Applying CRISPR-Cas9 Off-Target Editing on DNA based Steganography

Hong Zhou[1]
Department of Mathematical Sciences
University of Saint Joseph
West Hartford, Connecticut, USA

Xiaoli Huan[2]
Department of Computer Science
Troy University
Troy, Alabama, USA

***Abstract*—Different from cryptography which encodes data into an incomprehensible format difficult to decrypt, steganography hides the trace of data and therefore minimizes attention to the hidden data. To hide data, a carrier body must be utilized. In addition to the traditional data carriers including images, audios, and videos, DNA emerges as another promising data carrier due to its high capacity and complexity. Currently, DNA based steganography can be practiced with either biological DNA substances or digital DNA sequences. In this article, we present a digital DNA steganography approach that utilizes the CRISPR-Cas9 off-target editing such that the secret message is fragmented into multiple sgRNA homologous sequences in the genome. Retrieval of the hidden message mimics the Cas9 off-target detection process which can be accelerated by computer software. The feasibility of this approach is analyzed, and practical concerns are discussed.**

Keywords—DNA; steganography; CRISPR; Cas9; sgRNA; off-target; substitution

I. Introduction

Steganography is a technology that integrates the secret data into a common message seamlessly to avoid any attention to the data (note that data and message are two exchangeable concepts in this article). It is different from cryptography. If the data need to be secured from decryption, then cryptography technology is required. The power of cryptography is that even the encrypted message is left on the open ground, the meaning of the message cannot be revealed without the proper key. Its drawback lies in the fact that the data can catch attention. In contrast, steganography hides the message in a data carrier and makes minimal modifications to the carrier so that most people won't imagine any secret messages inside. However, hiding plain text messages via steganography is not recommended given the great advances in digital technologies. Today, steganography is used to place another layer of protection above cryptography in many applications.

Traditionally steganography makes use of either image, audio or video media. One of the most adopted steganography techniques is the Least Significant Bit (LSB) image steganography in which the least significant bits of the pixels of the cover image embed the secret message. The difference between the cover image and the stego-image (the modified image carrying the secret message) is imperceptible to human visual system [1]. Due to its complexity and large capacity, DNA, either digital DNA sequence, or DNA substance, is becoming another promising data carrier [2, 3, 4]. For example, human genome is about three billion base pairs which can store a large amount of information, making it a powerful data carrier in DNA based steganography.

II. Related Work

A single CPU of modern electronic computer works in a linear fashion, which limits its power in solving NP-complete problems when the size of the problem becomes large, for instance, the directed Hamiltonian path problem. However, in 1994, Leonard Adleman demonstrated that by applying step-by-step DNA biochemical reactions, a process like a computer science algorithm, the directed Hamiltonian path problem could be solved efficiently due to the large number of simultaneous DNA reactions [5]. Since then, DNA computing has become an eye-catching branch in computer science [2, 3, 6, 7, 8, 9, 10, 11, 12]. In 1999, Clelland et al. encoded a secret message as a DNA fragment (the secret-message DNA) and hid it among many other “junk” DNA fragments. Such DNA samples can be prepared as microdots to be delivered to recipients who can only retrieve the secret message by applying polymerase chain reaction (PCR) on the DNA sample with the knowledge of the two PCR primers followed by DNA sequencing [2]. This work started DNA based steganography. However, one drawback of this technique is that the recipient must be given the primer information. Leaking of the primer information can cause data insecurity. A recent study adds one extra layer of protection to the PCR primer by applying the trans-cleavage activity of CRISPR-Cas12a nuclease [13]. Note that CRISPR stands for Clustered Regularly Interspaced Short Palindromic Repeats, and Cas stands for CRISPR Associated System. In this study, fake primers or redundant DNA sequences were pre-ligated to the real PCR primers so that the real primers are concealed. To obtain the real primers, the recipient must apply CRISPR-Cas12a to cut off the fake primers or the redundant sequences [13].

A critical disadvantage of hiding the secret message into a biological DNA sample is the difficulty in preparing the sample and retrieving the secret message through a series of biological experiments. Such experiments are likely to be proceeded by experienced workers in labs armed with modern and expensive equipment. In addition, carrying biological samples may be considered illegal under some circumstances. Steganography based on digital DNA sequences however, provides much better feasibility.

There are three classical methods to hide data into DNA sequences, namely the insertion method, the complementary pair method, and the substitution method [4]. Several revisions of these methods were proposed, but the fundamental techniques keep the same [7, 9, 11]. All the methods require a binary coding rule to convert between binary digits and nucleotide bases (A, C, G, T). A key requirement of the substitution method is that a nucleotide can be substituted by another specific nucleotide, and the mapping is one-to-one [4]. Another requirement is that the length of the secret message must be no longer than the reference DNA (the data carrier) [4].

In the substitution method, the reference sequence R is known to the sender and the recipient along with many other public reference sequences. In the basic scenario, the message M is converted into a binary string and each bit is randomly but sequentially stored by a user-defined substitution function. The modified DNA sequence R’ is delivered to the recipient. The recipient retrieves the secret message by comparing R and R’ through the substitution function. The cracking of the substitution method requires the knowledge of the reference sequence R and the substitution function, and the cracking probability by a random guess is computed as $\frac{1}{6N}$, where N is the number of public reference sequences available [4]. Note that there are six different one-to-one substitution functions available [4], and currently there are about one billion publicly available reference DNA sequences [14].

This article presents an improved substitution method that can render a much lower cracking probability. This method is inspired by the ground-breaking biotechnology CRISPR-Cas9. It simulates the bacteria adaptive immune system and the evolutional arm-races between bacteria and phages.

CRISPR-Cas9 has become a widely adopted tool capable of gene editing, gene expression suppression/activation, and epigenetic modifications [15, 16]. Its gene editing function can usually be achieved by two approaches respectively. One approach introduces frame-shift insertions and/or deletions inside a targeted gene to generate gene-specific knock-outs. The success of this application depends on the error-prone non-homologous end joining (NHEJ) pathway that repairs the double strand DNA break introduced by the Cas9 nuclease. During the NHEJ repairing process, various mutations can be randomly generated and a few of them may result in loss-of-function knockouts [17]. The second approach relies on base-editor platforms which do not generate any double strand DNA breaks [18]. The beauty of CRISPR-Cas9 lies in the fact that it is programmable [15]. The core of CRISPR-Cas9 system includes the Cas9 endonuclease and a single guide RNA (sgRNA). sgRNA has two sequence regions, one is the scaffold to which Cas9 binds, the other is the spacer sequence of about 20 bases. The spacer sequence can be user defined and it can lead the Cas9 to any a genomic locus where the target DNA sequence matches the spacer and has a protospacer adjacent motif (PAM) immediately downstream [15, 19]. The PAM sequences vary depending on the types of Cas nucleases. The most commonly used Cas protein is SpCas9 whose PAM sequences are NGG (AGG, CGG, GGG, and TGG). A critical concern in the CRISPR-Cas9 technology is however, the sgRNA can also lead the Cas9 to other genomic loci that share sequence similarity with the sgRNA spacer, causing off-target genome editing [19]. This scenario is explained in Fig. 1 and 2.

It is understood that Cas9 off-target nuclease activity is likely a result of the evolutional arm-races between bacteria and infection virus. Bacteria that survived virus infection store characteristic virus genetic sequences as spacers between repeated sequences. These spacers can be transcribed into RNA to guide the Cas nuclease to degrade virus genetic elements containing the same sequence. This process provides bacteria an adaptive immune system to resist repeated phage invasion. In response to the selection pressure of the host bacteria, the phages’ genetic sequence evolves with variation(s) to escape Cas nuclease degradation. In return, bacteria CRISPR-Cas system evolves by allowing Cas nuclease to degrade sequences sharing certain degrees of homology with the spacer sequence.

The larger a genome’s size, the more the potential off-target loci for a given sgRNA spacer sequence. For genomes as large as the human genome, identifying the off-target sites becomes a time-consuming process. Different computational tools have been developed to help predict potential genome off-target loci [20, 21]. The early computational methods are mostly built upon the found sgRNA sequence features regarding SpCas9 [15, 22, 23]. However, some discoveries from different research groups are not in agreement with each other, though certain general understandings have reached consensus. 1) Off-target effect decreases when the number of mismatches (including both base mismatches and bulges) between sgRNA and target sequence increases; 2) Cas9 is less tolerant with mismatches proximal to PAM. Later methods incorporate Cas9 domain knowledge, especially energetics parameters, and therefore can achieve better predication results [24].

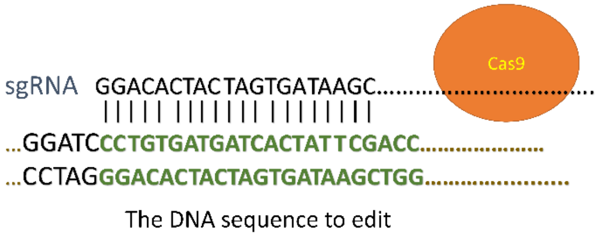
Cas9
sgRNA GGACACTACTAGTGATAAGC.................................
||||| |||||||| ||||||||
...GGATCCCTGTGATGATCACTATTCGACC.....................
...CCTAGGGACACTACTAGTGATAAGCTGG.......................
The DNA sequence to edit

Fig. 1. CRISPR-Cas9 on-Target Editing in which the sgRNA Sequence has a Perfect Match with the DNA Sequence.

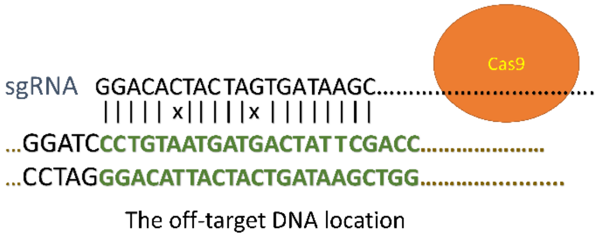
Cas9
sgRNA GGACACTACTAGTGATAAGC.................................
|||||x||||||x ||||||||
...GGATCCCTGTAATGATGACTATTCGACC.....................
...CCTAGGGACATTACTACTGATAAGCTGG.......................
The off-target DNA location

Fig. 2. CRISPR-Cas9 off-Target Editing in which the sgRNA Sequence is Homologous to the DNA Sequence.

III. Proposed Algorithm

Our proposed algorithm is based on the following assumptions regarding off-target homology search:

1) All off-target sites must have a primary NGG PAM immediately downstream the sgRNA spacer binding location.

2) All off-target sites can have up to five base mismatches within a given sgRNA spacer sequence. If there are at least six base mismatches, the DNA sequence in study is not considered an off-target homology.

3) Off-target sites cannot have indels, either DNA or RNA bulge. This assumption is against the existing biological discoveries, which will be discussed later.

In the proposed algorithm, the message recipient is the "bacteria" while the sender is the "phage". After some time of "evolution", the phage is aware of what sgRNA spacers the bacteria can recognize, a knowledge between the phage and bacteria only. For demonstration purpose, assume that the bacteria by far can recognize the following spacer sequence:

← ← ←

Position – 0 9 8 7 6 5 4 3 2 1 0 9 8 7 6 5 4 3 2 1

Spacer (S) – ACGTCGTAACGCGTATATGC

Using the same binary coding rule ((A:00), (C:01), (G:10), (T:11)), four nucleotide bases are required to define an eight-bit character. Thus, a 20-base sequence can code five characters. The substitution function is defined as:

1. T-A → 11
2. T-C, G-A → 10
3. T-G, G-C, C-A → 01
4. All other substitutions → 00

Note, T-A indicates that T substitutes for A. Suppose the secret message M is 01001110, let R be a large DNA sequence that has no homologous sequences of the spacer S (if there are, such homologous sequences must be either deleted or changed). To integrate M into R by substitution, the following algorithm is followed:

Step 1: Pick a large DNA sequence R, confirm there is at least one S inside R (must have PAM downstream). If there is none, substitute some nucleotides in R to generate S or select another valid R.

Step 2: compute the value of S by addition operation. The above sequence S = 30. Let v = (S mod 5) + 1 = 1. Only off-target sequences bearing exactly v mismatches with S would be used for carrying message. In this specific example, v = 1. Similarly, if v = 5, then valid off-target sequences must bear 5 mismatches.

Step 3: Change existing off-target sequences in R that bear exactly 1 mismatch with S so that they won't be mistaken as valid off-target sequences.

Step 4: Since M has eight bits, i.e. four nucleotides, four off-target sites are needed in R. Randomly but sequentially identify four non-overlapping sgRNA spacer sites in R, modify the site sequences according to the substitution function so that they become valid off-target sites of S. For instance, the message 01001110 can be fragmented into four off-target sites in order (PAM is attached):

Spacer: ACGTCGTAACGCGTATATGC

Site 1: ACGTCGTAACGCGTA**G**ATGC-GGG

Site 2: ACGTCGTAACG**G**GTATATGC-TGG

Site 3: ACG**A**CGTAACGCGTATATGC-CGG

Site 4: ACGTCG**C**AACGCGTATATGC-AGG

After the substitutions, carrier sequence R becomes R' and R' is sent to the recipient (bacterium) together with other noise DNA sequences (Note that these noise sequences do not contain S or any other bacteria-recognized CRISPR spacers). Once the bacterium receives R', it uses its existing CRISPR spacer sequence(s) to examine R'. If there is no recognized spacer in R', ignore it. In our case, as there is a recognized spacer S, the bacterium begins to process the R' sequence. The data retrieval reverses the protocol of data hiding.

IV. Results and Analysis

Unlike the classical substitution method, our method does not employ a one-to-one substitution function. We consider a one-to-one function is not necessary and one-to-one functions may present a more discernable pattern to intruders. The cracking probability of the above proposed method depends on the number of potential CRISPR spacers in the reference DNA sequence, the available substitution functions and binary coding rules. Since a spacer must be followed by an NGG PAM, the number of potential CRISPR spacers is about L/16, where L is the length of the reference sequence. The number of binary coding rules determines the available substitution functions, and it is $4 \times 3 \times 2 = 24$. Including the possible mismatch numbers, the cracking probability of the proposed method can be computed as:

$$\frac{16}{5 \times 24L} = \frac{2}{15L}$$

Thus, the cracking probability is a function of the size of the reference DNA sequence. This cracking probability is no better than that of the classical substitution method. To improve the cracking probability, we can remove the bacteria-recognized spacer S from R' because both the phage and bacteria are aware of the recognized spacers. In this revised scenario, to steal the secret message, the intruder must have knowledge of the spacer S. Therefore, the new cracking probability can be expressed as:

$$\frac{1}{120 \times 4^{20}} = 7.58 \times 10^{-15}$$

This cracking probability is much lower than that of the classical substitution method and it is independent of the reference DNA size. However, in computational practices, intruders can utilize sequence alignment techniques to guide their guessing directions and therefore greatly accelerate the cracking process. This is true to both our method and the classical method. To battle such a cracking strategy, the phage can make some random noise mismatches at other non-recognized CRISPR spacer locations.

In Table I, |M| = the length of message M in bits, L = the length of the reference DNA sequence, and *bpn* is the number of bits hidden per nucleotide. Note that our method doubles the *bpn* compared to the traditional substitution method. We also introduce a new concept “volume”, which represents the maximum number of bits that the method can integrate into the reference DNA sequence. An off-target site must be of 23-nucleotides, and 5 mismatches can record 10 bits, thus the volume of our method is 10L/23, less than half of the classical method.

Allowing an indel in off-target sites can increase the cracking difficulty significantly as locating bulges is a much more time-consuming process than locating base mismatches. It has been proven biologically that valid off-target sites can have indels [25]. The reason why we didn’t include indels is mostly for simplicity. A more sophisticated version can certainly include indels. For example, the 4th definition of the substitution function can be re-defined as: any a 1-nucleotide indel → 00.

We randomly generated 1000 sgRNA spacers, and then searched for their off-target sites in human genome. The result is summarized in Table II. The same computational experiment was conducted on a simulated genome of size 3 billion base pairs in which A, C, G, T are randomly distributed. The result is presented in Table III. The data in both Table II and Table III illustrate that the naturally-occurred off-target sites are not abundant, indicating noise mismatches must be added into the reference DNA sequence to distract the intruders. Otherwise, by sequence aligning the R and R’, if R is publicly available, the intruders can quickly identify the spacer and retrieve the hidden message. Note that a simulated DNA sequence R can be created randomly anytime and can be deleted after R’ has been generated because the recipient can retrieve the data from R’ alone. Therefore, deleting R can disable the use of sequence alignment in cracking. Thus, it can be concluded that our proposed method works better with a simulated DNA sequence.

TABLE. I. PERFORMANCE OF OUR METHOD AND THE CLASSICAL METHOD

Method type	Capacity	Payload	bpn	Volume
Traditional	L	0	|M|/L	L
Our method	L	0	|M|/2L	10L/23

TABLE. II. OFF-TARGET SITES IN HUMAN GENOME

Number of mismatches	Total number of off-target sites	Averaged number of off-target sites per million bases
0	0	0
1	25	0
2	511	0
3	3306	0.001
4	45684	0.015
5	474587	0.158

TABLE. III. OFF-TARGET SITES IN SIMULATED GENOME

Number of mismatches	Total number of off-target sites	Averaged number of off-target sites per million bases
0	1	0
1	50	0
2	825	0
3	5376	0.002
4	67210	0.022
5	643482	0.214

The proposed method is not limited to digital DNA sequence applications. It can be practiced with biological DNA substances, too. One exemplar is to construct a plasmid and hide the secret message into the plasmid DNA sequence. The plasmid can be easily delivered to the recipient. The recipient applies Cas9-sgRNA reaction on the DNA. By analyzing the modified DNA sequences (for instance, deep sequencing), the recipient can detect all the off-target sites in the plasmid DNA sequence and from there, the recipient can then retrieve the hidden message. However, when practicing with biological DNA substances, one caution we must take is that, a large portion of potential off-target sites identified by computational methods do not exhibit any off-target effect in biological experiments [22, 26, 27]. Thus, only confirmed off-target sites should be included in the plasmid sequence.

The proposed method cannot be applied on DNA sequences of living organisms. The small number of naturally occurred off-target sites for a single sgRNA spacer limits the size of data that can be hidden in the reference DNA sequence.

V. CONCLUSION

In this paper, an original DNA based steganography method is proposed. This method adopts the CRISPR-Cas9 off-target editing and can reach much lower cracking probability than the classical substitution method. While the off-target editing is a flaw in CRISPR-Cas9 biological and medical applications, it can be used to enhance the applications of DNA based steganography.

REFERENCES

[1] K. Bailey and K. Curran, "An Evaluation of Image Based Steganography," Multimedia Tools and Applications, vol. 30, no. 1, pp. 55-88, 2006.

[2] C. T. Clelland, V. Risca and C. Bancroft, "Hiding messages in DNA microdots," Nature, vol. 399, pp. 533-534, 1999.

[3] A. Leier, C. Richter, W. Banzhaf and H. Rauhe, "Cryptography with DNA binary strands," BioSystems, vol. 57, pp. 13-22, 2000.

[4] H. J. Shiu, K. L. Ng, J. F. Fang, R. C. Lee and C. H. Huang, "Data hiding methods based upon DNA sequences," Information Sciences, vol. 180, pp. 2196-2208, 2010.

[5] L. M. Adleman, "Molecular computation of solutions to combinatorial problems," Science, vol. 266, pp. 1021-1024, 1994.

[6] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm," BMC Bioinformatics, vol. 8, p. 176, 2007.

[7] A. Khalifa and A. Atito, "High-capacity DNA-based steganography," in The 8th International Conference on Informatics and Systems, Giza, 2012.

[8] P. Das, S. Deb, N. Kar and B. Bhattacharya, "An improved DNA based dual cover steganography," Procedia Computer Science, vol. 46, pp. 604-611, 2015.

[9] P. Malathi, M. Manoaj, R. Manoj, V. Raghavan and R. E. Vinodhini, "Highly improved DNA based steganography," Procedia, vol. 115, pp. 651-659, 2017.

[10] D. A. Zebari, H. H. Haron and S. R. Zeebaree, "Security issuesin DNA based on data hiding: a review," International Journal of Applied Engineering Research, vol. 12, pp. 15363-15377, 2017.

[11] G. Hamed, M. Marey, S. E.-S. Amin and M. F. Tolba, "Hybrid, Randomized and high capacity conservative mutations DNA-based steganography for large sized data," Biosystems, vol. 167, pp. 47-61, 2018.

[12] H. Bae, B. Lee, S. Kwon and S. Yoon, "DNA steganalysis using deep recurrent neural networks," in Pacific Symposium on Biocomputing, Hawaii, 2019.

[13] S.-Y. Li, J.-K. Liu, G.-P. Zhao and J. Wang, "CADS: CRISPR/Cas12a-assisted DNA steganography for securing the storage and transfer of DNA-encoded information," ACS Synthetic Biology, vol. 7, pp. 1174-1178, 2018.

[14] NCBI, [Online]. Available: https://www.ncbi.nlm.nih.gov/genbank/statistics/. [Accessed 22 June 2019].

[15] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna and E. Charpentier, "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity," Science, vol. 337, pp. 816-821, 2012.

[16] R. Herai, "Avoiding the off-target effect of CRISPR/cas9 system is still a challenging accomplishment for genetic transformation," Gene, vol. 700, pp. 176-178, 2019.

[17] H. Y. Shin, C. Wang, H. K. Lee, K. H. Yoo, X. Zeng, T. Kuhns, C. M. Yang, T. Mohr, C. Liu and L. Hennighausen, "CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome," Nature Communications, vol. 8, p. 15464, 2017.

[18] D. Kim, D.-e. Kim, G. Lee, S.-l. Cho and J.-S. Kim, "Genome-wide target specificity of CRISPR RNA-guided adenine base editors," Nature Biotechnology, vol. 37, pp. 430-435, 2019.

[19] J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten and D. E. Root, "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISCRISCRISCRISPR-Cas9," Nature Biotechnology, vol. 34, no. 2, pp. 184-191, 2016.

[20] H. Zhou, M. T. Zhou, D. Li, J. Manthey, E. Lioutikova, H. Wang and X. Zeng, "Whole genome analysis of CRISPR Cas9 sgRNA off-target homologies via an efficient computational algorithm," BMC Genomics, vol. 18, no. Suppl 9, p. 826, 2017.

[21] S. Bae, J. Park and J. S. Kim, "Cas-OFFinder: a fast and versatile algorithm that searches potential off-target sites of Cas9 RNA-guided endonuclease," Bioinformatics, vol. 30, pp. 1743-1745, 2014.

[22] P. D. Hsu, D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann, V. Agarwala, Y. Li, E. J. Fine, X. Wu, O. Shalem, T. J. Cradick, L. A. Marraffini, G. Bao and F. Zhang, "DNA targeting specificity of RNA-guided Cas9 nucleases," Nature Biotechnology, vol. 31, pp. 827-832, 2013.

[23] Y. Fu, J. D. Sander, D. Reyon, V. M. Cascio and J. K. Joung, "Improving CRISPR-Cas nuclease specificity using truncated guide RNAs," Nature Biotechnology, vol. 32, no. 3, pp. 279-284, 2014.

[24] D. Zhang, T. Hurst, D. Duand and S. J. Chen, "Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design," PNAS, vol. 116, no. 18, pp. 8693-8698, 2019.

[25] Y. Lin, T. J. Cradick, M. T. Brown, H. Deshmukh, P. Ranjan, N. Sarode, B. M. Wile, P. M. Vertino, F. J. Stewart and G. Bao, "CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences," Nucleic Acids Research, vol. 42, no. 11, p. 7473–7485, 2014.

[26] Y. Fu, J. A. Foden, C. Khayter, M. L. Maeder, D. Reyon, K. J. Joung and J. D. Sander, "High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells," Nature Biotechnology, vol. 31, no. 9, pp. 822-826, 2013.

[27] V. Pattanayak, S. Lin, J. P. Guilinger, E. Ma, J. A. Doudna and D. R. Liu, "High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity," Nature Biotechnology, vol. 31, no. 9, pp. 839-843, 2013.