(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 October 2015 (29.10.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/163733 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G06F 19/22* (2011.01)

(21) International Application Number:
PCT/KR2015/004132

(22) International Filing Date:
24 April 2015 (24.04.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/983,988 24 April 2014 (24.04.2014) US
10-2014-0101133 6 August 2014 (06.08.2014) KR

(71) Applicant: **INSTITUTE FOR BASIC SCIENCE** [KR/KR]; 70, Yuseong-daero 1689beon-gil, Yuseong-gu, Daejeon 305-811 (KR).

(72) Inventors: **KIM, Jin Soo**; 503-106, Department of Chemistry, College of Natural Sciences, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 151-015 (KR). **BAE, Sang Su**; 503-106, Department of Chemistry, College of Natural Sciences, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 151-015 (KR).

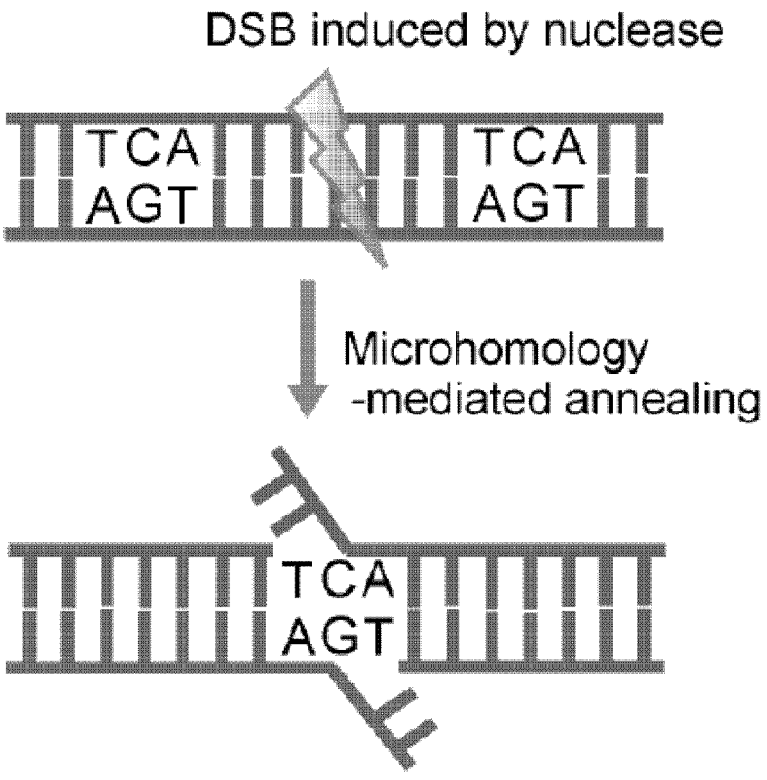
(74) Agent: **SON, Min**; (Hanol Intellectual Property & Law) STX R&D Center, 6th Floor, 163, YangJaeCheon-ro, Gangnam-gu, Seoul 135-855 (KR).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: A METHOD OF SELECTING A NUCLEASE TARGET SEQUENCE FOR GENE KNOCKOUT BASED ON MICROHOMOLOGY



(57) Abstract: The present invention relates to a method of selecting a nuclease target sequence for gene knockout based on microhomology.

WO 2015/163733 A1

Declarations under Rule 4.17:

— *as to non-prejudicial disclosures or exceptions to lack of novelty (Rule 4.17(v))*

Published:

— *with international search report (Art. 21(3))*

— *with sequence listing part of description (Rule 5.2(a))*

Description

Title of Invention: A METHOD OF SELECTING A NUCLEASE TARGET SEQUENCE FOR GENE KNOCKOUT BASED ON MICROHOMOLOGY

Technical Field

[1] The present invention relates to a method of selecting a nuclease target sequence for gene knockout based on microhomology.

Background Art

[2] Programmable nucleases, which include zinc finger nucleases (ZFNs), transcription-activator-like effector nucleases (TALENs), and RNA-guided engineered nucleases (RGENs) derived from the Type II CRISPR/Cas system, an adaptive immune response in bacteria and archaea, are now widely used for both gene knockout and knock-in in higher eukaryotic cells, animals, and plants. These nucleases induce DNA double-strand breaks (DSBs) at user-defined target sites in the genome, the repair of which via error-prone non-homologous end joining (NHEJ) or error-free homologous recombination (HR) gives rise to targeted mutagenesis and chromosomal rearrangements. Nuclease-mediated gene knockout is achieved preferentially via NHEJ rather than HR because NHEJ is a dominant DSB repair process over HR in higher eukaryotic cells and also because NHEJ does not require homologous donor DNA, fragments of which can be inserted at nuclease on-target and off-target sites. DSB repair by erroneous NHEJ is accompanied by small insertions and deletions (indels) at nuclease target sites, which can cause frameshift mutations in a protein-coding sequence. Inevitably, however, in-frame indels are also generated by this process, reducing the efficacy of nucleases in a population of cells and hampering the isolation of biallelic null clones. A recent study showed that RGENs induced in-frame deletions at frequencies up to 80%, resulting in incomplete gene disruption.

[3]

[4] It was reported that TALENs and RGENs produce deletions much more frequently than insertions and that nuclease-induced deletions are often associated with microhomology (Kim, Y. et al., Nature methods, 10:185, 2013), the presence of two identical short (2 to several base) sequences flanking a breakpoint junction: Apparently, microhomology stimulates nuclease-induced deletions via a DSB repair pathway known as microhomology-mediated end joining (MMEJ) (Fig. 1a), as observed in *C. elegans*, zebrafish, and human cell lines.

[5]

Disclosure of Invention

Technical Problem

[6] In this regard, the present inventors aimed to develop a technology for predicting a target sequence having a high probability of inducing out-of-frame mutations by an engineered nuclease. As a result, the present inventors developed a method and a program for providing useful information for selecting a nuclease target sequence via microhomology-mediated deletion prediction, and confirmed that these may be efficiently used in inducing effective gene disruptions in human cells, animals, etc., thereby completing the present invention.

[7]

Solution to Problem

[8] An objective of the present invention is to provide a method of selecting a nuclease target sequence for gene knockout.

[9] Another objective of the present invention is to provide a method of providing information for selecting a sequence having high efficiency of out-of-frame deletion by a nuclease.

[10] Still another objective of the present invention is to provide a computer program capable of performing the method.

[11] Still another objective of the present invention is to provide a computer-readable recording medium in which the program is recorded.

Advantageous Effects of Invention

[12] The method according to the present invention enables to identify or select a target site having a low probability of inducing in-frame mutations thus capable of easily producing mutants with knockout of a particular gene. Therefore, the method of increasing knockout efficiency using technologies such as the engineered nuclease technology can be efficiently used in the field of clinical research on life science.

[13]

Brief Description of Drawings

[14] Figs 1a to 1e show prediction of nuclease-induced deletion patterns that are associated with microhomology. (Fig 1a) Schematic representation of microhomology-mediated annealing at a nuclease target site. (Fig 1b) In silico-predicted deletion patterns that result from microhomology-associated DNA repair. Microhomologies are shown in underlined. The equation used for calculating pattern scores is shown below the table. (Fig 1c) Comparison of the pattern score with the experimentally-determined frequency of the deletion pattern found using the deep sequencing data. Arrows indicate the three most frequent deletion patterns correctly predicted by the scoring system. The Pearson correlation coefficient is shown. (Fig 1d) Comparison of microhomology scores with the experimentally-determined frequencies of microhomology-as-

sociated deletions. The microhomology score is the sum of all the pattern scores assigned to hypothetical deletion patterns at a given target site. (Fig 1e) Comparison of out-of-frame scores with the frequencies of frameshifting deletions observed in cells transfected with TALENs and RGENs.

[15] Figs 2a to 2d show Experimental validation of the scoring system. (Fig 2a) The distribution of out-of-frame scores associated with potential target sites in the BRCA1 gene. (Fig 2b) The frequencies of out-of-frame indels determined by deep sequencing at high-score and low-score sites. The dashed lines correspond to the peak value of the Gaussian distribution of out-of-frame scores shown in (Fig 2a). (Fig 2c) Correlation of the out-of-frame scores with the frequencies shown in (Fig 2b). (Fig 2d) Correlation of the out-of-frame scores with the frequencies of frameshifting indels (left) or deletions (right) induced by 68 RGENs.

[16] Fig 3 shows analysis of mutations induced by TALENs and RGENs. (a) The average frequencies of mutations induced by 10 TALENs in HEK293T cells and 10 RGENs in K562 cells. (b) Frequencies of deletions and insertions induced by TALENs and RGENs. Nuclease-induced mutations were classified as deletions or insertions relative to the wild-type sequences. Substitutions that may result from PCR or sequencing errors were obtained rarely (< 0.1%) and excluded in this analysis. (c) Frequencies of microhomology-associated deletions induced by TALENs and RGENs.

[17] Figs 4a to 4c show evaluation of weight factor for deletion length. The weight factor for deletion length was calculated by fitting the deep sequencing data obtained with TALENs (Fig 4a) and RGENs (Fig 4b) to a single-exponential function (shown as a line). (Fig 4c) The average weight factor for TALENs and RGENs.

[18] Figs 5a to 5c show source code for assigning a score to a hypothetical deletion pattern associated with microhomology.

[19] Figs 6a and 6b show comparison of the pattern score with the experimentally-determined frequency of the pattern using the deep sequencing data. Arrows indicate the most frequent deletion patterns correctly predicted by the scoring system. The Pearson correlation coefficient is shown.

[20] Fig 7 shows distribution of microhomology scores in the BRCA1 gene. Microhomology scores were assigned to all RGEN target sites in the human BRCA1 gene. The distribution of microhomology scores were fitted to a Gaussian function with a peak value at 4026 and a width of 1916.

[21] Fig 8 shows high-score and low-score sites. (a) Two RGEN target sites separated by 29 bp in the MCM6 gene. Out-of-frame scores at the two sites are shown in parentheses. (b) The most frequent deletion patterns obtained in cells transfected by the RGEN plasmids. Microhomologies are shown in underlined. The two PAM sequences are highlighted.

[22] Fig 9 shows comparison of out-of-frame scores with experimental data. (a) Genotype analysis of 81 live-born mice carrying mutations that had been produced via TALENs or RGENs in our previous studies. (b) Correlation of the out-of-frame scores with the frequencies of out-of-frame deletions (Pearson correlation coefficient = 0.996).

[23] Fig. 10 shows flow chart for system for selecting a target having high efficiency of gene knockout.

[24]

Best Mode for Carrying out the Invention

[25] In one aspect, the present invention provides a method of selecting a nuclease target sequence for gene knockout.

[26]

[27] The method according to the present invention may be used as a target-selecting system capable of pre-estimating the frequency of microhomology-associated deletion, may calculate the out-of-frame score of an *in silico* nuclease target site, and may help selecting an appropriate target site to enable gene knockout in cultured cells, plants, or animals using a scoring system. Therefore, the method may be used for predicting a frequency of out-of-frame deletions of a nuclease target sequence.

[28]

[29] In particular, the present invention provides a method of selecting a nuclease target sequence for gene knockout, which includes:

[30] (a) providing a nuclease target sequence candidate;

[31] (b) collecting information of microhomology present in the nuclease target sequence candidate; and

[32] (c) predicting frequency of microhomology-associated out-of-frame deletion of the nuclease target sequence candidate based on the information of microhomology collected in step (b).

[33]

[34] Further, the method further comprises a step of comparing the frequency of microhomology-associated out-of-frame deletion predicted in step (c) with frequency of microhomology-associated out-of-frame deletion of other nuclease target sequence candidate. Through this step, the nuclease target sequence having high efficiency of out-of-deletion frame deletion can be selected among the nuclease target sequence candidates.

[35]

[36] Further, the information of microhomology may comprise a size of microhomology sequence, a distance between two microhomology sequences, and sequence information of the microhomology sequence, but is not limited thereto.

[37]

[38] The nuclease target sequence candidate may include any sequence as long as it is a sequence in which deletion may be induced by microhomology. In particular, the sequence may be originated from human cells, zebrafish, *C. elengans,* etc., but is not limited thereto. Further, the sequence may be a sequence of mammalian cells, insect cells, plant cells, fish cells, or etc, but is not limited thereto.

[39] In the present invention, the microhomology sequence present in the target sequence refers to a sequence of at least 2bp having 100% identity with a sequence present in other region of the target sequence. In detail, the microhomogy sequences refer to identical sequences of at least 2bp flaking a position expected to be cleaved by a nuclease, but not limited thereto. For example, the microhomology sequence in the present invention may have a length of at least 2 bp, 3 bp, 4 bp, 5 bp, 6bp, 7bp, or 8bp, but is not limited thereto. The length of the microhomology sequence may vary depending on a given nuclease target sequence, and is preferably at least 2bp. Further, the length of the microhomology sequence is preferably shorter than the length from 5' or 3' end of the target sequence to a position expected to be cleaved by a nuclease of the nuclease target sequence. If microhomology sequences are present in both sides of a position cleaved by a nuclease, nuclease-induced deletion may be induced by micro-homology-mediated annealing (Fig. 1a).

[40]

[41] The nuclease target sequence candidate or nuclease target sequence according to the present invention may have an identical sequence length in both directions with respect to a position expected to be cleaved by a nuclease, but is not limited thereto.

[42] Bases which constitute the target sequence according to the present invention may be selected from the group consisting of A, T, G, and C, but are not limited thereto as long as they are bases which constitute the target sequence.

[43] The position expected to be cleaved by a nuclease according to the present invention refers to a position where the covalently bonded backbone of the nucleotide molecules is expected to be disrupted by a nuclease.

[44] The target sequence may be located in a gene regulatory region or a gene region, but is not limited thereto. The target sequence may be present within 10 kb, 5 kb, 3 kb, or 1 kb, or 500 bp, 300 bp, or 200 bp from the transcription start site of a gene, for example, upstream or downstream of the start site, but is not particularly limited as long as it is a target sequence for a nuclease.

[45] Meanwhile, the gene regulatory region according to the present invention may be selected from promoters, transcription enhancers, 5' non-coding regions, 3' non-coding regions, virus packaging sequences, and selectable markers, but is not limited thereto. Further, the gene region according to the present invention may be an exon or an

intron, but is not limited thereto.

[46]

[47] The nuclease according to the present invention may be selected from the group consisting of zinc finger nucleases (ZFNs), transcription-activator-like effector nucleases (TALENs), and RNA-guided engineered nucleases (RGENs), but is not limited thereto.

[48] ZFN may include a DNA-cleavage domain and a Zinc finger DNA-binding domain, and particularly, an integration of the two domains, which may be connected by a linker. Further, the zinc finger DNA-binding domain may be modified so that it can bind to a desired DNA sequence.

[49] Further, TALEN may include a DNA-cleavage domain and transcription activator-like effectors (TALE) DNA-binding domain, and particularly an integration of the two domains, which may be connected by a linker. Further, TALE may be modified so that it binds to a desired DNA sequence.

[50] RGEN refers to a nuclease containing a target DNA-specific guide RNA and Cas protein as components. The term "guide RNA" refers an RNA specific to a target DNA, which binds to Cas protein, thereby guiding the Cas protein to the target DNA.

[51] Further, the guide RNA may be composed of two RNAs such as CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA), or may be a single-chain RNA (sgRNA) produced by the integration of main parts of crRNA and tracrRNA.

[52] The guide RNA may be a dual RNA including crRNA and tracrRNA, and crRNA may bind to a target DNA.

[53] Examples of the nuclease are not limited thereto, but may include any nuclease capable of inducing microhomology-associated deletion reflecting the objectives of the present invention, without limitations.

[54]

[55] Further, in order to predict the frequency of microhomology-associated out-of-frame deletion of the nuclease target sequence candidate, step (c) may comprise calculating a pattern score, which is a score assigned to an expected deletion pattern of each of microhomologies present in the given nuclease target sequence candidate; and calculating (i) a microhomology score, which is a sum of the pattern scores of all microhomologies in the given nuclease target sequence candidate and (ii) a out-of-frame score, which is a ratio of a score which is a sum of the pattern scores of microhomologies associated with out-of-frame deletion to the microhomology score, based on the calculated pattern score.

[56]

[57] The method according to the present invention may comprise the following steps, but it not limited thereto:

[58] i) providing a nuclease target sequence candidate;

[59] ii) examining, in the given nuclease target sequence, whether two identical sequences of at least 2 bp flanking a position expected to be cleaved by a nuclease are present in the target sequence to identify the presence of microhomology;

[60] iii) obtaining information of microhomology, when the microhomology is present in the target sequence, and repeating steps ii) and iii) one or more times;

[61] iv) calculating a pattern score, which is a score assigned to an expected deletion pattern of each of microhomologies present in the given nuclease target sequence candidate; and

[62] v) calculating (i) a microhomology score, which is a sum of the pattern scores of all microhomologies in the given nuclease target sequence candidate and (ii) a out-of-frame score, which is a ratio of a score which is a sum of the pattern scores of microhomologies associated with out-of-frame deletion to the microhomology score.

[63]

[64] Step ii) is a step of obtaining information of microhomology, *e.g.*, a distance between 5' positions of the microhomology sequences or a distance between 3' positions of the microhomology sequences, and sequence information of the microhomology sequence, when the microhomology is present in the target sequence. Further, step iii) may further comprise a step of repeating step ii) and iii) one or more times to obtain information on all microhomologies.

[65] In particular, step iii) may be for obtaining information about a deletion length when nuclease-induced deletion is induced by MMEJ, and microhomology sequence, location, etc.

[66] All microhomogy patterns present in the given nuclease target sequence can be obtained via step iii).

[67]

[68] Step iv) refers to calculating a pattern score based on the information obtained from step iii).

[69]

[70] In an embodiment, the present invention confirmed that microhomology-associated deletion depends on the size and deletion length of microhomology. In particular, it was confirmed that as the size of microhomology increases, the frequency of deletion increase, while as the deletion length increases, the frequency of deletion decreases. In this regard, an equation for scoring a hypothetical deletion pattern (herein, also referred to as "pattern score") of a given nuclease target sequence was induced based on the results.

[71]

[72] In particular, a pattern score may be calculated by the following Equation 1.

[73] [Equation 1]

[74] Pattern score = S X exp(-△ / W_{length}),

[75] wherein:

[76] S is a microhomology index that corresponds to the size and base pairing energy of the microhomology sequence;

[77] △ is a distance between 5' positions of the microhomology sequences or a distance between 3' positions of the microhomology sequences (deletion length); and

[78] W_{length} is a weight factor on a distance between the microhomology sequences.

[79]

[80] More particularly, S is an index which corresponds to the size of a microhomology sequence and the base pairing energy which constitutes the same, and for example, may be calculated using Equation 4.

[81] [Equation 4]

[82] Microhomology index = (number of G and C in a microhomology sequence)*2 + (number of A and T bases in a microhomology sequence).

[83]

[84] Considering that G:C pairs are more stable than A:T pairs, +2 was assigned for the number of GC, and +1 was assigned for the number of AT, but are not limited thereto. It may be calculated by various methods which put more weight on the number of GC.

[85]

[86] Further, in the equation,

[87]

[88] W_{length} is a weight factor on a distance between the two sequence fragments, and may be 20 for example. However it is not limited thereto.

[89]

[90] Furthermore, in one embodiment, the present invention may perform calculating a pattern score by classifying step iv) into either when a deletion length is a multiple of 3 or when it is not a multiple of 3, but is not limited thereto.

[91] Here, when a distance between sequence fragments, thus a deletion length, is a multiple of 3, it may be determined that an in-frame deletion will be induced. On the other hand, when the deletion length is not a multiple of 3, it may be determined that an out-of-frame deletion will be induced.

[92]

[93] Further, prior to performing step iv), eliminating of overlapping information obtained from step iii) may be included, but is not limited thereto.

[94]

[95] Step v) of the method is a step of calculating a microhomology score, an out-of-frame score, or both based on the pattern score from iv). Further, more particularly,

the microhomology score and out-of-frame score may be calculated by the following Equations 2 and 3, respectively.

[96] [Equation 2]

[97] Microhomology score = ∑ pattern score,

[98] wherein the microhomology score is a sum of pattern scores of the obtained all microhomologies;

[99]

[100] [Equation 3]

[101] Out-of-frame score = ∑ pattern score of out-of-frame deletion / microhomology score (∑ pattern score),

[102] wherein ∑ pattern score of out-of-frame deletion is a sum of pattern scores of relevant microhomologies whose a deletion length is not a multiple of 3.

[103]

[104] Based on the microhomology score and the out-of-frame score calculated in the step above, the frequency of microhomology-associated deletion and frame shifting mutation regarding a nuclease target sequence may be predicted.

[105]

[106] The method according to the present invention may be implemented as a computer program, and be used to easily select a target having high efficiency of gene knockout. Computer programming languages capable of implementing the method according to the present invention are Python, C, C++, Java, Fortran, Visual basic, etc., but are not limited thereto. Each of the programs may be saved in a compact disc read only memory (CD-ROM), a hard disk, a magnetic diskette, or a similar recording medium tools, etc., and may be connected to intra- or internetwork systems. For example, the computer system may search the nucleotide sequences of a target gene or a regulatory region thereof by connecting to a sequence data base such as GenBank (http://www.ncbi.nlm.nih.gov/nucleotide) using HTTP, HTTPS, or XML protocols.

[107]

[108] The method according to the present invention may be used to help selecting an appropriate target site for knockout in cultured cells, plants, and animals by effectively predicting the frequency of microhomology-associated deletion of a nuclease target sequence. Further, the method may significantly increase efficiency not only in gene knockout cell clones and animals such as livestock, but also in nuclease-mediated genes or cellular therapies.

[109]

[110] In another aspect, the present invention provides a method of providing information for selecting a sequence having a high efficiency of out-of-frame deletion by a nuclease.

[111]

[112] In particular, it provides a method of providing information for selecting a sequence having high efficiency of out-of-frame deletion by a nuclease, including:

[113] (a) providing a nuclease target sequence candidate;

[114] (b) collecting information of microhomology present in the nuclease target sequence candidate; and

[115] (c) predicting frequency of microhomology-associated out-of-frame deletion of the nuclease target sequence candidate based on the information of microhomology collected in step (b).

[116] Steps (a) to (c) and each term are the same as described above.

[117]

[118] In another aspect, the present invention provides a computer program performing the steps of the method according to the present invention.

[119]

[120] The method, each step, and the computer program are the same as previously described above.

[121]

[122] In another aspect, the present invention provides a computer-readable recording medium in which the program is recorded.

[123] The program, the recording medium, etc., are the same as previously described above.

[124]

Mode for the Invention

[125] Hereinafter, the present invention will be described in more detail with reference to Examples. It is to be understood, however, that these examples are for illustrative purposes only and are not intended to limit the scope of the present invention.

[126]

[127] **Example 1: Materials & Methods**

[128]

[129] (1) Cell culture and transfection

[130]

[131] K562 (ATCC, CCL-243) cells were grown in RPMI-1640 with 10% FBS and a penicillin/streptomycin mix (100 units/mL and 100 mg/mL, respectively). To induce mutations in human cells using RGENs, $2x10^6$ K562 cells were transfected with 20 μg of Cas9-encoding plasmid using Amaxa SF Cell Line 4D-Nucleofector Kit (Lonza) according to the manufacturer's protocol. After 24 h, 60 mg and 120 mg of *in vitro* transcribed crRNA and tracrRNA, respectively, were transfected into 1 x 10^6 K562

cells. Genomic DNA was isolated at 48 h post-transfection. HEK293T/17 (ATCC, CRL-11268) and HeLa (ATCC, CCL-2) cells were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 100 units/mL penicillin, 100 μg/mL streptomycin, 0.1 mM nonessential amino acids, and 10% fetal bovine serum (FBS). To induce mutations in HEK 293T cells using TALENs, $2x10^5$ HEK293T cells were transfected with TALEN-encoding plasmids (500 ng) using lipofectamine 2000 (Invitrogen, Carlsbad, CA) according to the manufacturer s protocol. Genomic DNA was isolated at 72 h post-transfection. 1.6 x 10^4 HeLa cells were transfected with Cas9-encoding plasmid (0.1 μg) and sgRNA expression plasmid (0.1 μg) using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. Cells were collected 72 h after transfection and lysed with cell lysis buffer (0.005% SDS containing Proteinase K from Tritirachium album (1:50; Sigma-Aldrich)).

[132]

[133] (2) Construction of TALEN-encoding plasmids

[134]

[135] TALENs were designed to target sites shown in Tables 1 and 2. TALEN-encoding plasmids were assembled using the one-step Golden-Gate cloning system that we described previously.

[136] Table 1

[Table 1]

Nuclease (cell type)	Gene	Name	Target site (5′ to 3′)*	SEQ ID NO
TALEN (HEK293T)	*APP*	APP_1	TAGACCCCCGCCACAGCAGC ctctgaagttgg ACAGCAAAACCATTGCTTCA	1
	CD4	CD4_1	TGTCTCAGCTGGAGCTCCAG gatagtggcacc TGGACATGCACTGTCTTGCA	2
	CREBBP	CREB_1	TGTCCAATGACCTGTCCCAG aagctgtatgcc ACCATGGAGAAGCACAAGGA	3
	TP53	TP53_1	TACAACTACATGTGTAACAG ttcctgcatggg CGGCATGAACCGGAGGCCCA	4
	CFTR	CFTR_1	TCGGAAGGCAGCCTATGTGA gatacttcaata GCTCAGCCTTCTTCTTCTCA	5
	CFTR	CFTR_2	TCTCTTACTGGGAAGAATCA tagcttcctatg ACCCGGATAACAAGGAGGAA	6
	DROSHA	DROS_1	TGAGGAGGAGATTGCCAATA tgcttcagtggg AGGAGCTGGAGTGGCAGAAA	7
	DROSHA	DROS_2	TGAAGGATACAGAAATGACT gtgaatcaaccc ATATCATCAAGGAGCTGATA	8
	NFKB1	NFKB_1	TATGTATGTGAAGGCCCATC ccatggtggact ACCTGGTGCCTCTAGTGAAA	9
	NFKB1	NFKB_2	TTGTCATTGCTGTTGTCCCT ctgctacgttcc TATTGTCATTAAAGGTATCA	10
RGEN (K562)	*C4BPB*	C4BP_1	AATGACCACTACATCCTCAAGGG	11
	CCR5	CCR5_1	TGACATCAATTATTATACATCGG	12
	DROSHA	DROS_1	GATTGCCAATATGCTTCAGTGGG	13
	CCR5	CCR5_2	CCTCCGCTCTACTCACTGGTGTT	14
	CCR5	CCR5_3	CCTGCCTCCGCTCTACTCACTGG	15
	CCR5	CCR5_4	GAATCCTAAAAACTCTGCTTCGG	16
	CCR5	CCR5_5	CCTAAAAACTCTGCTTCGGTGTC	17
	CCR5	CCR5_6	AAATGAGAAGAAGAGGCACAGGG	18
	AAVS1	AAVS1_1	CTCCCTCCCAGGATCCTCTCTGG	19
	EMX1	EMX1	GAGTCCGAGCAGAAGAAGAAGGG	20

*A TALEN site consists of the left-half site (upper-case letters), spacer (lower-case letters), and the right-half site (upper-case letters). PAM sequences are shown in underlined.

[137]

[138] Table 2

[Table 2]

Nuclease (cell type)	Gene	Name	Target site (5′ to 3′)*	SEQ ID NO
TALEN (HEK293T)	BRCA1	BRCA1_l	TCCAGCTGCTGCTCATACTA ctgatactgctg GGTATAATGCAATGGAAGAA	21
	BRCA1	BRCA1_h	TCCTGAACATCTAAAAGATG aagtttctatca TCCAAAGTATGGGCTACAGA	22
	CXCR4	CXCR4_l	TCTTCCTGCCCACCATCTAC tccatcatcttc TTAACTGGCATTGTGGGCAA	23
	CXCR4	CXCR4_h	TGGGTTGATTTCAGCACCTA cagtgtacagtc TTGTATTAAGTTGTTAATAA	24
	MCM6	MCM6_l	TTAGAAGTAATTTTAAGGGC tgaagctgtgga ATCAGCTCAAGCTGGTGACA	25
	MCM6	MCM6_h	TGGAATCAACTTGTATGAAA ccttgtcaaaat GTACTCCACAAGTATGTACA	26
	PHF8	PHF8_l	TACAGAAGGCCCAAAAGAAG aaatatatcaag AAGAAGCCTTTGCTGAAGGA	27
	PHF8	PHF8_h	TACAGCCTGCTTGCTCCGCC tataccacagag CACAGCCTGGACATTATGGA	28
	SLC18A2	SLC18_l	TCCAGTCATATCCGATAGGT gaagatgaagaa TCTGAAAGTGACTGAGATGA	29
	SLC18A2	SLC18_h	TGTATAAAACAGTGTTTCCA gtgacacaactc ATCCAGAACTGTCTTAGTCA	30
	TP53	TP53_l	TGTACCACCATCCACTACAA ctacatgtgtaa CAGTTCCTGCATGGGCGGCA	31
	TP53	TP53_h	TTGTGAGCCACCACGTCCAG ctggaagggtca ACATCTTTTACATTCTGCAA	32
RGEN (K562)	APP	APP_l	AGAGGAGGAAGAAGTGGCTGAGG	33
	APP	APP_h	GCCACAGCAGCCTCTGAAGTTGG	34
	BRCA1	BRCA1_l	GCTCATACTACTGATACTGCTGG	35
	BRCA1	BRCA1_h	ATTGACAGCTTCAACAGAAAGGG	36
	MCM6	MCM6_l	GCTAGGGACAGAAGTGTTTCTGG	37
	MCM6	MCM6_h	CTCGTGGCCTGGAGCCTGGCTGG	38

*A TALEN site consists of the left-half site (upper-case letters), spacer (lower-case letters), and the right-half site (upper-case letters). PAM sequences are shown in underlined.

[139]

[140] (3) Construction of Cas9-encoding plasmids.

[141]

[142] The Cas9-encoding plasmid and sgRNA-encoding plasmids were constructed. The Cas9 protein is expressed under the control of the CMV promoter and fused to a peptide tag (NH_3-GGSGPPKKKRKVYPYDVPDYA-COOH, SEQ ID NO: 39)

containing the HA epitope and a nuclear localization signal (NLS) at the C-terminus.

[143]

[144] (4) RNA preparation

[145]

[146] RNAs used in K562 cells were *in vitro* transcribed through run-off reactions by T7 RNA polymerase using a MEGAshortscript T7 kit (Ambion) according to the manufacturer's manual. Templates for sgRNA or crRNA were generated by annealing and extension of two complementary oligonuceotides (Tables 1 or 2). Transcribed RNA was purified by phenol:chloroform extraction, chloroform extraction, and ethanol precipitation. Purified RNA was quantified by spectrometry.

[147]

[148] (5) Targeted deep sequencing

[149]

[150] Genomic DNA segments that encompass the nuclease target sites were amplified using Phusion polymerase (New England Biolabs). Equal amounts of the PCR amplicons were subjected to paired-end read sequencing using Illumina MiSeq at Bio-Medical Science Co. (South Korea). Rare sequence reads that constituted less than 0.005% of the total reads were excluded. Indels located around the RGEN cleavage site (3 bp upstream of the PAM) and around the TALEN target site (spacer) were considered to be mutations induced by RGENs and TALENs, respectively.

[151]

[152] **Example 2: determination of mutant sequences induced by TALENs and RGENs in human cells**

[153]

[154] The mutant sequences induced by 10 TALENs and 10 RGENs in human cells using deep sequencing were determined. TALENs and RGENs induced mutations at frequencies of 19.7±3.6% (mean±s.e.m) in HEK293T cells and 47.0±5.9% in K562 cells, respectively (Figure 3, Tables 1 and 3).

[155]

[156] Analysis was focused on deletions and excluded insertions because deletions are much more prevalent than are insertions (98.7% vs. 1.3% for TALENs and 75.1% vs. 24.9% for RGENs) and because microhomology is irrelevant to insertions. In aggregate, deletions were associated with microhomology at a frequency of 44.3% for TALENs and 52.7% for RGENs (Figure 3, Table 3). Thus, 43.7% (= 0.987 x 0.443) and 39.6% (= 0.751 x 0.527) of all the indels induced by TALENs and RGENs, respectively, were associated with microhomology. At a given nuclease target site, these microhomology-associated deletions can be predicted. In an extreme case, all or none of these deletions can cause frameshift in a protein-coding gene. In contrast, one third

of microhomology-independent indels result in in-frame mutations. Assuming that ~60% of indels are microhomology-independent on average, the fraction of in-frame mutations at a given site can range from 20% (= 60%/3 + 0%) to 60% (= 60%/3 + 40%), a three-fold difference between the two extreme cases. Because most eukaryotic cells are diploid rather than haploid, the fraction of null cells carrying two out-of-frame mutations can range from 16% (= 0.40 x 0.40) to 64% (= 0.80 x 0.80), depending on the choice of target sites.

[157] Table 3

[Table 3]

Nuclease (cell type)	Gene	Name	Number of sequence reads	Insertion	Deletion	Frequency of out-of-frame deletions (%)	Frequency of out-of-frame indels (%)	Frequency of microhomology-associated deletions (%)	Microhomology score[a]	Out-of-frame score[b]
TALEN (HEK293T)	*APP*	APP_1	58822	148	24260	74.18796373	74.22976073	45.08326	3930	73.61323155
	CD4	CD4_1	130890	221	15863	79.56250394	79.66923651	45.04633	3915	85.84929757
	CREBBP	CREB_1	146455	524	46455	72.3065332	72.41959173	48.77021	4184	48.11185468
	TP53	TP53_1	104451	216	13619	58.7561495	59.02421395	37.33461	2704	44.41568047
	CFTR	CFTR_1	133089	191	11835	57.82847486	58.21553301	40.79425	3171	48.53358562
	CFTR	CFTR_2	122477	90	9239	80.14936681	80.26583771	47.2129	3399	83.81877023
	DROSHA	DROS_1	218200	360	34204	61.34370249	61.23423215	42.91603	4195	46.79380215
	DROSHA	DROS_2	240203	1455	74503	69.29251171	69.37649754	39.50177	3400	81.05882353
	NFKB1	NFKB_1	107680	189	14017	57.95105943	57.90511052	44.29835	4111	43.29846753
	NFKB1	NFKB_2	235082	748	47387	80.92514825	80.69595928	52.7383	3642	93.49258649
RGEN (K562)	*C4BPB*	C4BP_1	47856	21247	11768	38.9785 86	76.08662729	46.46924	2969	40.99022324
	CCR5	CCR5_1	200645	10727	94967	83.49216043	83.75877533	47.60201	3316	71.26055489
	DROSHA	DROS_1	251509	15723	106834	56.85549544	60.24217303	40.52596	4530	46.55629139
	CCR5	CCR5_2	76347	1723	26406	74.16496251	75.49148566	47.13929	3772	65.16436904
	CCR5	CCR5_3	73367	2511	10001	62.34376562	69.46131714	55.49345	5118	57.44431419
	CCR5	CCR5_4	69780	1325	17745	65.08312201	67.29417934	59.77289	4148	68.63548698
	CCR5	CCR5_5	99571	3256	29392	80.3041644	82.11529037	62.9491	4569	76.01225651
	CCR5	CCR5_6	106450	22712	25837	68.4754422	83.03363612	44.9402	3660	60.51912565
	AAVS1	AAVS1_1	43249	7812	18964	86.24762708	93.29997012	57.83959	5894	72.344 76
	EMX1	EMX1	52945	16745	22858	47.30072622	69.47453476	64.47283	4756	50.75694

[158]

[159] A careful analysis of indel sequences also revealed that the frequency of micro-homology-associated deletions depends on both the size of the microhomology and the length of the deletions. Thus, as the microhomology size increased, the deletion frequency also increased. In addition, as the length of deletions increased, the deletion frequency decreased exponentially (Fig. 4). For example, the two most frequent

deletions induced by a TALEN pair specific to the human *APP* gene were associated with 5- and 4-nucleotide sequences separated by 20 and 17 bp, respectively, near the target site (Fig. 1b).

[160]

[161] **Example 3: Formula to predict microhology-associated deletions**

[162]

[163] Based on these observations, a simple formula to predict microhology-associated deletions was developed. First, deletion patterns at a given nuclease target site that are associated with microhomology of at least 2 bases *in silico* were predicted and then a score was assigned to each hypothetical deletion pattern using a computer program written in Python (Figs. 5a to 5c), according to the following equation 1 that accounts for both the size of microhomology and the deletion length (Fig. 1b).

[164]

[165] [Equation 5]

[166] A pattern score = S X exp(-△/20),

[167] where S is the microhomology index that corresponds to the size of microhomology and base pairing energy and

[168] △ is the deletion length in base pairs (bp).

[169]

[170] Because G:C base pairs are more stable than are A:T pairs, each A:T pair and each G:C pair in the microhomology sequence were arbitrarily assigned to +1 and +2, respectively, to obtain the microhomology index. This simple formula accurately predicted the three most frequent deletion patterns at the TALEN site (Fig. 1c). The program was used to assign scores to the other 19 sites. The program accurately predicted the most frequent deletion pattern at 5 TALEN sites and 8 RGEN sites (Figs. 6a and 6b). Overall, the scores correlated well with the deep sequencing data: The Pearson correlation coefficient ranged from 0.411 to 0.945 at the 20 sites with a mean value of 0.727.

[171]

[172] **Example 4: Evaluation of utility of scoring system**

[173]

[174] To choose nuclease target sites that are prone to forming microhomology-mediated deletions and out-of-frame mutations, two scores were assigned to each target site. A microhomology score is the sum of all the scores assigned to hypothetical deletion patterns at a given site: ∑ pattern score. An out-of-frame score assigned to each target site is calculated by the following equation 2:

[175]

[176] [Equation 3]

[177] Out-of-frame score = ∑ pattern score of an out-of-frame deletion/ ∑ pattern score

[178]

[179] The distance between the target sites was ±30bp. Then, the predicted scores were compared with the experimental data at the 20 sites. Both the microhomology scores and the out-of-frame scores were statistically significant predictors of the frequencies of microhomology-associated deletions and frame shifting mutations, respectively (Pearson coefficient = 0.635 and 0.797, respectively) (Figs. 1d and e). These results suggest that one can use the scoring system to choose sites appropriate for targeted gene disruption.

[180]

[181] To evaluate the utility of our scoring system, two target sites, one with a high score and the other with a low score, in each of 9 human genes were chosen. To this end, all RGEN target sites (5'-X20NGG-3', where X20 corresponds to the crRNA or sgRNA sequence and NGG is the protospacer-adjacent motif (PAM) recognized by Cas9) in the human *BRCA1* gene (9,494 sites in exons and introns) were firstly identified and the microhomology score and the out-of-frame score were assigned to each target site. Interestingly, the out-of-frame scores were distributed according to a Gaussian function with a peak value at 65.9 (Fig. 2a). This is expected because two thirds of all the microhomology-associated deletions would result in frame-shift mutations. Two target sites in exons, one from the top 20% of the scores and the other from the bottom 20%, were arbitrarily chosen. Likewise, high-score sites and low-score sites in 8 other genes were chosen. A total of 6 or 12 sites were targeted by RGENs or TALENs, respectively (Table 2). Then, mutations in human cells by transfecting cells with plasmids encoding these nucleases were induced, regions containing the target sites were amplified, and the PCR amplicons were deeply sequenced to obtain the fraction of out-of-frame indels at each target site (Table 4).

[182] Table 4

[Table 4]

Nuclease (Cell type)	Gene	Name	Number of sequence reads	Insertion	Deletion	Frequency of out-of-frame deletions (%)	Frequency of out-of-frame indels (%)	Microhomology score*	Out-of-frame score[b]
TALEN (HEK293T)	BRCA1	BRCA1_l	77583	795	32519	39.10479085	39.61392158	4303	21.77551
	BRCA1	BRCA1_h	122533	871	62077	81.10301121	81.08068489	3045	80.42693
	CXCR4	CXCR4_l	117578	417	42130	45.26139826	45.26186207	3903	37.56085
	CXCR4	CXCR4_h	280176	882	52068	83.71982103	83.71436317	4061	84.73282
	MCM6	MCM6_l	191096	3459	131302	43.83248991	44.57927991	3759	41.63341
	MCM6	MCM6_h	267702	941	19526	80.00247724	80.4623862	3812	79.56453
	PHF8	PHF8_l	253216	1071	87348	41.78051364	42.10553931	4765	42.70724
	PHF8	PHF8_h	264899	1811	75500	72.27631047	72.47083002	3267	78.29813
	SLC18A2	SLC18_l	356244	2773	147564	39.79381922	40.00610221	4816	45.72259
	SLC18A2	SLC18_h	374261	2427	98331	75.64093697	75.76827054	4220	85.92417
	TP53	TP53_l	84253	342	15334	48.18713345	48.46955659	3236	31.33498
	TP53	TP53_h	176325	1210	28962	79.16705144	78.8308357	3769	86.35421
RGEN (K562)	APP	APP_l	68578	559	6112	34.55981506	38.37524878	7565	23.91276
	APP	APP_h	278349	2952	23162	76.58807947	77.76956436	4180	73.37321
	BRCA1	BRCA1_l	143960	10054	30439	34.66284963	47.56692842	3658	23.75615
	BRCA1	BRCA1_h	102903	3066	15415	88.16393382	88.66998256	4432	79.62545
	MCM6	MCM6_l	273431	3304	93399	34.19839631	36.18849409	4359	38.74742
	MCM6	MCM6_h	167502	6026	14745	65.16221147	74.78114478	6330	71.87994

[a]Microhomology score = $\sum$pattern score.

[b]Out-of-frame score = $\frac{\sum \text{pattern score of an out-of-frame deletion}}{\sum \text{pattern score}}$. (±30bp between target sites)

[183]

[184] High-score sites produced out-of-frame indels much more frequently than did low-score sites in all of the 9 pairs (Fig. 2b). Thus, all 9 high-score sites produced frameshifting indels at frequencies higher than 66%, the mean value of predicted scores. In contrast, all 9 low-score sites produced out-of-frame mutations at frequencies much lower than the mean. For example, two RGENs induced out-of-frame

indels at frequencies of 36.2% and 74.8% at two adjacent low-score and high-score sites, respectively, in the *MCM6* gene; the sites were separated by merely 29 bp (Fig. 8), highlighting the importance of target site choice. On average, the high-score sites and low-score sites produced frameshifting indels at frequencies of 79.3% and 42.5%, respectively (Student's t-test, $p < 0.001$). In a diploid cell or organism, the probability of obtaining null clones would be 62.8% (= 0.793 x 0.793) and 18.1% (= 0.425 x 0.425), respectively, strikingly similar to our two extreme-case estimations of 64% and 16% described above. As expected, the out-of-frame scores were reliable predictors of the frequencies of frameshifting indels (Pearson coefficient = 0.934) (Fig. 2c). To demonstrate the usefulness of our scoring system further, we tested 68 new RGENs that target different genes in yet another human cell line, HeLa (Table 5).

[185] Table 5

[Table 5]

Gene	Target site (5' to 3')	Number of sequence reads	Insertion	Deletion	Frequency of out-of-frame deletions (%)	Frequency of out-of-frame indels (%)	Microho-mology score[a]	Out-of-frame score[b]
ABL1	TGGGGCTGGATAATGGAG CGTGG (SEQ ID NO: 40)	3777	630	848	89.8704	93.712	5895	67.68447837
ACK	CGGTCCAACAACGATCCC AGAGG (SEQ ID NO: 41)	2374	306	1112	74.1007	79.2666	4429	61.21020546
ALK	CTGTGACCACGGGACGGT GCTGG (SEQ ID NO: 42)	4753	905	2248	66.1922	74.3102	5617	66.22752359
ARG	TCCATCTCGCTCAGGTAC GAGGG (SEQ ID NO: 43)	4316	985	2188	80.8044	86.0384	4220	69.43127962
AXL	GTCCCGTGTCGGAAAGCT GCAGG (SEQ ID NO: 44)	3514	494	1870	61.6043	68.5702	4729	55.25481074
BLK	ACTACACCGCTATGAATG ATCGG (SEQ ID NO: 45)	4121	1286	1280	81.4844	90.0624	4684	56.85311699
BRK	CCCAGAGGCCCACATACT TGGGG (SEQ ID NO: 46)	3380	913	1229	55.9805	74.2297	5984	61.1631016
CCK4	ACATGCCGCTATTTGAGC CACGG (SEQ ID NO: 47)	3946	133	794	55.9194	60.1942	4259	62.15073961
CSK	CTGACCGACCCCTAGACC GCAGG (SEQ ID NO: 48)	4102	1053	1715	82.7405	88.7283	5058	64.84776592
CTK	GCGGAAACACGGGACCAA GTCGG (SEQ ID NO: 49)	4469	376	1571	78.9306	81.1505	6340	69.95268139
DDR2	CCCCAGTGCTCGGTTTGT CACGG (SEQ ID NO: 50)	6486	1082	3531	84.3104	87.5569	5379	63.32031976
EGFR	CAAAGCTGTATTTGCCCT CGGGG (SEQ ID NO: 51)	4302	194	688	67.0058	73.2426	3892	57.34840699
EphA1	GCTCCAATTGGATCTACC GCGGG (SEQ ID NO: 52)	3762	317	2322	70.801	73.7779	4049	67.64633243
EphA10	TGGACCGGCGCAGGTCTC CATGG (SEQ ID NO: 53)	3575	754	774	71.3178	85.0785	5892	64.69789545
EphA2	AGGCTCCGAGTAGCGCAC ACTGG (SEQ ID NO: 54)	3700	696	727	77.7166	88.2642	5328	73.40465465
EphA3	TTGTCGACCAGGTTTCTA CAAGG (SEQ ID NO: 55)	2132	608	636	87.1069	92.0418	3497	69.48813269
EphA4	AACACCGAGATCCGGGAT GTAGG (SEQ ID NO: 56)	5136	287	2520	85.2381	85.1087	4003	68.99825131
EphA5	ACTGCAGCGCCGAAGGGG AGTGG (SEQ ID NO: 57)	4830	109	1800	67.2778	67.7842	6062	62.27317717
EphA6	TCTCTCAATACGAATTCT TGAGG (SEQ ID NO: 58)	3660	344	1357	52.5424	59.3768	4342	63.79548595
EphA7	CACCTGGTATGTTCGTAT CGGGG (SEQ ID NO: 59)	6125	1850	2738	89.2988	92.6548	4648	74.44061962
EphB	CACATGCATCCCCAACGC	3688	361	2105	71.6865	74.2092	4395	61.592719

[186]

1	AGAGG (SEQ ID NO: 60)							
EphB2	GGCTACGGACCAAGTTTATCCGG (SEQ ID NO: 61)	3553	49	537	68.9013	70.9898	3974	59.33568193
EphB4	GCAGAATATTCGGACAAACACGG (SEQ ID NO: 62)	4113	1337	1722	90.0697	93.9523	4455	77.08193042
EphB6	CTTCACCCTTTACTACCGTCAGG (SEQ ID NO: 63)	4867	472	2010	89.7512	90.5318	4798	67.27803251
FER	AGACTGGGAATTACGGTTACTGG (SEQ ID NO: 64)	4619	172	2246	67.4978	67.9487	4468	61.01163832
FES	GGAGGCCGAGCTTCGTCTACTGG (SEQ ID NO: 65)	3287	75	756	32.8042	38.7485	4584	48.58202443
FGFR1	CTCTGCATGGTTGACCGTTCTGG (SEQ ID NO: 66)	4070	210	1386	83.4776	83.7719	4649	67.84254678
FGFR3	CGGCAACTACACCTGCGTCGTGG (SEQ ID NO: 67)	2250	299	1171	65.585	70.9524	4392	48.13296903
FGFR4	AACTCCCATAGTGGGTCGAGAGG (SEQ ID NO: 68)	6126	204	659	62.3672	70.2202	4744	57.25126476
FGR	GCAGCTGTACGCCGTGGTGTCGG (SEQ ID NO: 69)	4216	175	1686	45.255	49.2746	5234	36.35842568
FMS	ATCTACTTGATCGAGGTTGAGGG (SEQ ID NO: 70)	6805	467	2273	53.5416	60.9489	4919	48.34315918
FRK	CTGGTCAGTTTGGCGAAGTATGG (SEQ ID NO: 71)	4682	537	699	81.9742	89.4013	4712	72.24108659
FYN	GGGACCTTGCGTACGAGAGGAGG (SEQ ID NO: 72)	4055	130	1897	66.5788	67.8836	4443	66.93675445
HCK	TGTCGCCCGCGTTGACTCTCTGG (SEQ ID NO: 73)	4822	200	420	86.6667	89.5161	3736	72.88543897
HER2/ErbB2	AGCTGGCGCCGAATGTATACCGG (SEQ ID NO: 74)	4921	121	1935	76.1757	77.0914	5021	69.94622585
IGF1R	TCAGTACGCCGTTTACGTCAAGG (SEQ ID NO: 75)	4857	1117	2543	65.0806	74.7268	3991	55.14908544
INSR	GAGAATTGCTCTGTCATCGAAGG (SEQ ID NO: 76)	5838	924	920	84.8913	91.5944	4280	67.52336449
ITK	AAGCGGACTTTAAAGTTCGAGGG (SEQ ID NO: 77)	5075	125	472	80.5085	84.0871	4851	78.51989281
JAK2	AGCAACAGAGCCTATCGGCATGG (SEQ ID NO: 78)	4060	254	1473	67.2098	70.3532	4379	66.31651062
JAK3	CTGGAAAGTCGCAGAAGGGCTGG (SEQ ID NO: 79)	3349	102	574	86.2369	86.9822	4551	74.29136454
KDR	TCCAGGTTTCCTGTGATCGTGGG (SEQ ID NO: 80)	5604	988	1684	61.1045	75	3825	63.34640523
KIT	TATTCTCATTCGTTTCATCCAGG (SEQ ID NO: 81)	5426	428	1633	55.2358	61.8147	5110	56.53620352
LCK	GAGCCTTCGTAGGTAACCAGTGG (SEQ ID NO: 82)	3159	141	680	82.9412	83.8002	4884	73.42342342
LMR1	GCCACCCGTCGACGTCCCCTGGG	3363	236	1810	78.5083	80.2053	8541	61.9716660

[187]

	(SEQ ID NO: 83)							8
LMR2	GCTCAGGAGCGTTGAACT TGAGG (SEQ ID NO: 84)	4756	1648	1807	68.9541	83.3864	4369	58.41153582
LTK	TGGCTCCAAGATACTAGG CGGGG (SEQ ID NO: 85)	4131	172	1195	82.3431	80.9802	5454	65.52988632
MER	CTATTCCCGGGACCTTTT CCAGG (SEQ ID NO: 86)	2890	135	1320	81.3636	82.6804	5269	58.94856709
MUSK	GCATAGCTACCAATAAGC ATGGG (SEQ ID NO: 87)	4871	154	2709	65.2639	66.2592	4309	54.42097935
PDGFRa	CAGCCTAAGACCAGGAAC GCCGG (SEQ ID NO: 88)	4452	353	2708	84.8227	85.7563	5043	71.30676185
PDGFRb	AGGGAACGTAGTTATCGT AAGGG (SEQ ID NO: 89)	3996	149	2407	55.7541	57.903	4091	53.99657785
PYK2	GGTCCTGAATCGTATTCT TGGGG (SEQ ID NO: 90)	4180	695	1995	77.594	82.3792	3720	57.31182796
RET	TGCTGGGTGATGCGGCCG GTGGG (SEQ ID NO: 91)	3179	305	1027	69.2308	75.0751	5776	63.78116343
RON	GTCATCGGGCCGGTTATG GTGGG (SEQ ID NO: 92)	3350	1133	1326	78.9593	88.2066	6432	62.18905473
ROR1	GCCATAGATGGTGGACCG AAAGG (SEQ ID NO: 93)	5172	571	2748	82.2416	84.9654	6204	57.62411348
ROS	TGAGGTGCACTAATAGAG GGTGG (SEQ ID NO: 94)	4098	503	1663	44.979	56.5559	3834	53.5732916
RYK	TATTGCCTTACATGAATT GGGGG (SEQ ID NO: 95)	6079	753	2584	67.8406	74.1984	4018	67.86958686
SRC	GTCTGACTTCGACAACGC CAAGG (SEQ ID NO: 96)	4141	232	1700	35.0588	41.2526	4157	44.84002887
SRM	CCACACTCCGAATTCGCC CTTGG (SEQ ID NO: 97)	1423	73	722	75.2078	77.1069	4392	73.97540984
SYK	GGTGATGTTGCCGAAAAA GAAGG (SEQ ID NO: 98)	3825	368	1474	57.9376	65.5809	4424	51.37884268
TIE1	CGCCTGTGGGACGGGACA CGGGG (SEQ ID NO: 99)	2050	437	657	64.5358	77.5137	9164	63.74945439
TIE2	CAGAGTTCATATTCTGTC CGAGG (SEQ ID NO: 100)	5063	1238	2267	68.8134	78.9444	4027	60.44201639
TNK1	GCAGTAGGTTGCGCGTAG CGAGG (SEQ ID NO: 101)	3497	1307	725	69.931	89.2224	7094	65.21003665
TRKB	GCCGTGGTACTCCGTGTG ATTGG (SEQ ID NO: 102)	4525	1080	1973	62.3923	74.8772	3748	68.72998933
TRKC	CATCAGCGTTGATGCAGT AGAGG (SEQ ID NO: 103)	5151	83	876	48.0594	50.9906	5474	54.74972598
TXK	GTTGTTTACCAGCCACAG CTGGG (SEQ ID NO: 104)	5371	1954	1682	66.4685	83.8284	4931	66.98438451
TYK2	GAACCGGCTGTGTACCGT TGTGG (SEQ ID NO: 105)	4569	87	466	86.0515	86.9801	5638	75.8957077
TYRO3	GGCCACACTAGCGTTGCT GCTGG (SEQ ID NO: 106)	4466	345	2254	60.9583	65.0635	4665	58.17792069
YES	TCAGGTCTGTATTTAATG GCTGG (SEQ ID NO: 107)	5584	1157	1364	80.9384	88.8933	4727	62.83054792

[a] Microhomology score $= \sum$ pattern score.

[b] Out-of-frame score $= \frac{\sum \text{pattern score of an out-of-frame deletion}}{\sum \text{pattern score}}$. (±35bp between target sites)

[188]

[189] Again, out-of-frame scores correlated well with the frequencies of frame shifting indels or deletions (Pearson coefficient = 0.717 or 0.732, respectively) (Fig. 2d). The frequencies of out-of-frame indels ranged from 38.7% to 94.0%. In a diploid human cell, the probability of obtaining null clones would range from 15.0% (= 0.387 x 0.387) to 88.4%, a 5.9-fold difference between the extreme cases. Most cancer cell lines including HeLa are multi-ploid (> 3n), making it more important to choose high-

score sites. It is expected that the scoring system would work even better for TALENs because TALENs induce microhomology-independent insertions much less frequently than do RGENs, as shown above. In addition, it was analyzed that the genotypes of 81 live-born mice carrying mutations that had been produced via TALENs or RGENs in our previous studies (Sung, Y.H. et al. *Genome research***24**, 125-131 (2014); Sung, Y.H. et al. *Nature biotechnology***31**, 23-24 (2013)). The frequencies of out-of-frame deletions correlated well with predicted scores (Pearson coefficient = 0.996) (Fig.9).

[190]

[191] Those skilled in the art will appreciate that the conceptions and specific embodiments disclosed in the foregoing description may be readily utilized as a basis for modifying or designing other embodiments for carrying out the same purposes of the present invention. Those skilled in the art will also appreciate that such equivalent embodiments do not depart from the spirit and scope of the invention as set forth in the appended Claims.

Claims

[Claim 1] A method of selecting a nuclease target sequence for gene knockout, comprising:
(a) providing a nuclease target sequence candidate;
(b) collecting information of microhomology present in the nuclease target sequence candidate; and
(c) predicting frequency of microhomology-associated out-of-frame deletion of the nuclease target sequence candidate based on the information of microhomology collected in step (b).

[Claim 2] The method according to claim 1, further comprising a step of comparing the frequency of microhomology-associated out-of-frame deletion predicted in step (c) with frequency of microhomology-associated out-of-frame deletion of other nuclease target sequence candidate.

[Claim 3] The method according to claim 1, wherein the information of microhomology comprises a size of microhomology sequence, a distance between two microhomology sequences, and sequence information of the microhomology sequence.

[Claim 4] The method according to claim 1, wherein the nuclease is selected from the group consisting of zinc finger nucleases (ZFNs), transcription-activator-like effector nucleases (TALENs), and RNA-guided engineered nucleases (RGENs).

[Claim 5] The method according to claim 1, wherein step (c) comprises:
calculating a pattern score, which is a score assigned to an expected deletion pattern of each of microhomologies present in the given nuclease target sequence candidate; and
calculating (i) a microhomology score, which is a sum of the pattern scores of all microhomologies in the given nuclease target sequence candidate and (ii) a out-of-frame score, which is a ratio of a score which is a sum of the pattern scores of microhomologies associated with out-of-frame deletion to the microhomology score, based on the calculated pattern score.

[Claim 6] The method according to claim 1, wherein the method comprises:
i) providing a nuclease target sequence candidate;
ii) examining, in the given nuclease target sequence, whether two identical sequences of at least 2 bp flanking a position expected to be cleaved by a nuclease are present in the target sequence to identify the

presence of microhomology;

iii) obtaining information of microhomology, when the microhomology is present in the target sequence, and repeating steps ii) and iii) one or more times;

iv) calculating a pattern score, which is a score assigned to an expected deletion pattern of each of microhomologies present in the given nuclease target sequence candidate; and

v) calculating (i) a microhomology score, which is a sum of the pattern scores of all microhomologies in the given nuclease target sequence candidate and (ii) a out-of-frame score, which is a ratio of a score which is a sum of the pattern scores of microhomologies associated with out-of-frame deletion to the microhomology score.

[Claim 7] The method according to claim 5 or 6, wherein the pattern score is calculated using Equation 1:

[Equation 1]

Pattern score = S X $\exp(-\triangle / W_{length})$,

wherein,

S is a microhomology index that corresponds to the size and base pairing energy of the microhomology sequence;

$\triangle$ is a distance between initiation sites located at 5' position of each microhomology sequence or a distance between terminal sites located at 3' position of each microhomology sequence of the two microhomology sequences (deletion length); and

W_{length} is a weight factor on a distance between the microhomology sequences.

[Claim 8] The method according to claim 5 or 6, wherein the microhomology score is calculated using Equation 2, and the out-of-frame score is calculated using Equation 3:

[Equation 2]

Microhomology score = $\sum$ pattern score,

wherein the microhomology score is a sum of pattern scores of the obtained all microhomologies;

[Equation 3]

Out-of-frame score = $\sum$ pattern score of out-of-frame deletion / Microhomology score ($\sum$ pattern score),

wherein $\sum$ pattern score of out-of-frame deletion is a sum of pattern scores of relevant microhomologies whose deletion length is not a multiple of 3.

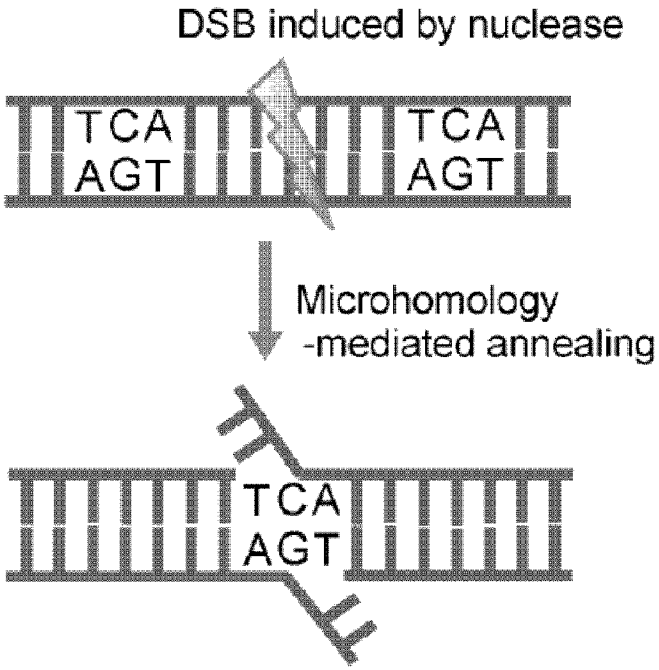
[Claim 9] The method according to claim 7, wherein, in Equation 1,
a)the microhomology index (S) is calculated by Equation 4 below; and
b)W_{length} is 20:
[Equation 4]
Microhomology index = (number of G and C in the microhomology sequence)*2 + (number of A and T bases in the microhomology sequence).

[Claim 10] A method of providing information for selecting a sequence having high efficiency of out-of-frame deletion by a nuclease, comprising:
(a) providing a nuclease target sequence candidate;
(b) collecting information of microhomology present in the nuclease target sequence candidate; and
(c) predicting frequency of microhomology-associated out-of-frame deletion of the nuclease target sequence candidate based on the information of microhomology collected in step (b).

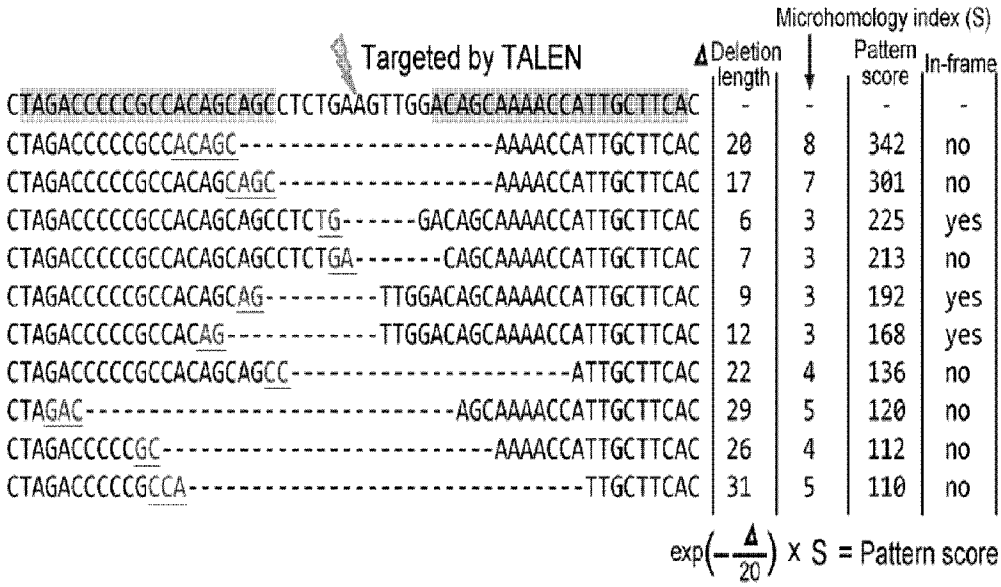
[Claim 11] A computer program capable of performing a method according to any one of claims 1 to 6.

[Claim 12] A computer-readable recording medium in which the program according to claim 11 is recorded.

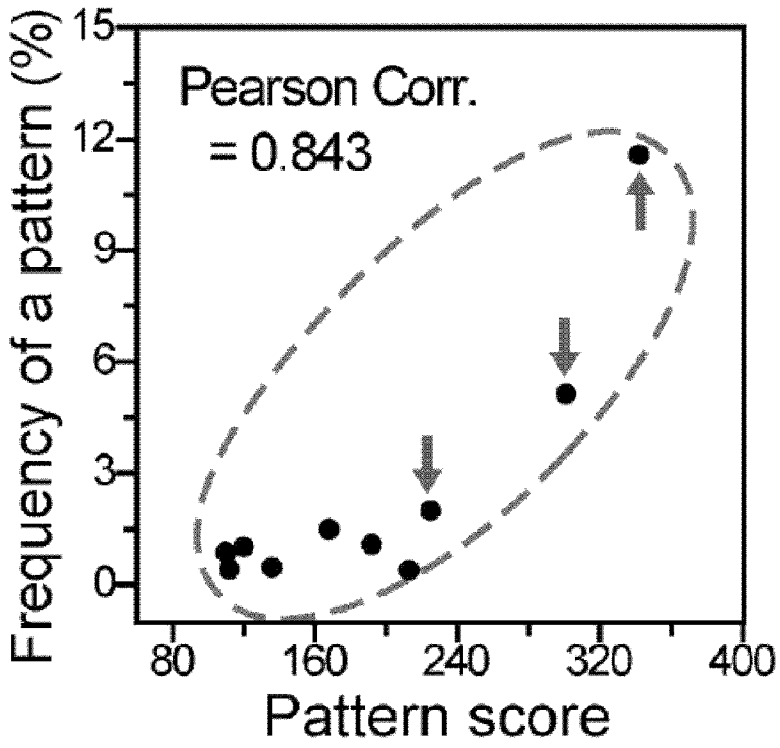
[Fig. 1a]


DSB induced by nuclease
TCA
AGT
TCA
AGT
Microhomology
-mediated annealing
TCA
AGT

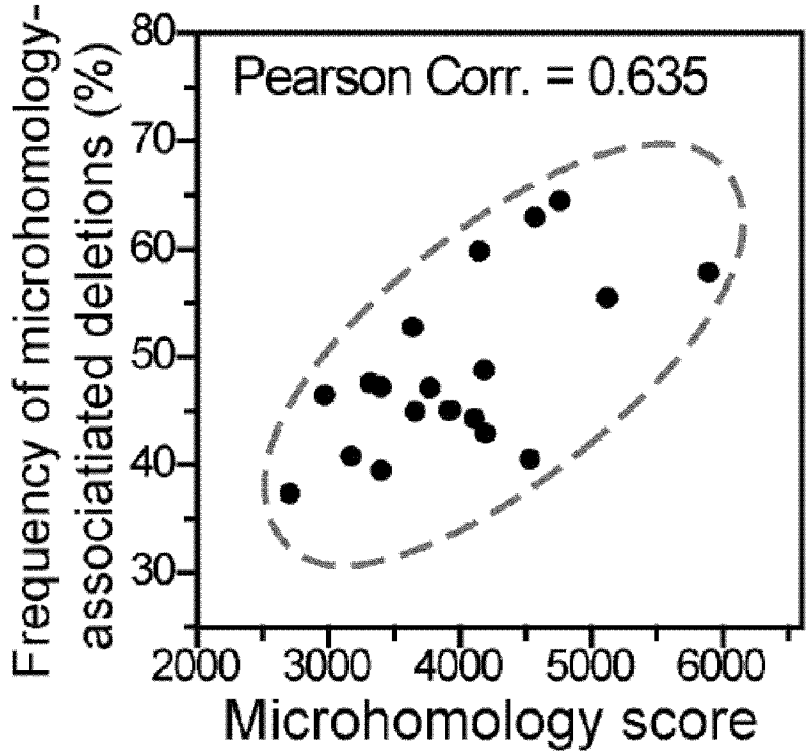
[Fig. 1b]


Targeted by TALEN
Microhomology index (S)
Δ Deletion length
Pattern score
In-frame
CTAGACCCCCGCCACAGCAGCCTCTGAAGTTGGACAGCAAAACCATTGCTTCAC - - - -
CTAGACCCCCGCCACAGC----------AAAACCATTGCTTCAC 20 8 342 no
CTAGACCCCCGCCACAGCAGC----------AAAACCATTGCTTCAC 17 7 301 no
CTAGACCCCCGCCACAGCAGCCTCTG------GACAGCAAAACCATTGCTTCAC 6 3 225 yes
CTAGACCCCCGCCACAGCAGCCTCTGA-------CAGCAAAACCATTGCTTCAC 7 3 213 no
CTAGACCCCCGCCACAGCAG---------TTGGACAGCAAAACCATTGCTTCAC 9 3 192 yes
CTAGACCCCCGCCACAG------------TTGGACAGCAAAACCATTGCTTCAC 12 3 168 yes
CTAGACCCCCGCCACAGCAGCC----------ATTGCTTCAC 22 4 136 no
CTAGAC----------AGCAAAACCATTGCTTCAC 29 5 120 no
CTAGACCCCCGC----------AAAACCATTGCTTCAC 26 4 112 no
CTAGACCCCCGCCA----------TTGCTTCAC 31 5 110 no
exp(-Δ/20) x S = Pattern score

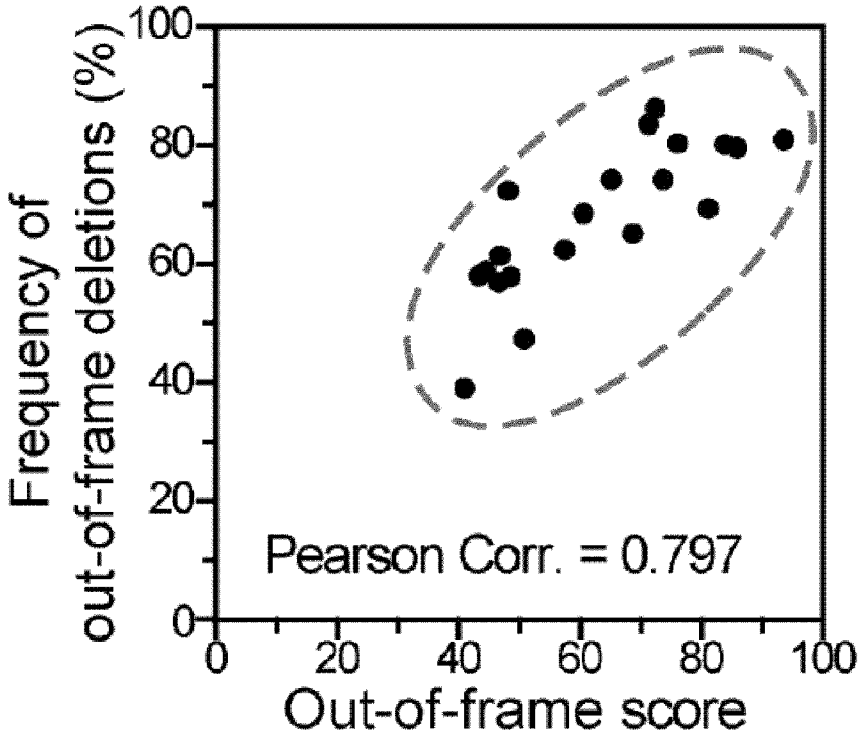
[Fig. 1c]


Pearson Corr.
= 0.843
Frequency of a pattern (%)
0
3
6
9
12
15
80
160
240
320
400
Pattern score

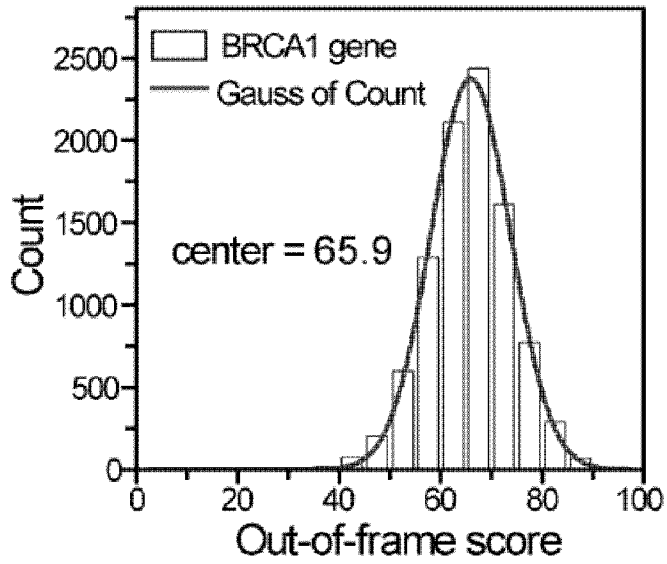
[Fig. 1d]


Pearson Corr. = 0.635
Frequency of microhomology-associatiated deletions (%)
30
40
50
60
70
80
2000
3000
4000
5000
6000
Microhomology score

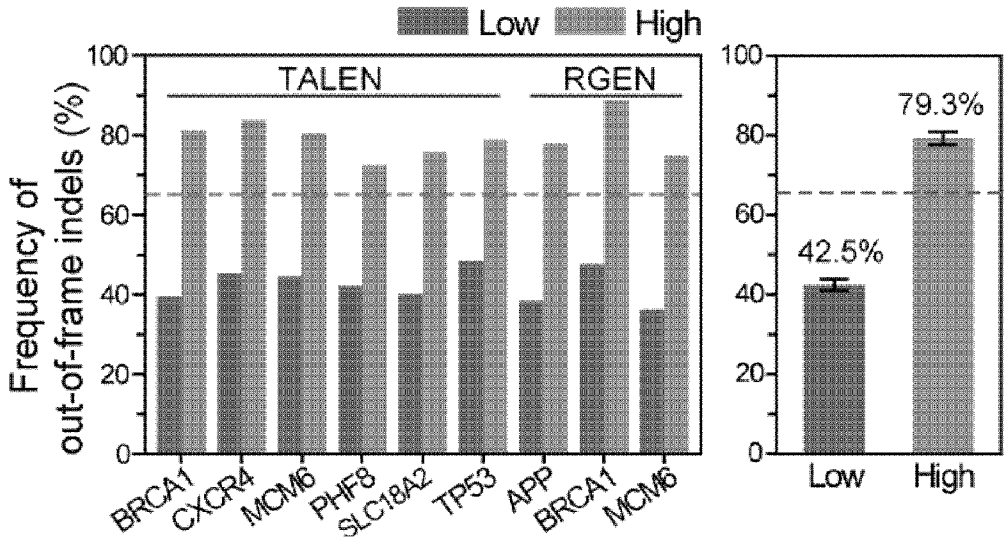
[Fig. 1e]


Frequency of out-of-frame deletions (%)
100
80
60
40
20
0
Pearson Corr. = 0.797
0
20
40
60
80
100
Out-of-frame score

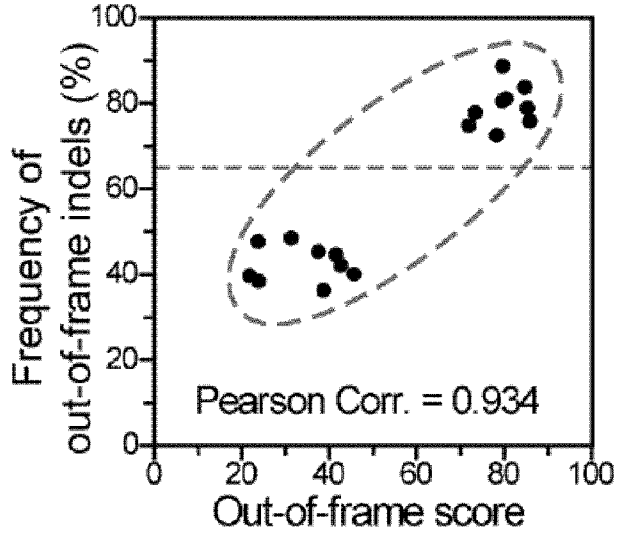
[Fig. 2a]


BRCA1 gene
Gauss of Count
Count
2500
2000
1500
1000
500
0
center = 65.9
0
20
40
60
80
100
Out-of-frame score

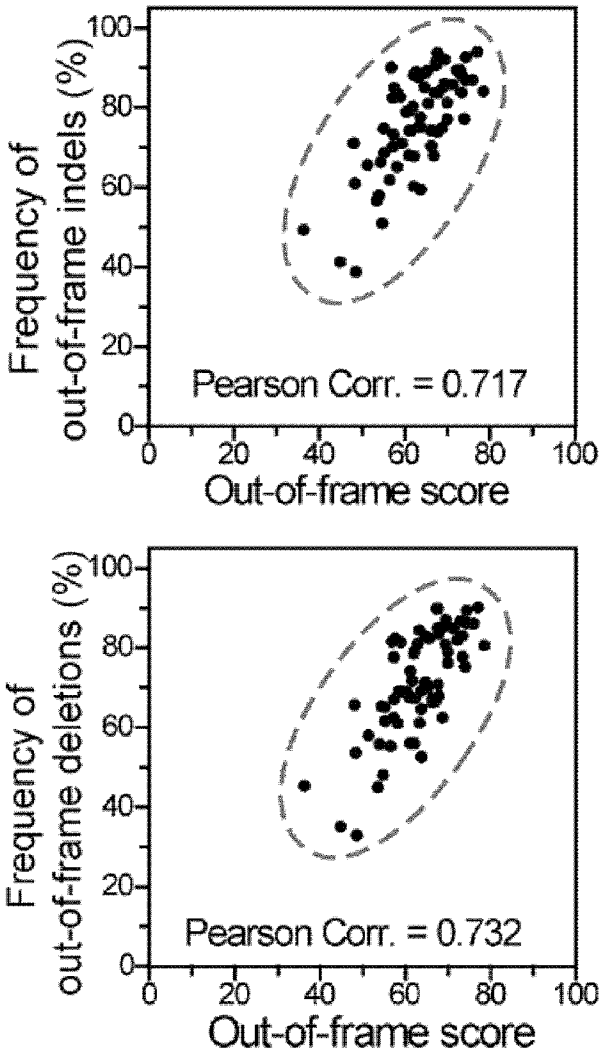
[Fig. 2b]


Low
High
TALEN
RGEN
Frequency of out-of-frame indels (%)
100
80
60
40
20
0
BRCA1
CXCR4
MCM6
PHF8
SLC18A2
TP53
APP
BRCA1
MCM6
79.3%
42.5%
Low
High

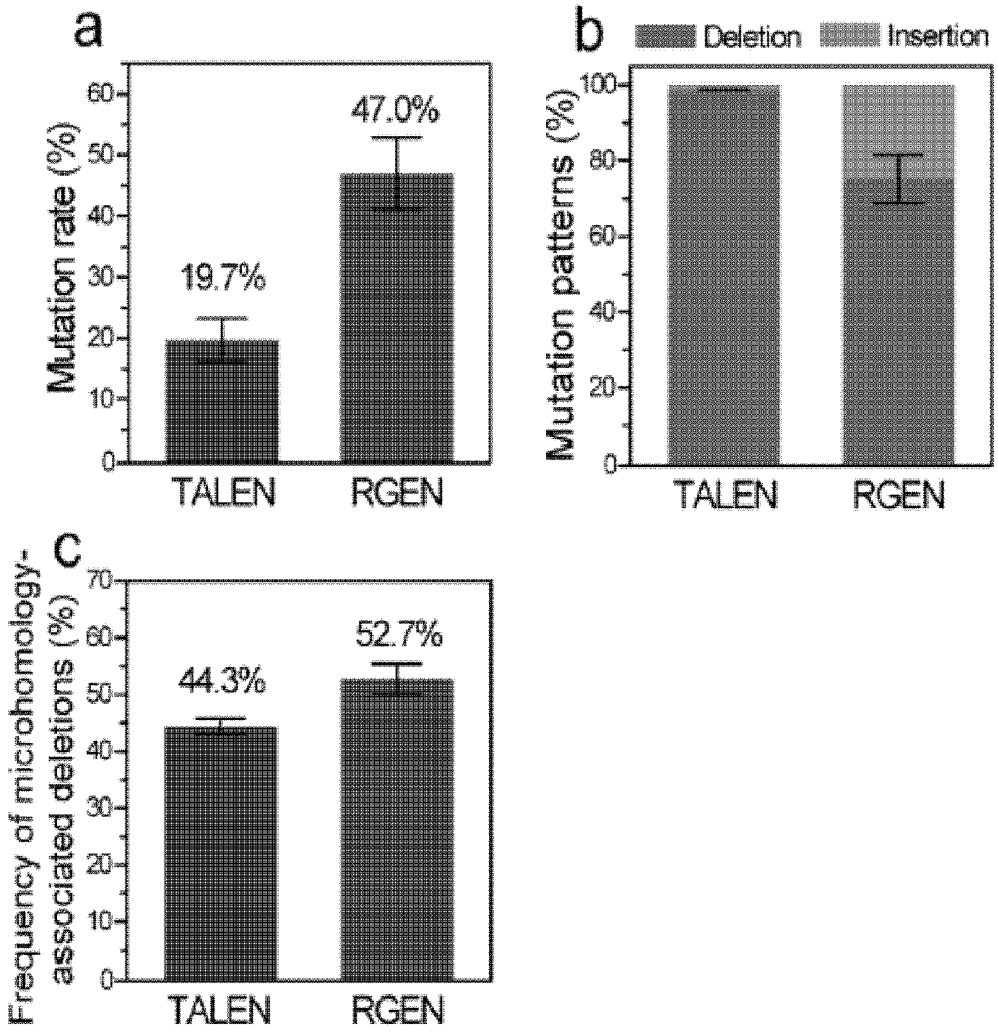
[Fig. 2c]


Frequency of out-of-frame indels (%)
100
80
60
40
20
0
Pearson Corr. = 0.934
0
20
40
60
80
100
Out-of-frame score

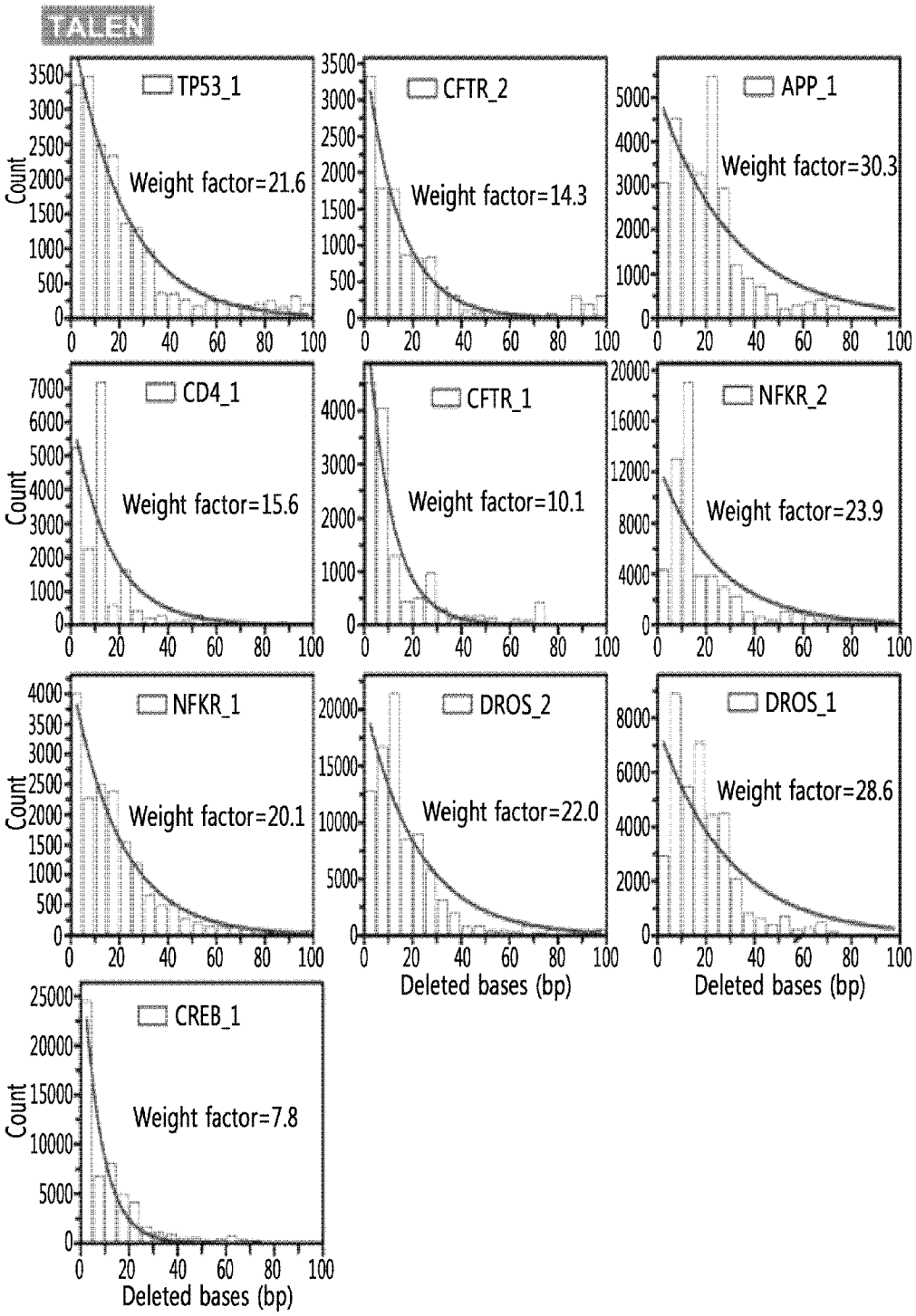
[Fig. 2d]


Frequency of out-of-frame indels (%)
Pearson Corr. = 0.717
Out-of-frame score
Frequency of out-of-frame deletions (%)
Pearson Corr. = 0.732
Out-of-frame score

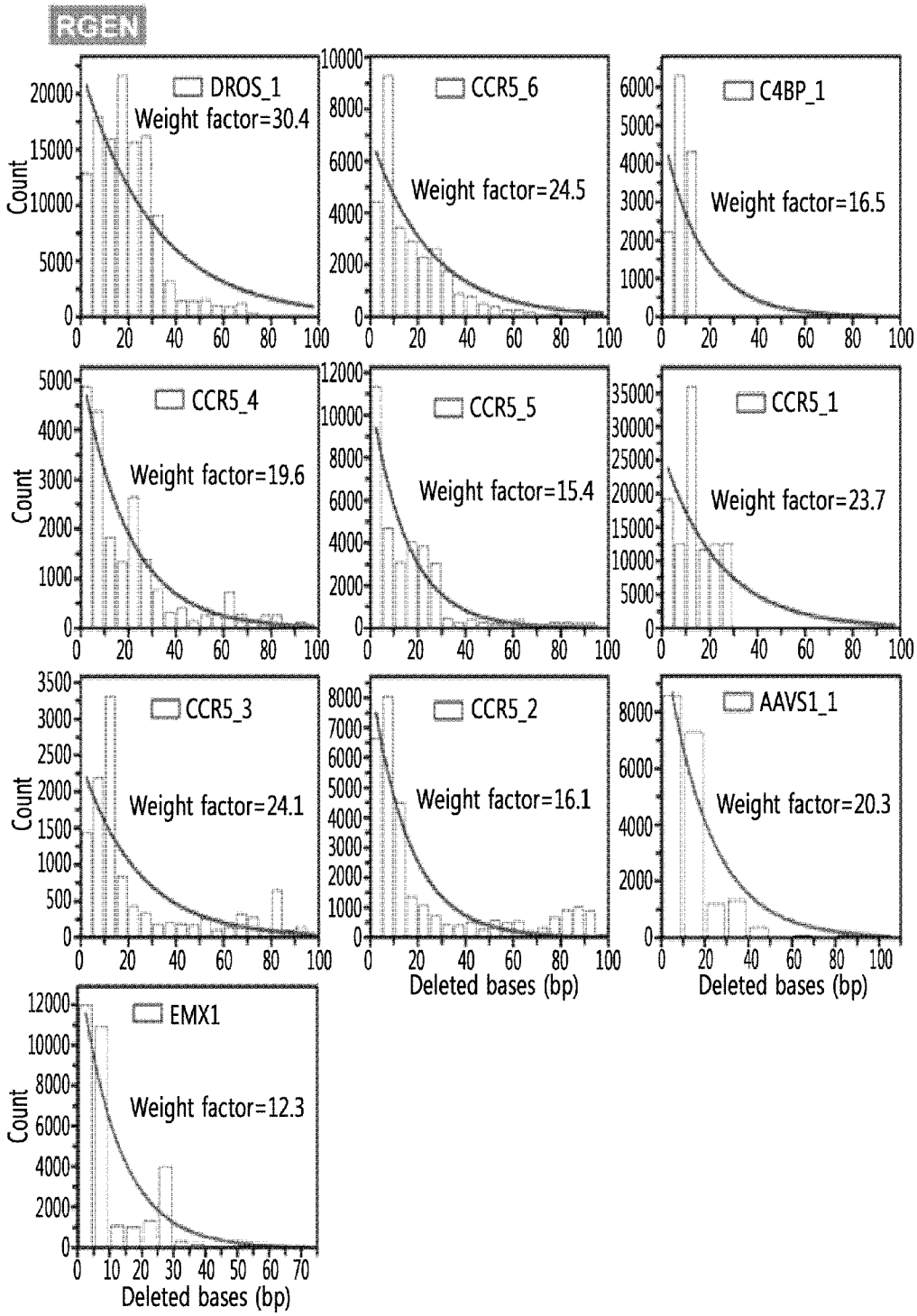
[Fig. 3]


a
Mutation rate (%)
60
50
40
30
20
10
0
19.7%
47.0%
TALEN
RGEN
b
Deletion
Insertion
Mutation patterns (%)
100
80
60
40
20
0
TALEN
RGEN
c
Frequency of microhomology-associated deletions (%)
70
60
50
40
30
20
10
0
44.3%
52.7%
TALEN
RGEN

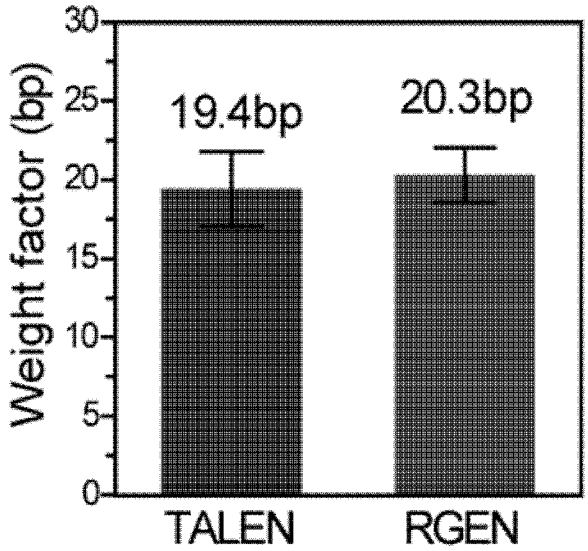
[Fig. 4a]


TALEN
TP53_1
Weight factor=21.6
CFTR_2
Weight factor=14.3
APP_1
Weight factor=30.3
CD4_1
Weight factor=15.6
CFTR_1
Weight factor=10.1
NFKR_2
Weight factor=23.9
NFKR_1
Weight factor=20.1
DROS_2
Weight factor=22.0
DROS_1
Weight factor=28.6
CREB_1
Weight factor=7.8
Count
Deleted bases (bp)

[Fig. 4b]


RGEN
DROS_1
Weight factor=30.4
CCR5_6
Weight factor=24.5
C4BP_1
Weight factor=16.5
CCR5_4
Weight factor=19.6
CCR5_5
Weight factor=15.4
CCR5_1
Weight factor=23.7
CCR5_3
Weight factor=24.1
CCR5_2
Weight factor=16.1
AAVS1_1
Weight factor=20.3
EMX1
Weight factor=12.3
Count
Deleted bases (bp)

[Fig. 4c]



[Fig. 5a]

```
#!/usr/bin/python2.7
from math import exp
from re import findall

seq='GGAGGAAGGGCCTGAGTCCGAGCAGAAGAAGAAGGGCTCCCATCACATCAACCGGTGGCG'   # The
length of sequence is recommend within 60~80 bases.
print seq

length_weight=20.0
left=30      # Insert the position expected to be broken.
right=len(seq)-int(left)
print 'length of seq = '+str(len(seq))

file_temp=open("1.before removing duplication.txt", "w")
for k in range(2,left)[::-1]:
    for j in range(left,left+right-k+1):
        for i in range(0,left-k+1):
            if seq[i:i+k]==seq[j:j+k]:
                length=j-i
                file_temp.write(seq[i:i+k]+'\t'+str(i)+'\t'+str(i+k)+'\t'+str(j)+'\t'+str(j+k)+'\t'+str(length)+'\n')
file_temp.close()

### After searching out all microhomology patterns, duplication should be removed!!
f1=open("1.before removing duplication.txt", "r")
s1=f1.read()
```

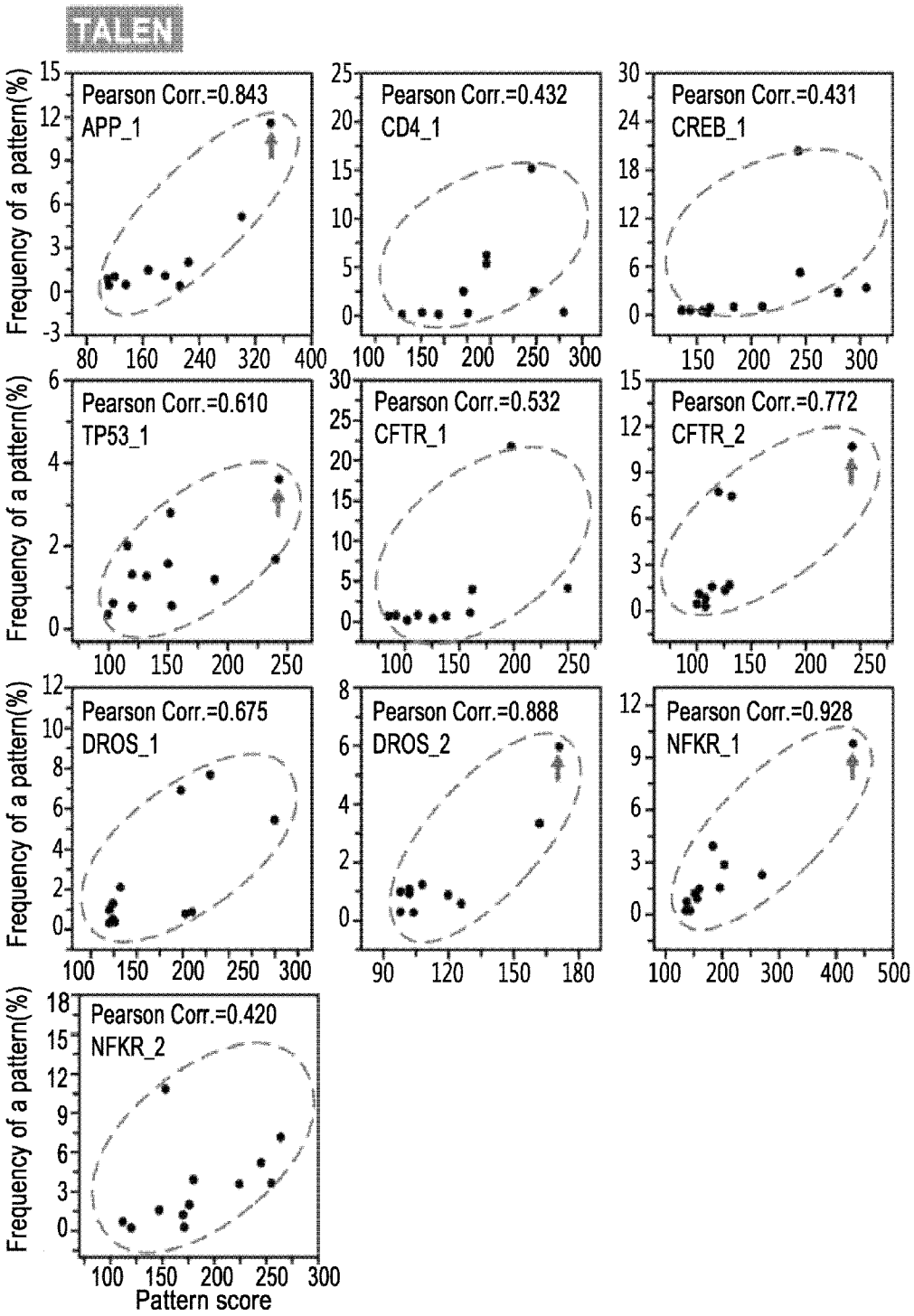
[Fig. 5b]

```
f2=open("2.all microhomology patterns.txt", "w") #After removing duplication
f2.write(seq+'\t'+'microhomology\t'+'deletion length\t'+'score of a pattern\n')

if s1!="":
    list_f1=s1.strip().split('\n')
    sum_score_3=0
    sum_score_not_3=0

    for i in range(len(list_f1)):
        n=0
        score_3=0
        score_not_3=0
        line=list_f1[i].split('\t')
        scrap=line[0]

        left_start=int(line[1])
        left_end=int(line[2])
        right_start=int(line[3])
        right_end=int(line[4])
        length=int(line[5])

    for j in range(i):
        line_ref=list_f1[j].split('\t')
        left_start_ref=int(line_ref[1])
        left_end_ref=int(line_ref[2])
        right_start_ref=int(line_ref[3])
        right_end_ref=int(line_ref[4])
```

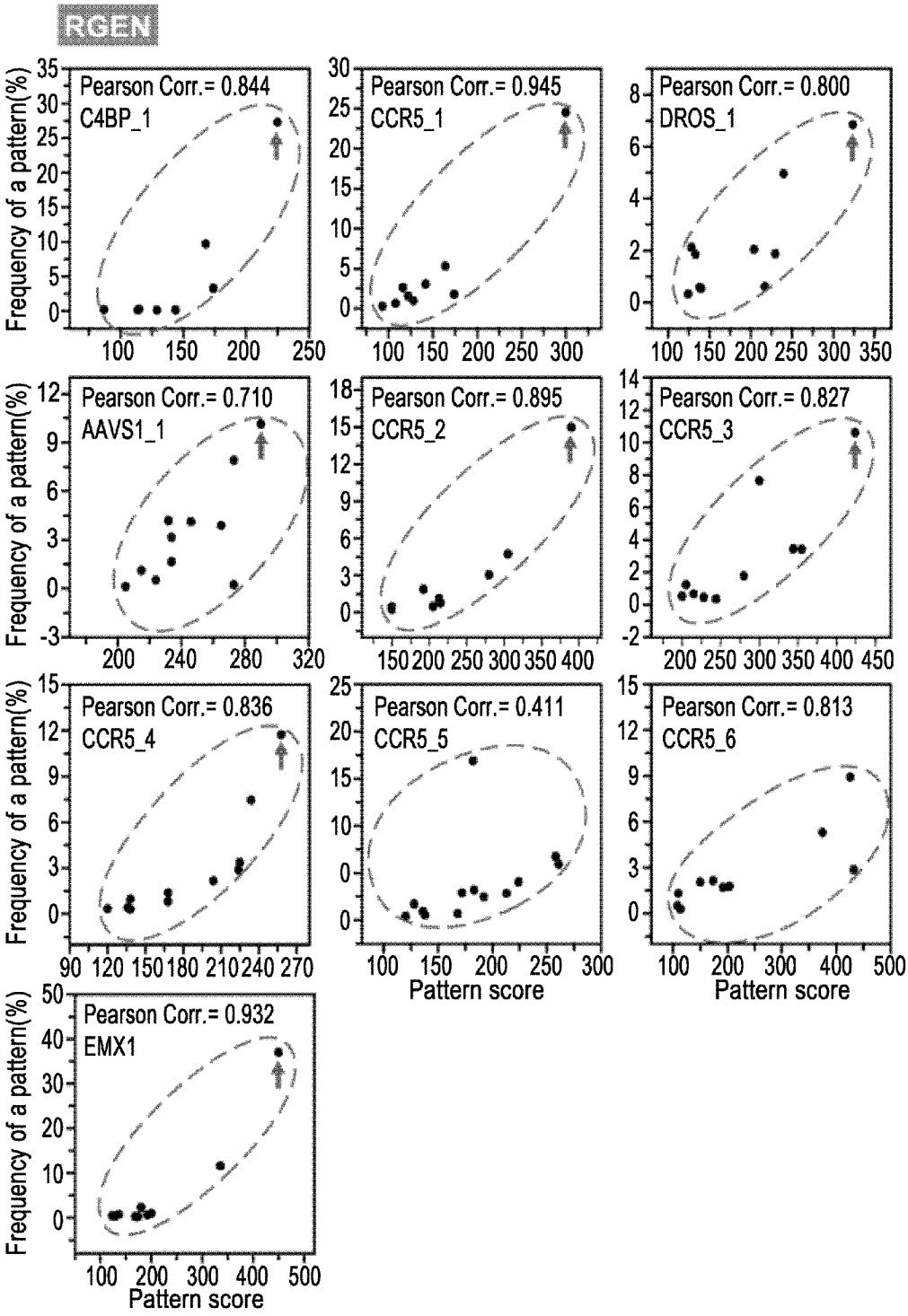
[Fig. 5c]

```
            if (left_start >= left_start_ref) and (left_end <= left_end_ref) and (right_start >= right_start_ref) and
(right_end <= right_end_ref):
                if (left_start - left_start_ref)==(right_start - right_start_ref) and (left_end -
left_end_ref)==(right_end - right_end_ref):
                    n+=1
            else: pass

        if n == 0:
            if (length % 3)==0:
                length_factor = round(1/exp((length)/(length_weight)),3)
                num_GC=len(findall('G',scrap))+len(findall('C',scrap))
                score_3=100*length_factor*((len(scrap)-num_GC)+(num_GC*2))

            elif (length % 3)!=0:
                length_factor = round(1/exp((length)/(length_weight)),3)
                num_GC=len(findall('G',scrap))+len(findall('C',scrap))
                score_not_3=100*length_factor*((len(scrap)-num_GC)+(num_GC*2))

            f2.write(seq[0:left_end]+'-
'*length+seq[right_end:]+'\t'+scrap+'\t'+str(length)+'\t'+str(100*length_factor*((len(scrap)-
num_GC)+(num_GC*2)))+'\n')
        sum_score_3+=score_3
        sum_score_not_3+=score_not_3

    print 'Microhomology score = ' + str(sum_score_3+sum_score_not_3)
    print 'Out-of-frame score = ' + str((sum_score_not_3)*100/(sum_score_3+sum_score_not_3))
f1.close()
f2.close()

# The row of output file is consist of (full sequence, microhomology scrap, deletion length, score of pattern).
```

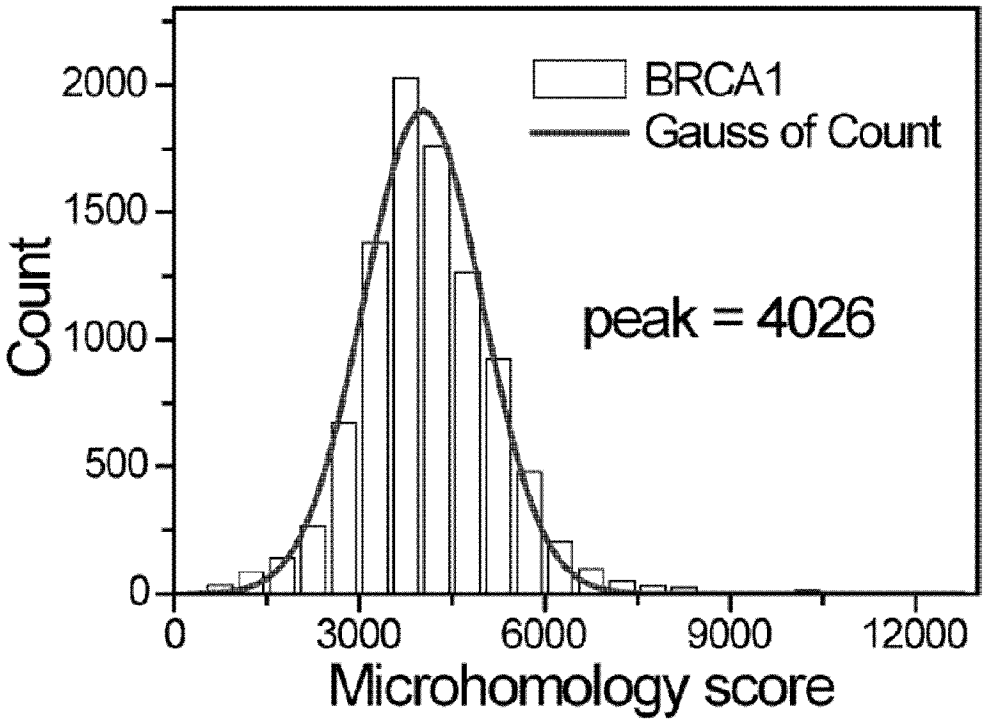
[Fig. 6a]


TALEN
Pearson Corr.=0.843
APP_1
Pearson Corr.=0.432
CD4_1
Pearson Corr.=0.431
CREB_1
Pearson Corr.=0.610
TP53_1
Pearson Corr.=0.532
CFTR_1
Pearson Corr.=0.772
CFTR_2
Pearson Corr.=0.675
DROS_1
Pearson Corr.=0.888
DROS_2
Pearson Corr.=0.928
NFKR_1
Pearson Corr.=0.420
NFKR_2
Frequency of a pattern(%)
Pattern score

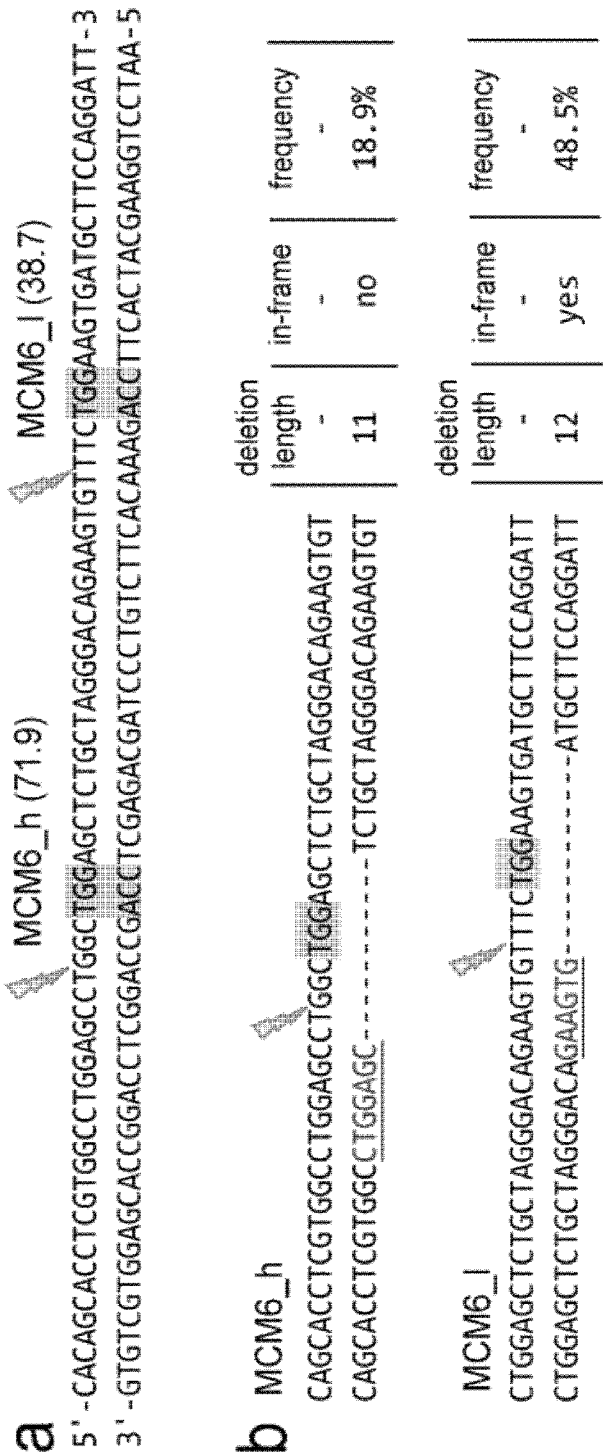
[Fig. 6b]


RGEN
Pearson Corr.= 0.844
C4BP_1
Pearson Corr.= 0.945
CCR5_1
Pearson Corr.= 0.800
DROS_1
Pearson Corr.= 0.710
AAVS1_1
Pearson Corr.= 0.895
CCR5_2
Pearson Corr.= 0.827
CCR5_3
Pearson Corr.= 0.836
CCR5_4
Pearson Corr.= 0.411
CCR5_5
Pearson Corr.= 0.813
CCR5_6
Pearson Corr.= 0.932
EMX1
Frequency of a pattern(%)
Pattern score

[Fig. 7]


BRCA1
Gauss of Count
peak = 4026
Count
2000
1500
1000
500
0
0
3000
6000
9000
12000
Microhomology score

[Fig. 8]

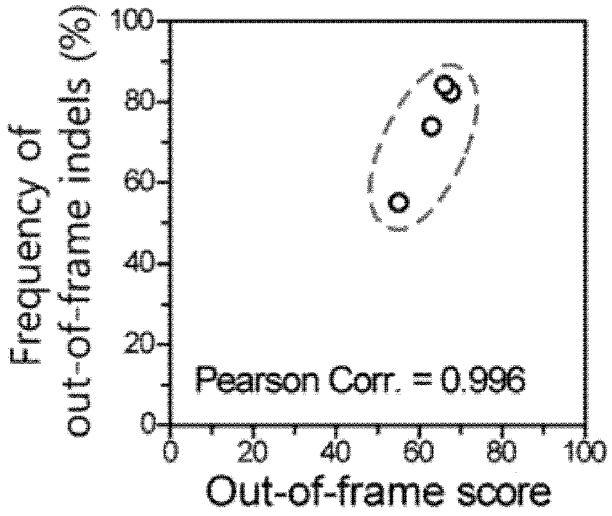

a
MCM6_h (71.9)
MCM6_I (38.7)
5'-CACAGCACCTCGTGGCCTGGAGCCTGGCTGGAGCTCTGCTAGGGACAGAAGTGTTCTGGAAGTGATGCTTCCAGGATT-3
3'-GTGTCGTGGAGCACCGGACCTCGGACCGACCTCGAGACGATCCCTGTCTTCACAAAGACCTTCACTACGAAGGTCCTAA-5
b
MCM6_h
deletion length
in-frame
frequency
CAGCACCTCGTGGCCTGGAGCCTGGCTGGAGCTCTGCTAGGGACAGAAGTGT
-
-
-
CAGCACCTCGTGGCCTGGAGC----------TCTGCTAGGGACAGAAGTGT
11
no
18.9%
MCM6_I
deletion length
in-frame
frequency
CTGGAGCTCTGCTAGGGACAGAAGTGTTCTGGAAGTGATGCTTCCAGGATT
-
-
-
CTGGAGCTCTGCTAGGGACAGAAGTG----------ATGCTTCCAGGATT
12
yes
48.5%

[Fig. 9]

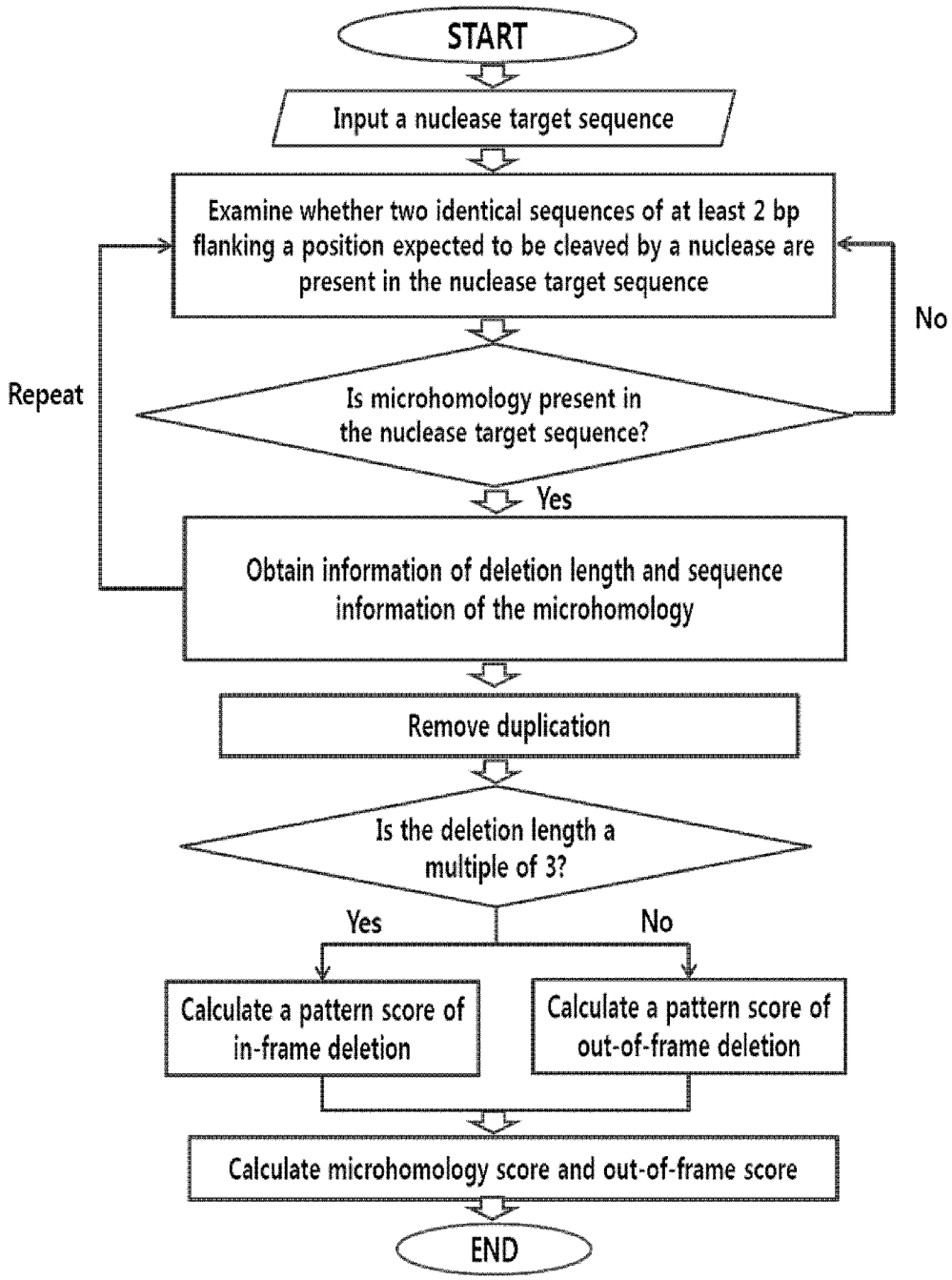
a

Engineered nucleases	TALENs		RGENs	
Gene	Pibf1	Sepw1	Foxn1	Prkdc
Live-born mutant mice	12	6	33	30
Number of out-of-frame	11	10	58	34
Number of in-frame	9	2	10	12
Frequency of out-of-frame indels (%)	55%	83.3%	85.3%	73.9%
Out-of-frame score	55.1	66.9	67.7	63.0

b



[Fig. 10]


START
Input a nuclease target sequence
Examine whether two identical sequences of at least 2 bp flanking a position expected to be cleaved by a nuclease are present in the nuclease target sequence
No
Repeat
Is microhomology present in the nuclease target sequence?
Yes
Obtain information of deletion length and sequence information of the microhomology
Remove duplication
Is the deletion length a multiple of 3?
Yes
No
Calculate a pattern score of in-frame deletion
Calculate a pattern score of out-of-frame deletion
Calculate microhomology score and out-of-frame score
END

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2015/004132

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68(2006.01)i, G06F 19/22(2011.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q 1/68; C12N 15/01; C12N 5/04; G06F 19/22

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: microhomology, deletion, frameshift, nuclease, target, algorithm, site-specific, CAS9/CRISPR, RGENs, TALENs, ZFNs

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2012-0149115 A1 (KIM JIN SOO et al.) 14 June 2012 (see abstract, claims, [5], [19], [22], [23], [42], [57], Figs. 6, 9 & 10)	1-10
A	Morton, J., et al., Proc. Natl. Acad. Sci. USA, Vol.103, pp.16370-16375. (2006) "Induction and repair of zinc-finger nuclease-targeted double-strand breaks in Caenorhabditis elegans somatic cells" (see the entire document)	1-10
A	Qi, Y., et al., Genome Research, Vol.23, pp.547-554. (2013) "Increasing frequencies of site-specific mutagenesis and gene targeting in Arabidopsis by manipulating DNA repair pathways" (see the entire document)	1-10
A	McVey, M. & Lee, S. E., Trends Genet. Vol.24, pp.529-538. (2008) "MMEJ repair of double-strand breaks (director`s cut): deleted sequences and alternative endings" (see the entire document)	1-10
A	Stephens, PJ., et al., Nature, Vol.462, pp.24-31. (Dec 2009) "Complex landscapes of somatic rearrangement in human breast cancer genomes" (see abstract; pages 1007-1008)	1-10

☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents,such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 August 2015 (10.08.2015)	**10 August 2015 (10.08.2015)**

Name and mailing address of the ISA/KR
International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 35208, Republic of Korea
Facsimile No. +82-42-472-7140

Authorized officer
SHIN, Weon Hye
Telephone No. +82-42-481-8298

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2015/004132

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Schroeder, Jan., et al., Bioinformatics, (22 Jan 2014) "Socrates: Identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads" (see abstract; page 3, 2nd para of the right column; Fig. 1).	1-10
T	Bae, S., et al., Nat. Methods, Vol.11(7), pp.705-706. (Jul 2014) "Microhomology-based choice of Cas9 nuclease target sites" (see the entire document)	1-10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2015/004132

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.: 11
because they relate to subject matter not required to be searched by this Authority, namely:
Claim 11 is directed to a computer program which does not require search under Article 17(2)(a)(i) & Rule 39.1(vi) PCT.

2. ☒ Claims Nos.: 12
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
Claim 12 is worded in reference to claim 11, for which international search has not been carried out as described above. Claim 12 does not meet the requirement of PCT Article 6.

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required addtional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of any additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/KR2015/004132

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2012-0149115 A1	14/06/2012	JP 2012-529287A	22/11/2012
		KR 10-2010-0133319 A	21/12/2010
		WO 2010-143917 A2	16/12/2010
		WO 2010-143917 A3	21/04/2011