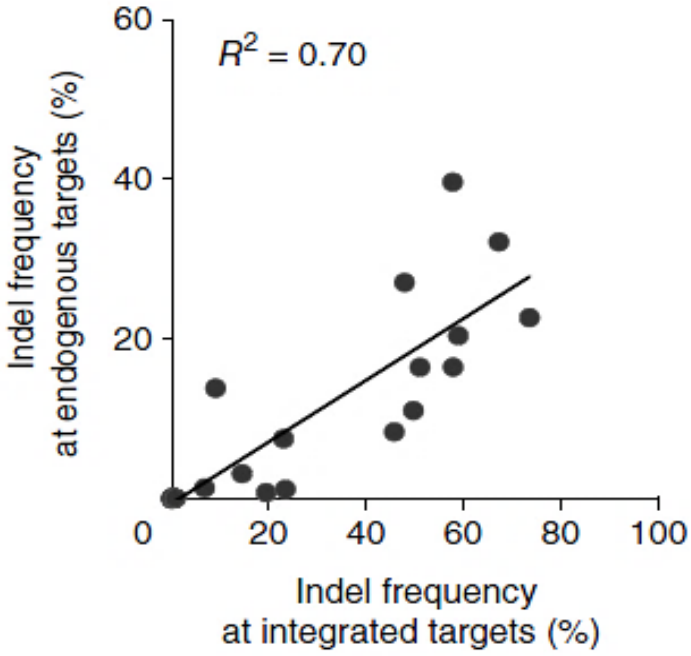


Figure 3. Correlation of previous library method with genome wide sequences. Correlation of integrated target sequences and endogenous target sites in Cpf1 library as we previous reported[37].

C. High-throughput targeted deep sequencing of off-target sites

After 7 days of treated lentiviral Cas9, we analyzed the indel frequency of synthetic on- and off- target sequences by targeted deep sequencing (Figure 4). We sequenced deeply specified on near target double stranded break regions. Each oligonucleotide of the library has the constant sequence at both 5’ end and 3’ end, therefore every guide-target pairs could be amplified with same primer. Our library also has unique barcode sequence of each target sequence. Therefore, we were able to sort exact guide-target pair twice first with both guide sequence and second with barcode sequence. Furthermore, we sequenced with paired end. For this reason, we could analyze indel frequencies correctly and deeply.

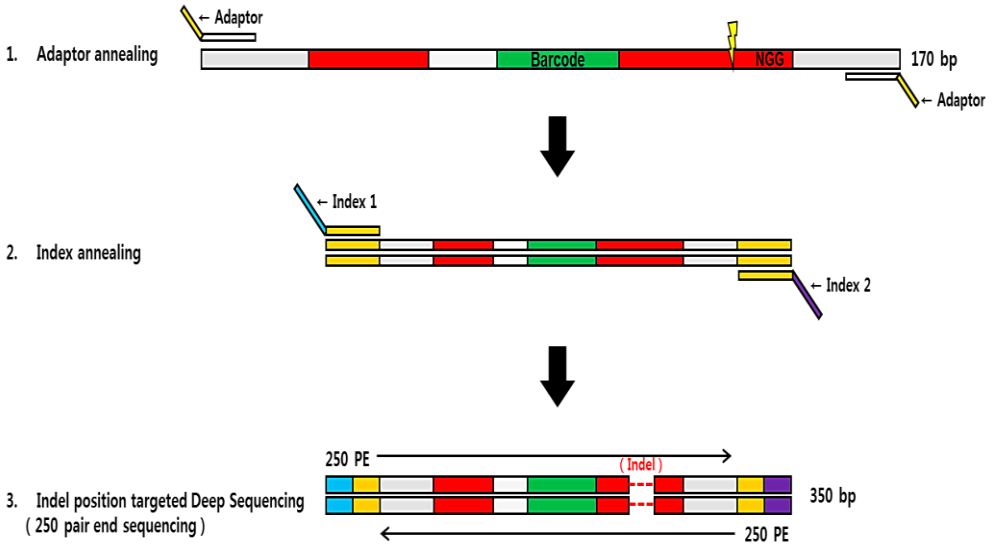


Figure 4. Overview of targeted Deep-sequencing. Mimetic diagram of Targeted Deep Sequencing. Paired end sequencing was able to raise the accuracy of indel analysis.

3. Verification of 10 on-targets activity

There are many factors that raise off-target effects. However, the most important thing is that high on-target activity is able to induce many off-target chances. And a goal of our research is to find as many off targets. For this reason, it is unnecessary to annalyze the off-targets of all 10 guides in terms of cost and labor. By the GUIDE-seq, the majority of off-targets of VEGFA site 2 and HEK239 site 4 were detected. We previous thought that we had to use these guide sequences for direct compare. However this approach is not appropppriate to profile novel off-targets. Therefore we changed our purpose of research to identify as many off-targets which had been not found. We first checked on-target activity of 10 guides with T7E1 assay. The best activity of on-target was FANCF (32%). Then we chose 3 guides having the best on-target activity, FANCF (32%), VEGFA site 1 (28%) and VEGFA site 3 (20%) (Figure 5). And we added a guide, RNF2, which GUIDE-seq could not detect an off-target. The on-target activity of RNF2 was poor. Therefore the reason that GUIDE-seq had not described the off-targets of RNF2 may be owing to little on-target activity.

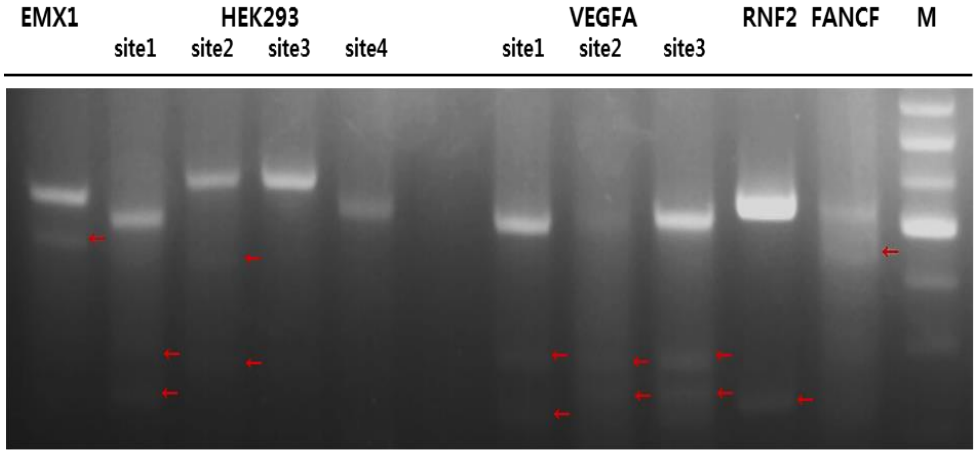


Figure 5. T7E1 assay for on-target activity of 10 guides used in GUIDE-Seq. Selection of guide. VEGFA1, VEGFA3 and FANCF has the best on-target activity in T7E1 assay (28%, 20%, 32%). The on target activity of RNF2 was 10%. The red arrow means the expected cleaved size by T7 endonuclease1.

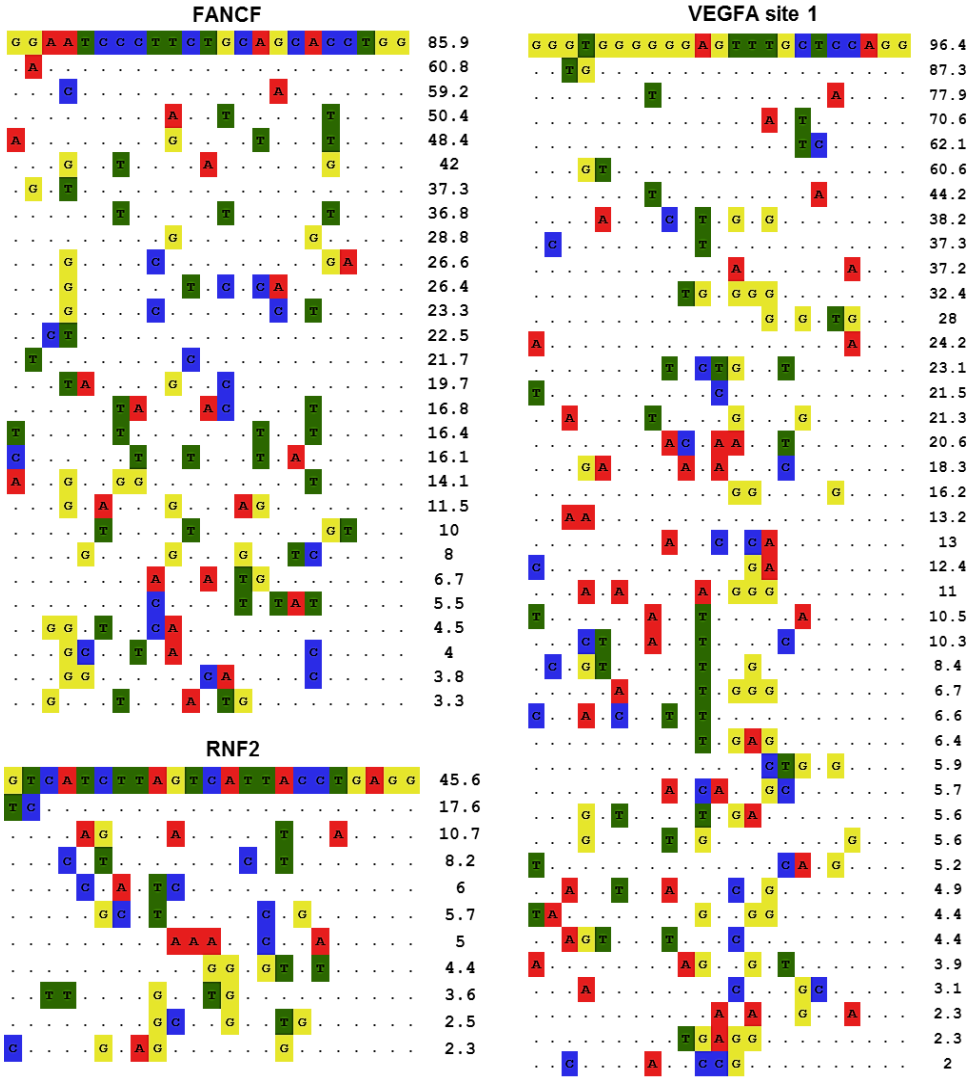
4. Genome-wide, off-target cleavage profiles in cells

A. Comparison of new method and GUIDE-Seq

We performed profiling off-targets of four different target sequences in HEK 293T cell. By analyzing the indel frequency of synthetic sequences, we identified the off-target sites. We extracted the off-target cleavage by the filtering with the difference of indel frequency between with and without Cas9 nuclease. We first screened the off-targets that had 0% of indel frequency of Cas9 untreated for the reason that avoid error pron oligos and back ground noise. Then we sought the off-targets on condition that indel frequency was up to 2% and a deep sequencing read count of both Cas9 treatment and Cas9 untreatment were over 100. And we investigated sequences of detected off-targets (Figure 6). Among the off-target sequences, we foud 248 off-targets of 4 target sequences. These off target sites harbored as many as 6 mismatches same as GUIDE-seq within the protospacer sequence. Interestingly, we detected the most off-targets of VEGFA site 3. However the best on-target activity was FANCF. It indicates that the on-target activity is not the only factor affecting the off-target effect.

We compared our method to GUIDE-Seq. We detected novel off-targets, not identified in GUIDE-Seq, and our method could demonstrate better genome wide off-targets than GUIDE-Seq although there are still remain undetected off-targets (Figure 7A). Unfortunately, our method could not cover all the off-targets found in GUIDE-Seq (Figure 7B). We hypothesized two reasons of this result. First, those off target sites might be undetected by our method, or second, it is possible that those sites were filtered by the condition of read count of either Cas9 untreated or Cas9 treated, or bacause of the back ground noise in spite of had been cleaved in cell. Therefore we propose that if the researchers want to profile the off-targets of interested guide RNA, they are able to find off-targets as many as possible in case of combining these two

methods, GUIDE-seq and our new method.



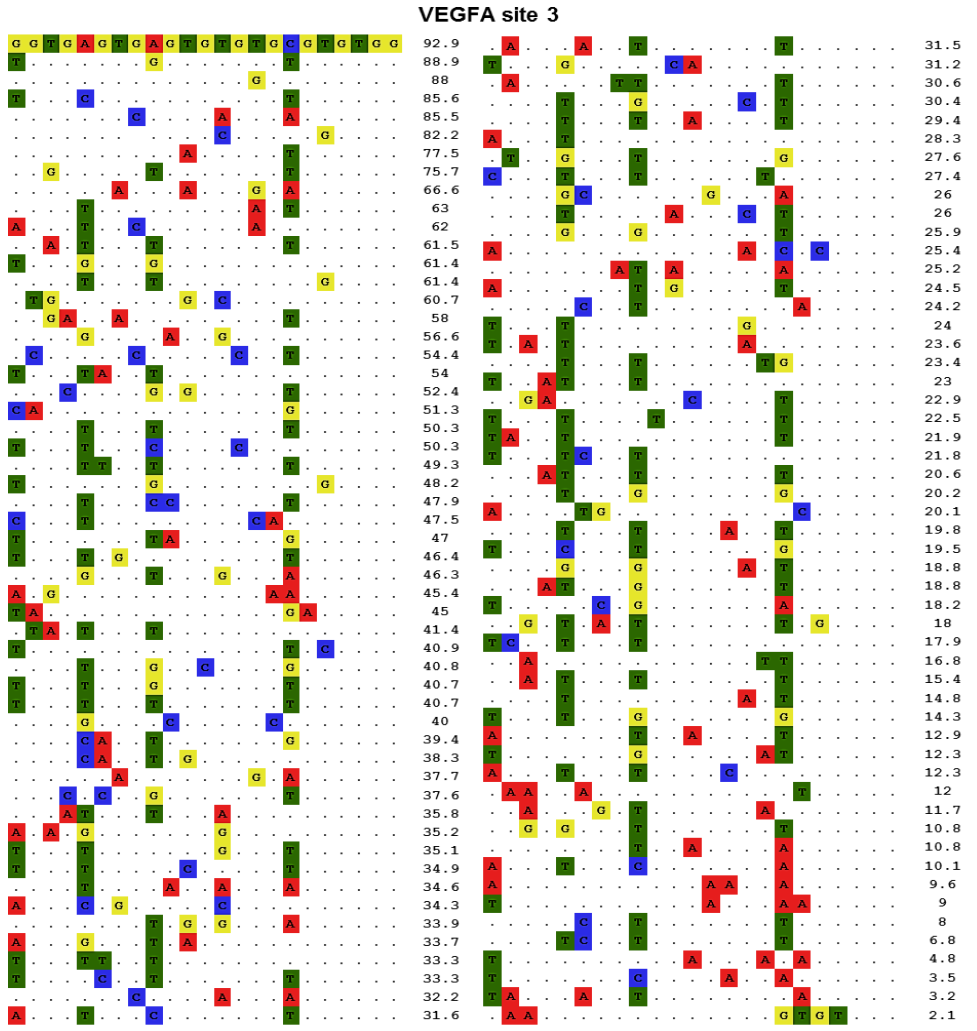


Figure 6. Identification the sequences of off-target sites. The on-target sequence is shown in the top line, and the mismatch nucleotides are highlighted in color in the below lines. The indel frequency is indicated on the right of each site. The PAM sequences of off-targets are not indicated, because we had fixed the pam with NGG**.** These off target sites harbored as many as 5 mismatches, within the protospacer sequence.

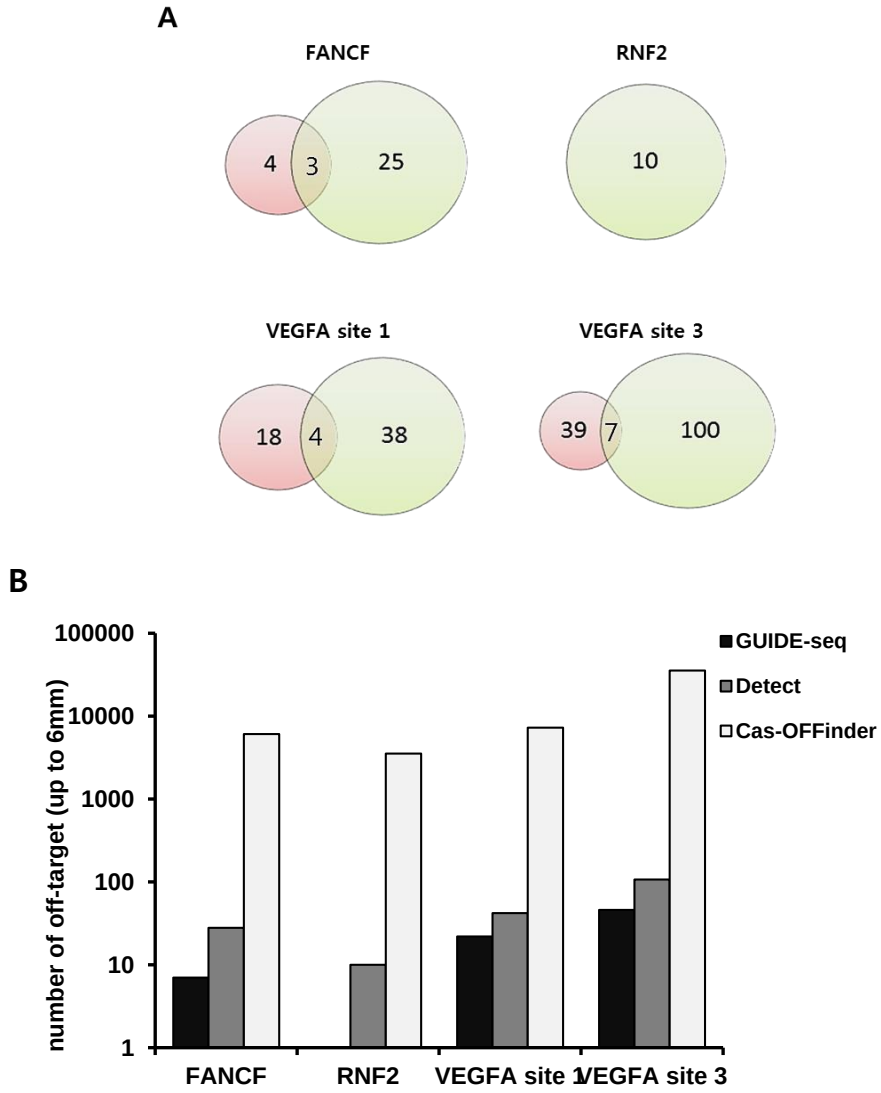


Figure 7. Comparison of new method with GUIDE-Seq. (A) Venn diagrams of overlap between off-target sites detected by the GUIDE-Seq and new method for 4 guides. (B) Comparison the number of detected off-targets of new method and GUIDE-Seq among the total potential off-targets up to 6 mismatches.

B. Relationship between Cas9 expression and detection of off-targets

We examined the effect of Cas9 delivery method on off-target measurement. We delivered Cas9 by either transient transfection of Cas9-expressing plasmid or transduction of Cas9-expressing lentiviral vector. 72 hours after transfection or 7days transduction, we analyzed indel frequency of on- and off-targets. We found that Cas9 delivery by transduction, more off-targets were found than when by transfection (Figure 8A). We sought that this difference was due to the Cas9 expression time. The exogenous expression of Cas9 plasmid, transfected in mammalian cells, is decreased over time whereas viral transduced Cas9 continuously expresses. When we measured the on-target activity of target sequence, the average indel frequency of 4 target sequences was 21.33 ± 3.25 % by transfected Cas9 and 80.2 ± 19.47 % by transduced Cas9. We also investigated whether the increasing cas9 expression time induced the number of detected off-targets. When we chose five off-targets with the lowest activity per target sequence and examined the indel frequency of off-targets at 7, 14 days and 21 days after Cas9 lentiviral delivery, the indel frequency of off-targets was increasing over time even though the indel frequency of the on-targets was saturating (Figure 8B). Therefore, we concluded that the longer time of cas9 expression, the more off-targets we could find with our method. However, since over-expression time of Cas9 saturates the on-target activity and there may be overfitting issues, we concluded that best cas9 expression time is 7 days. Although this condition reduces the number of off-targets that can be found, but is able to increase the accuracy to represent off-targets that may actually occur in the genome.

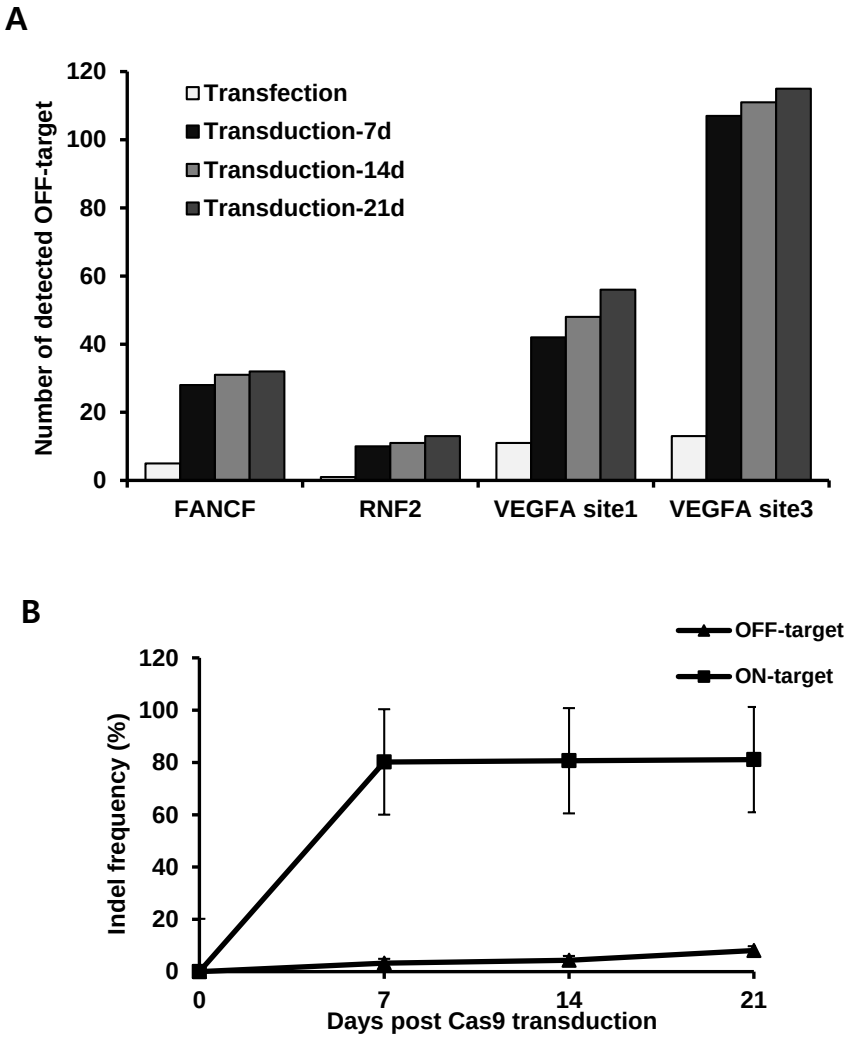


Figure 8. Relationship between Cas9 expression and detection of off-targets. (A)Comparison of Cas9 transfection and transduction. More off-targets were detected by Cas9 transduction than Cas9 transfection. (B)Comparison of the indel frequency change by Cas9 incubation time of on- and off-targets.

5. Analysis of off-target sequence characteristics

A. Characterization of detected off-targets

Our quantitative analysis of off-targets enabled us to hypothesize that the off-target effects could be impacted by variables such as mismatch number and position. We analyzed the relationship between activity and the number of mismatched bases. We found that RGENs could bear some mismatches in protospacer even detection rate was decreased as the number of mismatches was increased (Figure 9A). And we investigated the impact of mismatch position and off-target cleavages. We split the nucleotide positions by 5bp from 5' end. We sought that a mismatched position in protospacer was not associated with off-target cleavages (Figure 9B). And we also identified the frequency of indel pattern. When we analyzed the pattern of indels either deletion or insertion of detected off-targets, deletion occurred more often than insertion in double stranded break region(Figure 9C). This indel pattern rate is able to demonstrate the whole RGEN systems. When DNA strand repairs from break, joining of two strands occurrs more often.

We also analyzed endogenous off-targets corresponding detected synthetic off-target sequences. Off-target sequences were distributed throughout the genome in exon and intron (Figure 10A). This result is important of the reason that alteration of CDS region can influence to gene expression. Chromatin accessibility has been one of critical issues of RGEN activity in genome. Some reports suppose that open chromatin has more chances to bind with RGEN than nucleosome. We could define chromatin accessibility by profiling DNase hyperesensitivity (DHS). When we classified the endogenous sites, corresponing on detected off-target sequences, to chromatin accessibility in HEK293 T cell line, they were distributed around both DHS region, or non DHS region (Figure 10B). We are able to apply this data to our

prediction model development. Therefore, we are able to analyze the influence of chromatin accessibility to off-target effect.

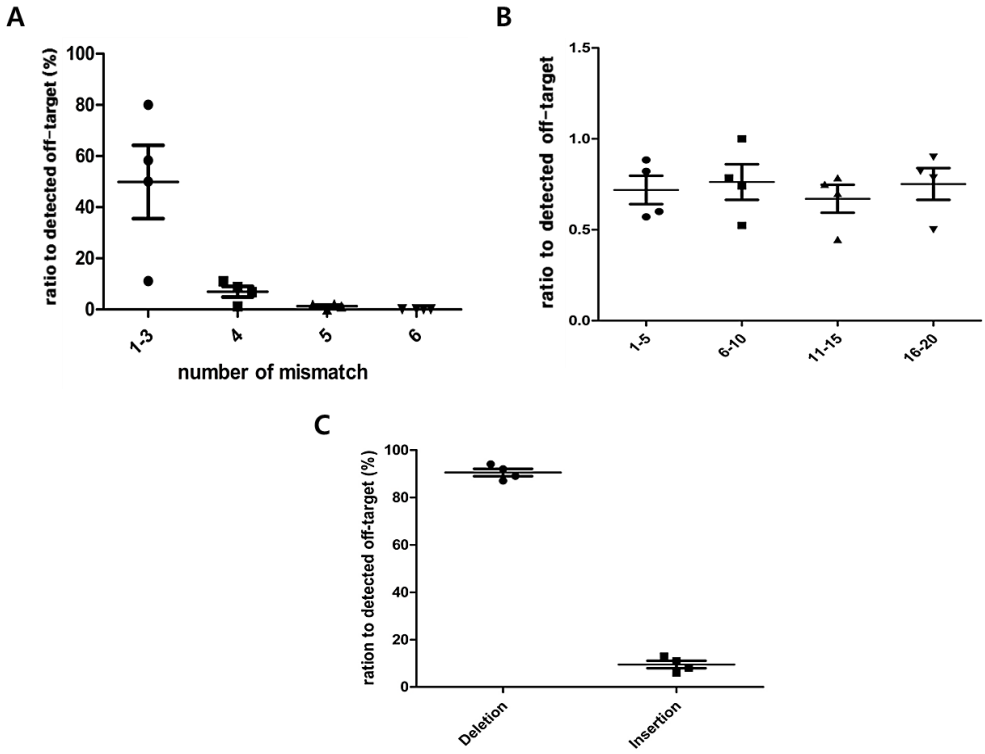


Figure 9. Characterization of detected off-targets. (A)(B)(C) each dots indicate 4 guides;FANCF, RNF2, VEGFA site1, VEGFA site3 (A) Ratio of off-targets according to the number of mismatch to total number of detected off-targets for 4 guides. (B) Effects of mismatch position within the protospacer of detected off-targets. It is indicated by ratio to total detected off-targets (the number of off-targets of each mismatch position/total detected off-targets). (C) Analysis of indel pattern of the detected off-targets.

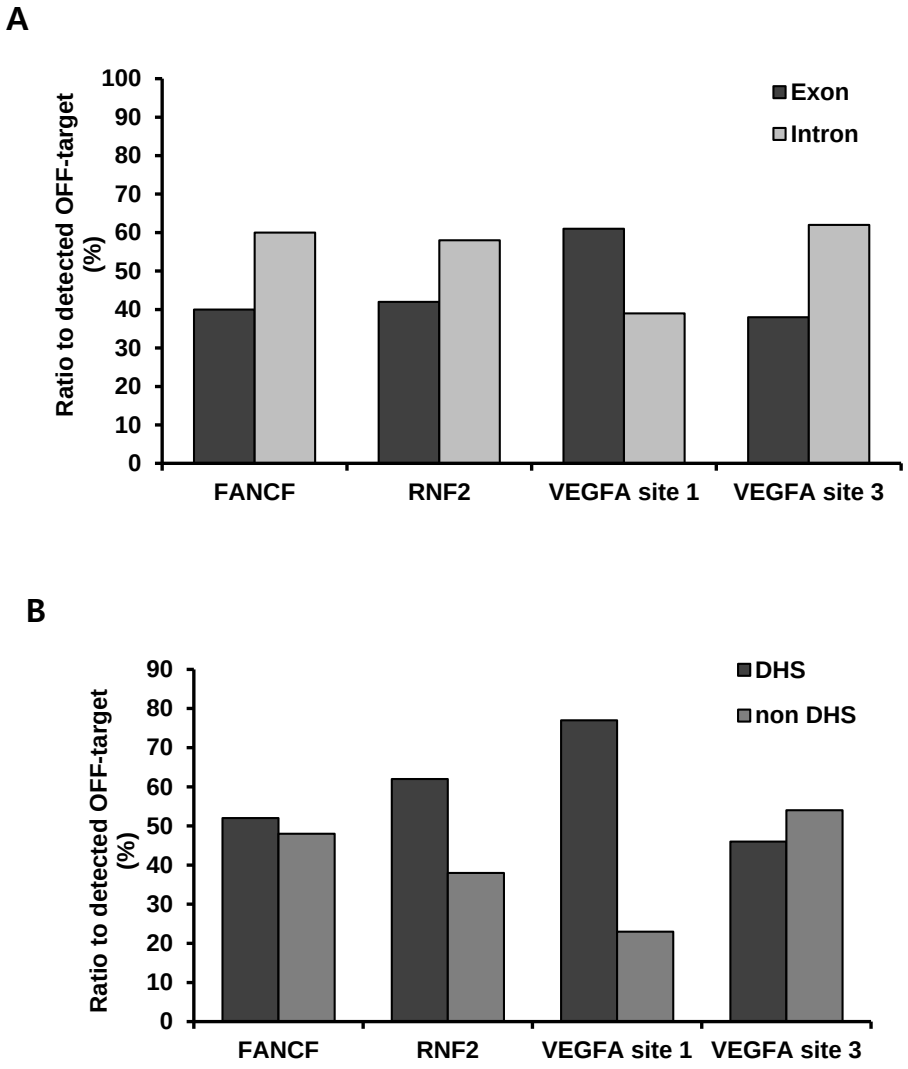


Figure 10. Analysis of edcogenous off-targets, corresponding detected synthetic off-target sequences. (A) Profiling the genomic location of the detected off-target sequences. (B) Classification of the detected off-target sequences in chromatin accessibility in HEK 293T cell line.

B. Correlation to endogenous off-target sites

We previously checked that our library system significantly correlated to endogenous target sites with Cpf1, another nuclease. We also checked our method with Cas9. When we analyzed our method, we found that the off-targets found in our method correlated (R^2=0.4765) with the off-tagets existing in the genome (Figure 11). Although this method was not highly correlated yet, however we expect that the correlation might be better if it is tested with more data; different target sequences. Therefore, we could conclude that our method could demonstrate the real off-targets in the genome. This result is important to develop our method to a new prediction tool. Therefore, we sought to create a new prediction tool using these well defined data, good representation of off-targets in the genome.

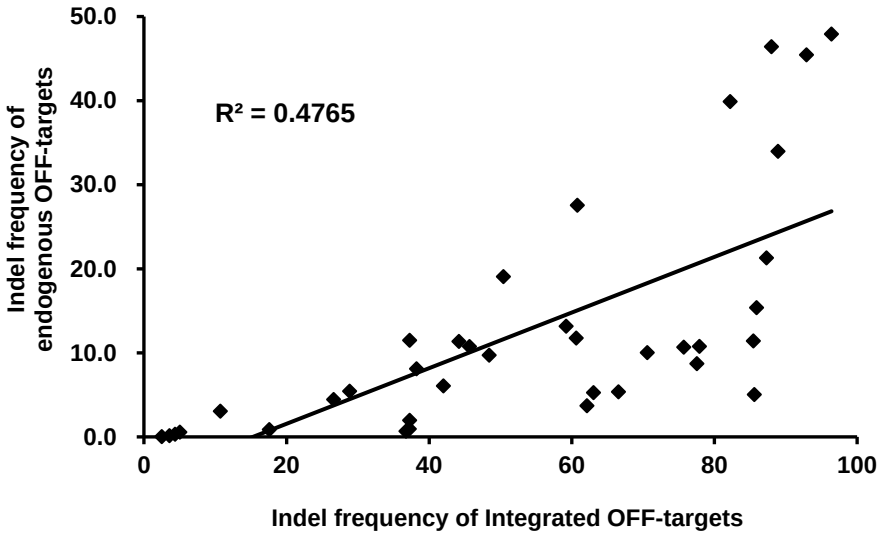


Figure11. Correlation of endogenous off-target sites and integrated off-target sites. As previously mentioned, we analyzed endogenous off-target sites corresponding integrated off-target sites.

6. Development of off-target prediction model based on Deep-Learning

We developed a Deep-Learning based Regression model (Figure 12). The new Deep-Learning method is based on a convolutional neural network for cas9 off-target prediction. A convolutional neural network is a type of feed-forward artificial neural network. This model predicts the off-target, using the following four stages. (1) The one-hot encoding input layer converts the sequence, into numerical representations for downstream processing. It encodes the nucleotide in each position, as a four-dimensional binary, in which each element represents the type of nucleotide: A, T, G and C. (2) The convolution layer performs one-dimensional convolution operations, with 80 filters of length 5. The filters, slide along only one axis of the one-hot encoded matrix, containing the 4-nt channels. The convolution layer, then applies the RLU nonlinear function, to the convolution outputs. the pooling layer computes the average of the values, providing invariance to local shifts. (3) our model uses three fully connected layers, with 80, 40, and 40 units. Each unit in the fully connected layers, performs linear transformations of the outputs of the previous layer, and applies the RLU nonlinear function. Multiple nonlinear layers, enable the model to increase levels of abstraction. (4) and then, The output layer makes the prediction of relative off-target activity of Cas9.

To demonstrate the reliability of the model selection process, we conducted nested cross-validation with data set (Figure 13). In each fold of the outer ten fold cross-validation, we randomly constructed training data sets to evaluate the performance improvements associated with training data sets of different sizes. After we split a data set to 10 inner separate data set, we used 9 data sets to training and used a remaining data to test. We repeated this 10 times with each data set.

We trained our model with several independent reported data sets. We collected data sets from CIRCLE-seq, GUIDE-seq and Digenome-seq. We got 9 guides datasets from GUIDE-seq, except RNF2, and 11 guides datasets from CIRCLE-seq and Digenome-seq. CIRCLE-seq and Digenome-seq used same guides, 10 guides used in GUIDE-seq and HBB guide. The differrences of these methods are the number of detected off-targets and cleavage situation. GUIDE-seq detected off-targets in cell level, however CIRCLE-seq and Digenome-seq detected *in vitro*. For this reason, GUIDE-seq datasets include chromatin accessibility of off-targets. We designed each dataset for cross-validation of our model. First, we searched potential off targets of on target sequences, up to 6 mismatches in Cas-OFFinder. And second, we classified potential off targets into active, which had been detected to real off-targets in genome, and non-active which had not been detected (Table3). And then, we performed 'leave one-guide out cross validation'. We trained model with 9 or 10 guides datasets and tested model with 1 independent guide datasets. We repeated this with each guide dataset. The performance of our model was best in CIRCLE-seq data sets (Figure 14). We can not find the reason yet. We assumed that the difference of performance might be because of size of datasets or quality of datasets. CIRCLE-seq detected a lot of off-targets because of *in vitro* system, even though the detected off-targets were not able to demonstrate genome. This result indicates that Deep-Learning model could affected by the size of dataset.

To compare the performance of our prediction tool with other tools previously reported, we checked other cas9 off-target prediction tools, CCTOP, CFD and Hsu Zhang_MIT. We used datasets from CFD paper[24]. CFD paper made off-target libraries of independent gene set, H2-D/H2-K and CD33 (Table 4). We changed nucleotides on target DNA sequence, however CFD mutated nucleotides on guide RNA sequence. This paper showed score of CCTOP, CFD and MIT with H2-D/H2-K data set.

Therefore we could directly compare our model to these tools. We reproduced these three algorithms nearly same as reported (Figure 15). And we pre-tested our deep-learning model. We trained our model with CD33 data sets and tested our model with H2-D/H2-K data sets. When we tested our model, our deep-learning model showed best performance than other prediction tools (Figure 16).

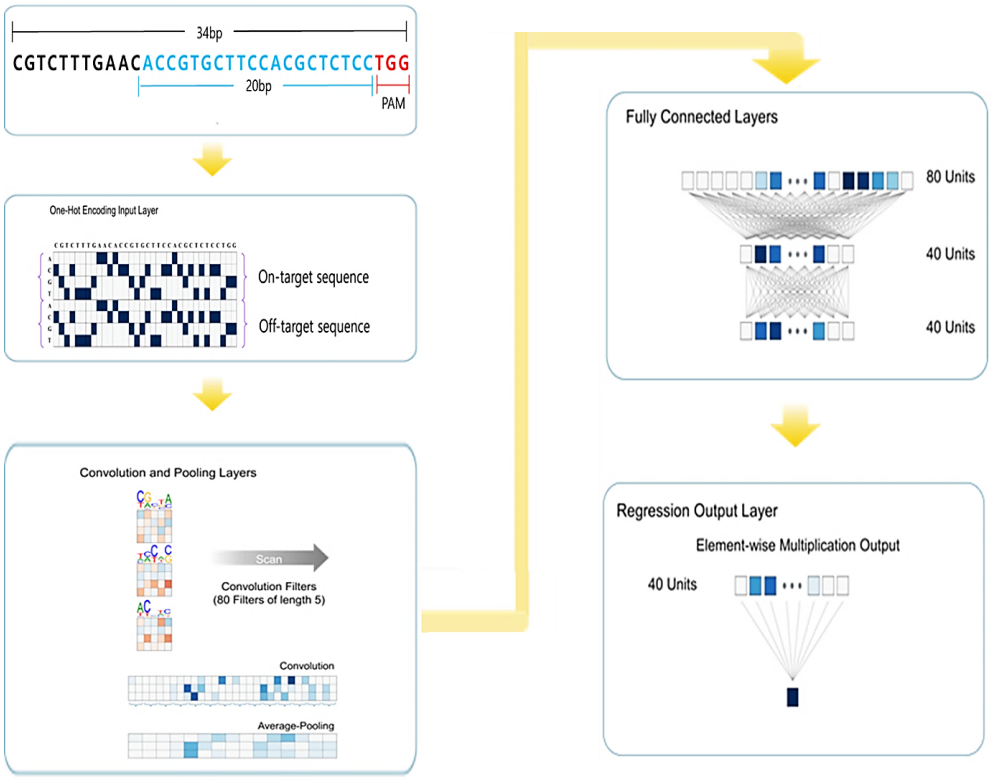


Figure 12. Overview of new Deep-Learning based off-target prediction. The new Deep-Learning method is an end-to-end deep learning framework based on a convolutional neural network (CNN) for cas9 off-target prediction. This model predicts the off-target using the following four stages. (This data was from co-work with Seonwoo Min in Seoul National university.)

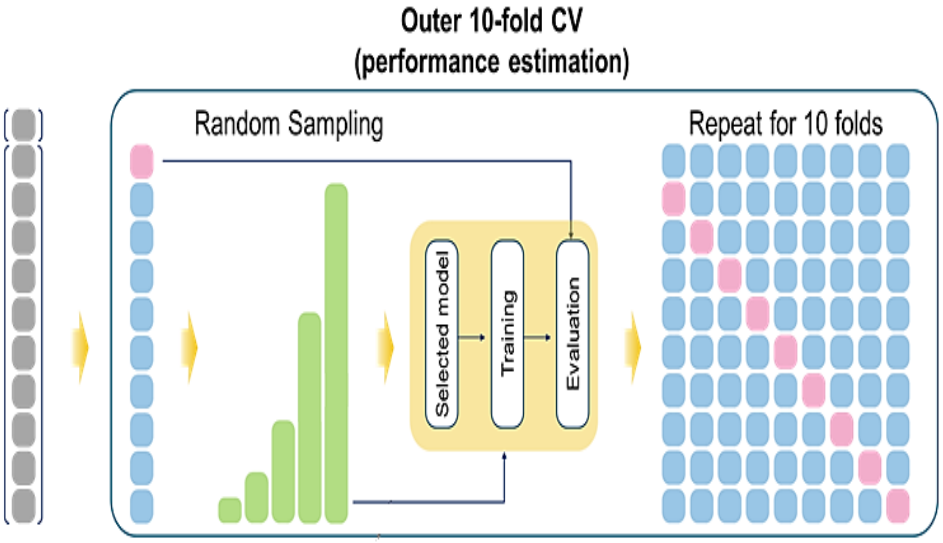


Figure 13. Diagram of nested cross-validation (CV). In each fold of the outer ten fold cross-validation, we randomly constructed training data sets (This data was from co-work with Seonwoo Min in Seoul National university.)

Table 3. Data sets used in training of model.

Method	Dataset	
	active	non active
CIRCLE seq (11 guides)	**6634**	**66078**
GUIDE seq (9 guides)	**402**	**63929**
Digenome seq (11 guides)	**963**	**68618**

( ) indicates the number of guides wich are used in each method.

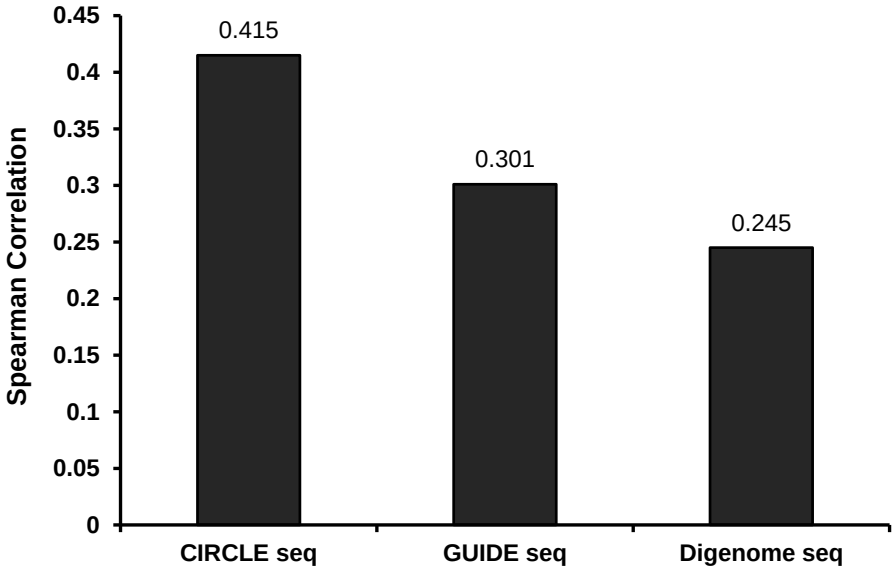


Figure 14. Leave-one-guide-out cross-validation. Performance of cross-validation of each datasets. (This data was from co-work with Seonwoo Min in Seoul National university.)

Table 4. The number of Datasets wich were obtained from CFD paper

Data source	Datasets
CD33_Doench2016	**3826** (65 guides)
H2-D/H2-K	**89**

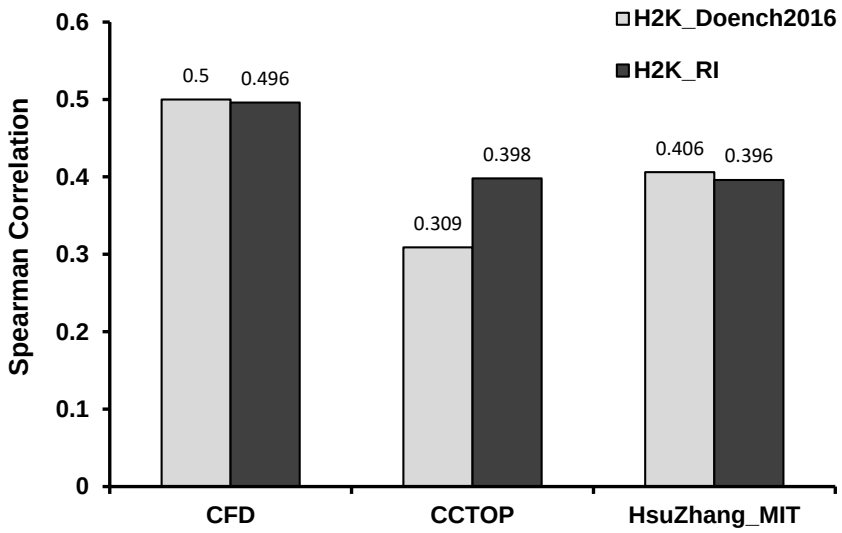


Figure 15. Reproduction of other Cas9 off-target prediction model. Reproduction of other tools previoulsy reported;CCTOP, CFD and Hsu Zhang_MIT. (This data was from co-work with Seonwoo Min in Seoul National university.)

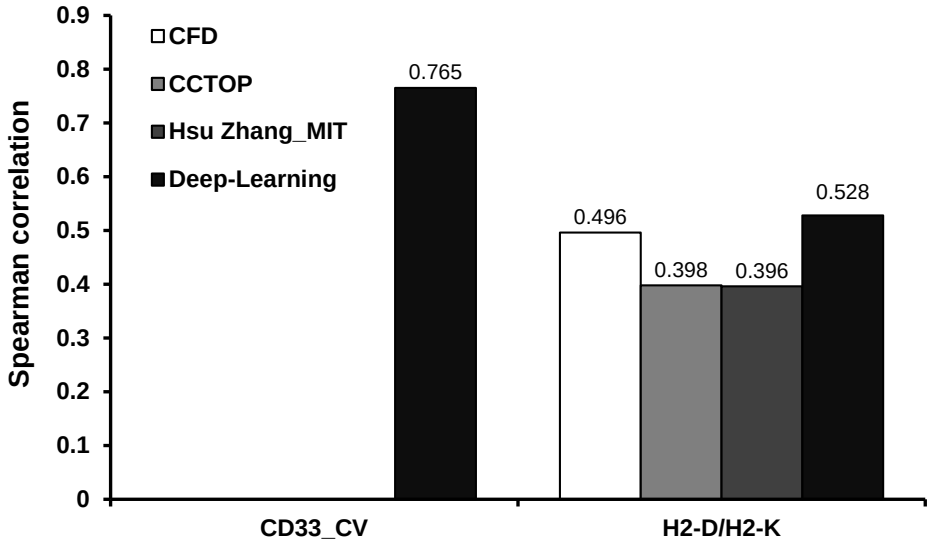


Figure 16. Performance comparison of new Deep-Learning based off-target prediction model with other prediction models. The bar graph shows Spearman correlations between measured off-target activity and predicted activity scores.

IV. DISCUSSION

CRISPR–Cas9 RNA-guided nuclease is a transformative technology for biology, genetics and medicine owing to the simplicity with which they can be programmed to cleave specific DNA target sites in organisms. However, it is seriously important to carefully define and improve genome-wide specificities of RGENs to apply into safe and effective clinical applications.

Therefore, there have been many challengies to identify and minimize off-target mutations induced by RGENs. A number of novel strategies to define and improve the genome-wide specificities of CRISPR–Cas9 nucleases have been published over the past few years. Although these worth metods are available for detecting and reducing off-target effects, it is still difficult to use in different experimental and therapeutic trials.

Our new method was able to identify genome-wide off-target sites more comprehensively than any other method. One limitation of our method is that we ranged the off-targets, therefore it could detect the off-targets only in library. For this reason, it is impossible to conclude that our method is unbiased. And there is another limitation of our method compared to GUIDE-seq. The main advantage of GUIDE-seq is that dsODNs are able to directly integrated in the DSB regions generated by Cas9 in the genome. Therefore, GUIDE-seq method is no need to be validated, unlike our method. However, there are more advantages of our method than limitation. Since GUIDE-seq is more susceptible to dsODN being inserted, it might not be able to detect the region which dsODN not integrate in despite of been cleaved by Cas9. Therefore, if we confirm the correlation with the endogenous off-target sites, our method might be simpler and more accurate than the GUIDE-seq. Our method is easy to profile and is more comprehensive, therefore it will be good assistant to researchers

who want to identify off-targets.

Furthermore, we developed a off-target prediction model based on Deep-Learning. There have been many prediction tools. However, they are not fair rool and just classify the off-targets without score. It is the first trial to predict off-targets with both classifying and scoring by Deep-Learning model. There still remains more validation and stabilization. We need to fine-tune the model with chromatin accessibility, gene expression and cell type specificity. We are sure the more training with lots of datasets, the better our method will performs.

We expect that our overall approach will prove to be very useful for the evaluation of off-target mutations and genomic rearrangements induced by RGENs. Our new high-throughput method most likely be extended for use in any cell in which NHEJ is active. This strategy can be used as part of a useful preclinical pathway for objectively assessing the potential off-target effects of any RGENs proposed for therapeutic use, thereby substantially improving the prospects for eventual translation of RGENs to the clinical region.

V. CONCLUSION

Genome editing with programmable nucleases (CRISPR-Cas9) is one of the most promising tool in the field of research and biomedical applications. Cas9 nuclease makes double-strand breaks (DSBs) at desired targeted sites in the genome and edits target region of target DNA. However, there is serious difficulty of using RGENs in genome therapy owing to off-target effects. In this study, we developed our own method to identify and predict the off-targets which are iduced by spCas9. And when we compared our method to others, our finding is more efficient than oother methods. Furthermore, our new tool is the first trial to predict off-targets using Deep-Learning model. Predicting off-target effects is essential to gene therapy application to avoid oncogenic mutation and unwanted mutation. Our High-throughput tool will be one of the solutions of clinical issues of CRISPR-Cas9.

REFERENCES

1. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA. et al. CRISPR RNA maturation by transencoded small RNA and host factor RNase III. Nature 471, 602–607 (2011).

2. Kim H. Kim JS. A guide to genome engineering with programmable nucleases. Nat. ReviewGenet. 15, 321-334 (2014).

3. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK. et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat. Biotechnol. 31, 822–826 (2013).

4. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. 31, 827–832 (2013).

5. Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat. Biotechnol. 31, 839–843 (2013).

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

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

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

(2014).

9. Ran FA, Hsu PD, Lin CY, Gootenberg JS, Konermann S, Trevino AE. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell 154, 1380–1389 (2013).

10. Mali P, Aach J, Stranges PB, Esvelt KM, Moosburner M, Kosuri S. et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat. Biotechnol. 31, 833–838 (2013).

11. Tsai SQ, Wyvekens N, Khayter C, Foden JA, Thapar V, Reyon D. et al. Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat. Biotechnol. 32, 569–576 (2014).

12. Guilinger JP, Thompson DB, Liu DR. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat. Biotechnol. 32, 577–582 (2014).

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

14. Kim S, Kim D, Cho SW, Kim J, Kim JS. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. Genome Res. 24, 1012–1019 (2014).

15. Sung YH, Kim JM, Kim HT, Lee J, Jeon J, Jin Y. et al. Highly efficient gene knockout in mice and zebrafish with RNA-guided endonucleases. Genome Res. 24, 125–131 (2014).

16. Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V. et al. GUIDE-

seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. Nat. Biotechnol. 33, 187–197 (2015).

17. Kim D, Bae S, Park J, Kim E, Kim S, Yu HR. et al. Digenome-seq: genome-wide profiling of CRISPR–Cas9 off-target effects in human cells. Nat. Methods 12, 237–243 (2015).

18. Kim D, Kim S, Kim S, Park J, Kim JS. Genome-wide target specificities of CRISPR–Cas9 nucleases revealed by multiplex Digenome-seq. Genome Res. 26, 406–415 (2016).

19. Wang X, Wang Y, Wu X, Wang J, Wang Y, Qiu Z. et al. Unbiased detection of off-target cleavage by CRISPR–Cas9 and TALENs using integrase-defective lentiviral vectors. Nat. Biotechnol. 33, 175–178 (2015).

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

21. Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ. et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature 520, 186–191 (2015).

22. Yan WX, Mirzazadeh R, Garnerone S, Scott D, Schneider MW, Kallas T. et al. BLISS is a versatile and quantitative method for genomewide profiling of DNA double-strand breaks. Nat. Commun. 8, 15058 (2017).

23. Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q. et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat. Methods 10, 361–365 (2013).

24. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF. et al.

Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9. Nat. Biotechnol. 34, 184–191 (2016).

25. Listgaten J, Weinstein M, Kleinstiver BP, Sousa AA, Joung JK, Crawford J. Et al. Prediction of off-target activities for the end-to end design of CRISPR guide RNAs. Nat. Biomedical Eng. 2, 38-47 (2018).

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

27. Haeussler M, Schönig K, Eckert H, Eschstruth A, Mianné J, Renaud JB. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. 17, 148 (2016).

28. Labun K, Montague TG, Gagnon JA, Thyme SB, Valen E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. Nucleic Acids Res. 44, W272–W276 (2016).

29. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat. Methods 11, 122–123 (2014).

30. Ma J, Köster J, Qin Q, Hu S, Li W, Chen C. et al. CRISPR-DO for genome-wide CRISPR design and optimization. Bioinformatics 32, 3336–3338 (2016).

31. Singh R, Kuscu C, Quinlan A, Qi Y, Adli M. Cas9–chromatin binding information enables more accurate CRISPR off-target prediction. Nucleic Acids Res. 43, e118 (2015).

32. Cradick TJ, Qiu P, Lee CM, Fine EJ, Bao G. COSMID: a web-based tool for identifying and validating CRISPR/Cas off-target sites. Mol. Ther. Nucleic Acids 3,

e214 (2014).

33. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. 31, 827–832 (2013).

34. Stemmer M, Thumberger T, del Sol Keyer M, Wittbrodt J, Mateo JL. CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. PLoS ONE 10, e0124633 (2015).

35. Kim HK, Min S, Song M, Jung S, Choi JW, Kim Y. et al. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. Nat. Biotechnol. (2018)

36. Tsai SQ, Nguyen NT, Lopez JM, Topkar VV, Aryee MJ, J Keith Joung. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. Nat. Methods 14, 607-614 (2017)

37. Kim H, Song M, Lee J, Menon V, Jung S, HK Kim et al. In vivo high-throughput profiling of CRISPR–Cpf1 activity. Nat. Methods 14, 153-159 (2017)

ABSTRACT (IN KOREAN)

유전체 내에서 CRISPR-Cas9 에 의한 비특이적 절단을 찾아내기 위한 새로운 방법의 개발

(지도교수 김형범)

연세대학교 대학원 의과학과

정수빈

유전자가위 CRISPR-Cas9은 현재 생물학분야에서 다양한 방법으로 많이 사용되고 있고 개발되고 있는 새로운 유전자교정 방법이다. 하지만 유전체 상에서 CRISPR-Cas9에 의해 발생하는 의도치 않은 비특이적 절단은 아직 잘 규명되어지지 않았다. 이러한 문제는 유전자가위를 임상분야에서 응용하려고 할 때 가장 우려되는 한계점 중 하나이다. 현재까지 이러한 Cas9에 의한 비특이적 절단을 찾아내는 여러 가지 방법들이 많이 보고되어 왔지만 상용화되기에는 많은 한계점을 가지고 있다. 따라서 우리는 이번 연구를 통해 CRISPR-Cas9에 의한 비특이적 절단을 좀 더 효율적이고 정확한 방법으로 검출하고, 더 나아가 인공지능을 바탕으로 한 컴퓨터상에서 Cas9에 의한 비특이적 절단을 미리 예측할 수 있는 방법을 개발하였다. 이러한 우리의 새로운 예측방법은 향후 생물학적 분야뿐 아니라 여러 질병을 치료하기 위해 유전자가위를 임상학적 분야에 사용하고자 할 때 연구자들에게 많은 도움을 줄 수 있을 것이다.

핵심되는 말: CRISPR-Cas9, 유전자 교정