



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2018년01월30일
 (11) 등록번호 10-1823661
 (24) 등록일자 2018년01월24일

- (51) 국제특허분류(Int. Cl.)
C12Q 1/68 (2006.01) *G06F 19/22* (2017.01)
- (52) CPC특허분류
C12Q 1/6869 (2013.01)
G06F 19/22 (2013.01)
- (21) 출원번호 10-2015-0058304
- (22) 출원일자 2015년04월24일
 심사청구일자 2015년04월24일
- (65) 공개번호 10-2015-0123195
- (43) 공개일자 2015년11월03일
- (30) 우선권주장
 61/983,988 2014년04월24일 미국(US)
 1020140101133 2014년08월06일 대한민국(KR)
- (56) 선행기술조사문헌
 Trends Genet., Vol. 30, No. 3, pp. 85-94
 (2014.03.)
 Mol. Ther. Nucleic Acids., Vol. 5, No. 6,
 e323 doi: 10.1038/mtna.2016.35. (2016.06.14.)
 Nucleic Acids Research, Vol. 41, No. 19, e181
 doi:10.1093/nar/gkt716 (2013)

- (73) 특허권자
 기초과학연구원
 대전광역시 유성구 엑스포로 55(도룡동)
- (72) 발명자
 김진수
 서울특별시 관악구 관악로 1, 서울대학교 자연과
 학대학 화학부 503동 106호
- 배상수
 서울특별시 관악구 관악로 1, 서울대학교 자연과
 학대학 화학부 503동 106호
- (74) 대리인
 유미특허법인

전체 청구항 수 : 총 11 항

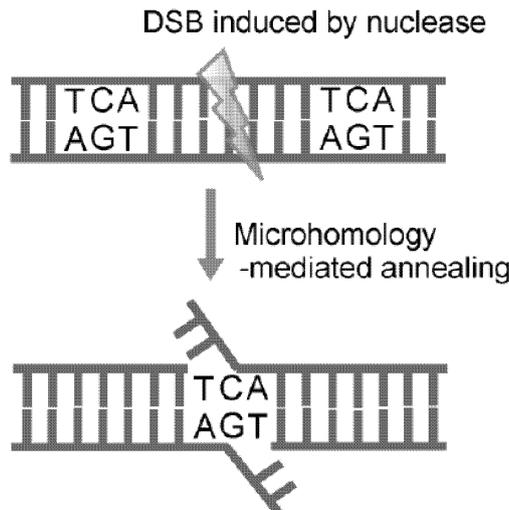
심사관 : 박정웅

(54) 발명의 명칭 **마이크로 상동 기반의 유전자 녹아웃을 위한 뉴클레아제 표적 서열을 확인하는 방법**

(57) 요약

본 발명은 마이크로 상동(microhomology)을 기반으로 유전자 녹아웃을 위한 뉴클레아제 표적 서열을 확인하는 방법에 관한 것이다.

대표도 - 도1a



이 발명을 지원한 국가연구개발사업

과제고유번호 IBS-R021-D1-2014-a00

부처명 미래창조과학부

연구관리전문기관 기초과학연구원

연구사업명 본원연구단 연구사업

연구과제명 인공제한효소를 이용한 동식물 및 인간 유전체 교정

기 여 율 1/1

주관기관 기초과학연구원

연구기간 2014.03.18 ~ 2014.12.31

공지예외적용 : 있음

명세서

청구범위

청구항 1

- (a) 뉴클레아제 표적 후보 서열을 제공하는 단계;
- (b) 주어진 뉴클레아제 표적 후보 서열에 존재하는 마이크로 상동 (microhomology)에 대한 정보를 수집하는 단계; 및
- (c) 상기 (b) 단계에서 수집된 마이크로 상동에 대한 정보를 바탕으로 마이크로 상동과 결부된 out-of-frame 형태의 결실(microhomology-associated out-of-frame deletion)의 빈도를 예측하는 단계를 포함하는, 유전자 녹아웃을 위한 뉴클레아제 표적 서열을 선택하는 방법.

청구항 2

제1항에 있어서,

상기 방법은 (c) 단계에서 예측된, 주어진 뉴클레아제 표적 후보 서열의 마이크로 상동과 결부된 out-of-frame 형태의 결실의 빈도와 다른 뉴클레아제 표적 후보 서열의 마이크로 상동과 결부된 out-of-frame 형태의 결실의 빈도를 비교하는 단계를 포함하는 것인, 방법.

청구항 3

제1항에 있어서,

상기 마이크로 상동에 대한 정보는 마이크로 상동 서열의 크기, 마이크로 상동 서열 간의 거리 및 마이크로 상동 서열의 서열 정보를 포함하는 것인, 방법.

청구항 4

제1항에 있어서, 상기 뉴클레아제는 ZFNs(zinc finger nucleases), TALENs(transcription-activator-like effector nucleases) 및 RGENs(RNA-guided engineered nucleases)로 구성된 군에서 선택되는 것인, 방법.

청구항 5

제1항에 있어서, 상기 (c) 단계는

주어진 뉴클레아제 표적 후보 서열에 존재하는 마이크로 상동의 예상 결실 패턴에 대한 스코어 (패턴 스코어)를 산출하는 단계; 및

산출된 패턴 스코어를 바탕으로, (i) 주어진 뉴클레아제 표적 후보 서열에 존재하는 전체 마이크로 상동의 패턴 스코어의 합인 마이크로 상동 스코어, 및 (ii) 이에 대한 out-of-frame 형태의 결실과 결부된 마이크로 상동의 패턴 스코어의 합의 비율 (out-of-frame 스코어)을 산출하는 단계를 포함하는, 방법.

청구항 6

i) 뉴클레아제 표적 후보 서열을 제공하는 단계;

ii) 주어진 뉴클레아제 표적 서열에서, 뉴클레아제 절단 예상 위치를 기준으로 양쪽 서열 부위에 2bp 이상의 동

일한 서열이 존재하는지를 확인하여 마이크로 상동의 존재를 확인하는 단계;

iii) 상기 표적 서열에 마이크로 상동이 존재하면, 존재하는 마이크로 상동에 대한 정보를 수집하고, 상기 ii) 및 iii) 단계를 1회 이상 반복하는 단계;

iv) 주어진 뉴클레아제 표적 후보 서열에 존재하는 마이크로 상동의 예상 결실 패턴에 대한 스코어 (패턴 스코어)를 산출하는 단계; 및

v) 산출된 패턴 스코어를 바탕으로, (i) 주어진 뉴클레아제 표적 후보 서열에 존재하는 전체 마이크로 상동의 패턴 스코어의 합인 마이크로 상동 스코어, 및 (ii) 이에 대한 out-of-frame 형태의 결실과 결부된 마이크로 상동의 패턴 스코어의 합의 비율 (out-of-frame 스코어)을 산출하는 단계를 포함하는,

유전자 녹아웃을 위한 뉴클레아제 표적 서열을 선택하는 방법.

청구항 7

제5항 또는 제6항에 있어서, 상기 패턴 스코어는 하기 수학적 식 1로 산출되는 것인, 방법:

[수학적 식 1]

$$\text{패턴 스코어} = S \times \exp(-\Delta / W_{\text{length}})$$

여기서,

S는 상기 마이크로 상동 서열의 크기 및 상기 마이크로 상동 서열의 염기쌍(base pairing) 에너지에 비례하는 마이크로 상동 인덱스(Microhomology index)이고,

Δ 는 두 마이크로 상동 서열의 5'에 위치한 시작점 사이의 거리 또는 두 마이크로 상동 서열의 3'에 위치한 말단 사이의 거리 (결실 길이)이고;

W_{length} 는 상기 마이크로 상동 서열 간의 거리에 대한 가중치임.

청구항 8

제5항 또는 제6항에 있어서, 상기 마이크로 상동 스코어는 하기 수학적 식 2로 산출되고, out-of-frame 스코어는 하기 수학적 식 3으로 산출되는 것인 방법:

[수학적 식 2]

$$\text{마이크로 상동 스코어} = \sum \text{패턴 스코어},$$

여기서, 상기 마이크로 상동 스코어는 얻어진 모든 마이크로 상동에 대한 패턴 스코어 값의 합임;

[수학적 식 3]

$$\text{Out-of-frame 스코어} = \sum \text{Out-of-frame 결실의 패턴 스코어} / \text{마이크로 상동 스코어} (\sum \text{패턴 스코어}),$$

여기서, \sum out-of-frame 결실의 패턴 스코어는 결실 길이가 3의 배수가 아닌 경우에 해당하는 마이크로 상동 서열의 패턴 스코어 값의 합임.

청구항 9

제7항에 있어서, 상기 수학적 식 1에서

a) 마이크로 상동 인덱스 (S)는 하기 수학적 식 4로 계산되는 것이고, b) W_{length} 는 20인, 방법:

[수학적 식 4]

마이크로 상동 인덱스(S) = (마이크로 상동 서열의 G 및 C의 수)*2 + (마이크로 상동 서열의 A 및 T의 수).

청구항 10

삭제

청구항 11

제1항 내지 제6항 중 어느 한 항의 단계를 수행하는, 컴퓨터 판독 가능한 기록 매체에 저장된 컴퓨터 프로그램.

청구항 12

제11항의 프로그램이 기록된 컴퓨터 판독 가능한 기록 매체.

발명의 설명

기술 분야

[0001] 본 발명은 마이크로 상동(microhomology)을 기반으로 유전자 녹아웃을 위한 뉴클레아제 표적 서열을 확인하는 방법에 관한 것이다.

배경 기술

[0002] 프로그램화될 수 있는 뉴클레아제인 ZFNs (zinc finger nucleases), TALENs (transcription-activator-like effector nucleases) 및 Type II CRISPR/Cas 시스템 (박테리아 및 archaea의 적응면역반응) 유래의 RGENs (RNA-guided engineered nucleases)은 고등 진핵세포, 동물 및 식물에서 녹아웃 및 녹인을 시행하는데 널리 사용되고 있다.

[0003] 상기 뉴클레아제는 사용자가 지정한 게놈의 표적 사이트에 DSBs (DNA double-strand breaks)를 유발하여 error-prone NHEJ (non-homologous end joining) 또는 error-free HR (homologous recombination)으로 타겟화된 돌연변이 및 염색체 재배열 (chromosomal rearrangement)을 유발한다.

[0004] 고등 진핵세포의 경우 NHEJ가 HR보다 우세한 DSB 수복 작용 (repair process)으로 작용하며, 온-타겟 및 오프-타겟 사이트에서 삽입 가능한 상동 기증자 (homologous donor) DNA를 필요로 하지 않기 때문에 뉴클레아제를 매개로 하는 유전자 녹아웃에 있어서 HR보다 NHEJ의 사용을 선호한다.

[0005] 상기 DSB 수복은 error-prone NHEJ로 이뤄지며 이는 뉴클레아제 표적 사이트에서 작은 삽입 또는 결실 (indels: insertions + deletions)을 유발하여, 단백질 코딩 서열에 프레임 이동 돌연변이 (frame shift mutation)를 일으킬 수 있다. 그러나, 불가피하게 인-프레임 indels (삽입 또는 결실) 또한 이 과정에서 생성될 수 있으며, 이는 세포 집단에서 뉴클레아제의 효능을 감소시키고 biallelic NULL 클론의 단리를 방해한다. 최근 연구에 의하면 RGENs를 사용할 경우 80% 빈도로 인-프레임 결실을 유발하여 불완전한 유전자 파괴를 야기하는 것으로 나타났다.

[0006] 상기 뉴클레아제를 이용하면 특정 유전자가 작동하지 못하도록 녹아웃 시킬 수 있다. 그러나 이 과정에서 마이크로 상동 염기서열 매개 인-프레임 돌연변이 (microhomology-mediated in-frame mutations)가 생기면 타겟 유전자를 녹아웃시키는 데 방해가 된다.

[0007] 선행 연구에 의하면 뉴클레아제 유발 삽입 또는 결실 (indel: insertions + deletions) 서열을 분석한 결과 TALENs 또는 RGENs는 결실 (deletion)을 삽입 (insertion)보다 훨씬 더 자주 생성시키며 (Kim, Y. et al., Nature methods, 10:185, 2013), 뉴클레아제-유발 결실은 브레이크포인트 교차점 (breakpoint junction) 측면에 위치한 2개의 동일한 짧은 서열 (최소 2 염기)로 구성되는 마이크로 상동과 결부되어 있는 것으로 확인되었다. 또한, 마이크로 상동이 DSB 수복 경로인 MMEJ (microhomologymediated end joining)(도 1a)를 통해 뉴클레

아제 유발 결실을 일으키는 것이 꼬마선충, 지브라피시 및 인간 세포에서 관찰된 바 있다.

[0008] 본 발명자들은, 유전자 가위에 의하여 out-of-frame 돌연변이가 유발될 확률이 높은 표적 서열을 예측할 수 있는 기술을 개발하고자 예의 노력한 결과, 마이크로 상동의 예측을 통하여 뉴클레아제 표적 사이트 선별에 유용한 정보를 제공하는 방법 및 프로그램을 개발하였고, 이를 인간 세포 및 동물 등에서 효과적인 유전자 파괴를 야기하는데 유용하게 이용할 수 있음을 확인하고, 본 발명을 완성하였다.

발명의 내용

해결하려는 과제

- [0009] 본 발명의 하나의 목적은 유전자 녹아웃을 위한 뉴클레아제 표적 서열을 선택하는 방법을 제공하는 것이다.
- [0010] 본 발명의 다른 목적은 뉴클레아제에 의한 Out-of-frame 결실 효율이 높은 서열의 선택을 위한 정보의 제공 방법을 제공하는 것이다.
- [0011] 본 발명의 또 다른 목적은 상기 방법을 수행하는, 컴퓨터 프로그램을 제공하는 것이다.
- [0012] 본 발명의 또 다른 목적은 상기 프로그램이 기록된 컴퓨터 판독 가능한 기록 매체를 제공하는 것이다.

과제의 해결 수단

- [0013] 하나의 양태로서, 본 발명은 유전자 녹아웃을 위한 뉴클레아제 표적 서열을 선택하는 방법을 제공한다.
- [0014] 본 발명에 따른 상기 방법은 마이크로 상동-결부 결실(Microhomology-associated deletion)의 빈도를 사전에 예측(preestimate)할 수 있는 타겟 선정 시스템으로 사용될 수 있으며, 인 실리코(*in silico*) 상에서 뉴클레아제 표적 사이트의 out-of-frame 스코어(score)를 계산하고 채점 시스템(scoring system)을 이용하여 배양세포, 식물 또는 동물에서의 효율적인 유전자 녹아웃이 가능하도록 적절한 표적 사이트를 선정하는데 도움을 줄 수 있다. 또한, 뉴클레아제 표적 서열의 Out-of-frame 결실의 빈도를 예측하기 위하여 사용될 수 있다.
- [0015]
- [0016] 보다 구체적으로, 본 발명은
- [0017] (a) 뉴클레아제 표적 후보 서열을 제공하는 단계;
- [0018] (b) 주어진 뉴클레아제 표적 후보 서열에 존재하는 마이크로 상동 (microhomology)에 대한 정보를 수집하는 단계; 및
- [0019] (c) 상기 (b) 단계에서 수집된 마이크로 상동에 대한 정보를 바탕으로 마이크로 상동과 결부된 out-of-frame 형태의 결실의 빈도를 예측하는 단계를 포함하는, 유전자 녹아웃을 위한 뉴클레아제 표적 서열을 선택하는 방법을 제공한다.
- [0020] 또한, 상기 방법은 추가로 (c) 단계에서 예측된, 주어진 뉴클레아제 표적 후보 서열의 마이크로 상동과 결부된 out-of-frame 형태의 결실의 빈도와 다른 뉴클레아제 표적 후보 서열의 마이크로 상동과 결부된 out-of-frame 형태의 결실의 빈도를 비교하는 단계를 포함할 수 있다. 이를 통하여, 뉴클레아제 표적 후보 서열들 중에서 out-of-frame 형태의 결실의 빈도가 상대적으로 높은 서열이 선택될 수 있다.
- [0021] 상기 마이크로 상동에 대한 정보는 마이크로 상동 서열의 크기, 마이크로 상동 서열 간의 거리 및 마이크로 상동 서열의 서열 정보를 포함할 수 있으나, 이에 제한되지 않는다.
- [0022] 상기 뉴클레아제 표적 후보 서열은 마이크로 상동(Microhomology)에 의해 결실(deletion)이 야기될 수 있는 서열이라면 어떠한 서열도 사용될 수 있다. 구체적으로, 상기 서열은 인간 세포, 지브라피시, 꼬마선충

(*C. elegans*) 등에서 유래될 수 있으며, 또한 그 예로 포유 동물 세포, 곤충 세포, 식물 세포, 어류 세포 등의 서열을 포함할 수 있으나, 이에 제한되지 않는다.

- [0023] 본 발명에서 상기 마이크로 상동 서열은 100% 동일성(identity)를 갖는 2bp 이상의 길이를 갖는 서열 부분을 의미한다. 구체적으로, 절단 부위의 양쪽에 위치한 동일한 2bp 이상의 길이를 갖는 서열 부분일 수 있다. 그 예로, 상기 마이크로 상동 서열은 2bp 이상, 3bp 이상, 4bp 이상, 5bp, 6bp 이상, 7bp 이상, 8bp 이상의 길이를 가질 수 있으나, 이에 제한되지 않는다. 상기 마이크로 상동 서열의 길이는 주어진 뉴클레아제 표적 서열의 길이에 따라 달라질 수 있으며, 2bp 이상의 길이를 가지는 것이 바람직하다. 또한, 상기 뉴클레아제 표적 서열의 5' 또는 3' 말단 부위에서 뉴클레아제 절단 예상 위치까지의 길이 미만인 것이 바람직하다. 뉴클레아제에 의하여 절단되는 위치 양쪽에 마이크로 상동 서열이 존재하는 경우, 마이크로 상동-매개 어닐링(Microhomology-mediated annealing)에 의해 뉴클레아제-유도 결실이 유발될 수 있다 (도 1a).
- [0024] 본 발명에서 상기 뉴클레아제 표적 후보 서열 또는 뉴클레아제 표적 서열은 뉴클레아제 절단 예상 위치를 중심으로 양쪽의 서열 길이가 동일할 수 있으나, 이에 제한되지 않는다.
- [0025] 본 발명에서 상기 표적 서열을 구성하는 염기는 A, T, G, 및 C 로 이루어진 군에서 선택된 것일 수 있으나, 표적 서열을 구성하는 염기라면 그 종류에 특별히 제한되지 않는다.
- [0026] 본 발명에서 상기 뉴클레아제 절단 예상 위치는 뉴클레아제에 의하여 뉴클레오티드 분자의 공유 결합된 백본이 파손될 것으로 예상되는 위치를 말한다.
- [0027] 상기 표적 서열은 유전자 조절부위 또는 유전자 부위에 위치할 수 있으나, 이에 제한되지 않는다. 상기 표적 서열은 유전자의 전사 개시 부위의 10, 5, 3, 또는 1kb 또는 500, 300, 또는 200bp 내에, 예를 들어, 개시 부위의 상류(upstream) 또는 하류(downstream)에 있을 수 있다. 다만, 뉴클레아제에 대한 표적 서열이라면, 그 종류 및 위치는 특별히 제한되지 않는다.
- [0028] 한편, 본 발명에 있어서 상기 유전자 조절부위는 프로모터, 전사 인핸서, 5' 비번역 영역, 3' 비번역 영역, 바이러스 핵키징 서열, 및 선택 마커에서 선택될 수 있으나, 이에 제한되지 않는다. 또한, 본 발명에 있어서, 상기 유전자 부위는 엑손 또는 인트론일 수 있으나, 이에 제한되지 않는다.
- [0029]
- [0030] 본 발명에 있어서, 상기 뉴클레아제는 ZFNs(zinc finger nucleases), TALENs(transcription-activator-like effector nucleases) 및 RGENs(RNA-guided engineered nucleases)로 구성된 군에서 선택될 수 있으나, 이에 제한되지는 않는다.
- [0031] 상기 ZFN은 DNA 절단 도메인 및 징크 핑거 (Zinc finger) DNA 결합 도메인을 포함할 수 있다. 구체적으로, 상기 두 도메인이 융합된 것일 수 있으며, 서로 링커로 연결될 수 있다. 또한, 상기 징크 핑거 DNA 결합 도메인은 목적하는 DNA 서열에 결합하도록 조작된 것일 수 있다.
- [0032] 또한, 상기 TALEN은 DNA 절단 도메인 및 TALE (Transcription activator-like effectors) DNA 결합 도메인을 포함할 수 있다. 구체적으로, 상기 두 도메인이 융합된 것일 수 있으며, 서로 링커로 연결될 수 있다. 상기 TALE는 목적하는 DNA 서열에 결합하도록 조작된 것일 수 있다.
- [0033] 상기 RGEN은 표적 DNA 특이적 가이드 RNA 및 Cas 단백질을 성분으로 포함하는 뉴클레아제를 의미한다. 상기 "가이드 RNA (guide RNA)"는 표적 DNA 특이적인 RNA를 의미하며, Cas 단백질과 결합하여 Cas 단백질을 표적 DNA로 인도한다.
- [0034] 또한, 상기 가이드 RNA는 두 개의 RNA, 즉, crRNA (CRISPR RNA) 및 tracrRNA (trans-activating crRNA)로 구성될 수 있다. 또는 crRNA 및 tracrRNA의 주요 부분의 융합으로 제조된 sgRNA (single-chain RNA)일 수 있다.
- [0035] 또한, 상기 가이드 RNA는 crRNA 및 tracrRNA를 포함하는 dualRNA 일 수 있고, crRNA는 표적 DNA와 결합할 수 있다.
- [0036] 상기 뉴클레아제의 예는 이에 제한되는 것이 아니며, 본 발명의 목적에 비추어, 마이크로 상동 매개 결실을 야기할 수 있는 뉴클레아제라면 제한 없이 포함되는 것이 분명하다.
- [0037] 상기 (c) 단계는 주어진 뉴클레아제 표적 후보 서열에 존재하는 마이크로 상동의 예상 결실 패턴에 대한 스코어

(패턴 스코어)를 산출하는 단계; 및 산출된 패턴 스코어를 바탕으로, (i) 주어진 뉴클레아제 표적 후보 서열에 존재하는 전체 마이크로 상동의 패턴 스코어의 합인 마이크로 상동 스코어, 및 (ii) 이에 대한 out-of-frame 형태의 결실과 결부된 마이크로 상동의 패턴 스코어의 합의 비율 (out-of-frame 스코어)을 산출하는 단계를 포함할 수 있다. 이를 통하여 마이크로 상동과 결부된 out-of-frame 형태의 결실의 빈도를 예측할 수 있으나, 상기 예에 의해 제한되는 것은 아니다.

[0038] 이에 제한되지 않으나, 본 발명의 방법은 보다 구체적으로 하기와 같은 단계를 포함할 수 있다:

[0039] i) 뉴클레아제 표적 후보 서열을 제공하는 단계;

[0040] ii) 주어진 뉴클레아제 표적 서열에서, 뉴클레아제 절단 예상 위치를 기준으로 양쪽 서열 부위에 2bp 이상의 동일한 서열이 존재하는지를 확인하여 마이크로 상동의 존재를 확인하는 단계;

[0041] iii) 상기 표적 서열에 마이크로 상동이 존재하면, 존재하는 마이크로 상동에 대한 정보를 수집하고, 상기 ii) 및 iii) 단계를 1회 이상 반복하는 단계;

[0042] iv) 주어진 뉴클레아제 표적 후보 서열에 존재하는 마이크로 상동의 예상 결실 패턴에 대한 스코어 (패턴 스코어)를 산출하는 단계; 및

[0043] v) 산출된 패턴 스코어를 바탕으로, (i) 주어진 뉴클레아제 표적 후보 서열에 존재하는 전체 마이크로 상동의 패턴 스코어의 합인 마이크로 상동 스코어, 및 (ii) 이에 대한 out-of-frame 형태의 결실과 결부된 마이크로 상동의 패턴 스코어의 합의 비율 (out-of-frame 스코어)을 산출하는 단계.

[0044]

[0045] 상기 ii) 단계는 주어진 뉴클레아제 표적 서열에서 뉴클레아제 절단 예상 위치를 기준으로 양쪽에 동일한 짧은 서열이 존재하는지, 즉 마이크로 상동 서열(microhomologous sequence)의 존재 여부를 확인하는 단계이다.

[0046] 상기 iii) 단계는 상기 표적 서열에 마이크로 상동 서열이 존재하면 마이크로 상동에 대한 정보, 구체적으로 상기 두 개의 마이크로 상동 서열의 각 서열의 5' 쪽에 위치한 시작점 간의 거리 또는 각 서열의 3' 쪽에 위치한 말단 간의 거리인 결실 길이, 및 상기 마이크로 상동 서열의 염기 서열 정보를 얻는 단계이다. 또한 상기 (b) 단계는 모든 마이크로 상동에 대한 정보를 수집하기 위하여 상기 ii) 및 iii) 단계를 1회 이상 반복하는 단계를 포함할 수 있다.

[0047] 구체적으로, MMEJ에 의하여 뉴클레아제-매개 결실이 야기될 경우의 결실 길이(deletion length) 및 마이크로 상동 서열의 서열, 위치 등에 대한 정보를 얻는 단계일 수 있다.

[0048] 상기 iii) 단계를 통하여 주어진 표적 서열에 존재하는 모든 마이크로 상동 패턴이 수집될 수 있다.

[0049] 상기 iv) 단계는 상기 iii) 단계에서 얻어진 정보를 바탕으로 패턴 스코어를 계산하는 단계이다.

[0050] 본 발명의 일 구현예에서는, 마이크로 상동-연관 결실이 마이크로 상동 서열의 크기 및 결실 길이에 의존적임을 확인하였다. 구체적으로, 마이크로 상동 서열의 크기가 증가할수록 결실 빈도가 증가하고, 반면 결실 길이가 증가할수록 결실 빈도가 감소함을 확인하였다. 이에, 상기 확인한 내용을 기초로 주어진 뉴클레아제 표적 서열의 가상적 결실 패턴(Hypothetical deletion pattern)에 대한 스코어 (본 명세서에서 "패턴 스코어"로도 명명됨)에 대한 수학적식을 도출하였다.

[0051] 구체적으로, 패턴 스코어는 하기 수학적 식 1로 계산될 수 있다.

[0052] [수학적 식 1]

[0053] 패턴 스코어 = $S \times \exp(-\Delta / W_{\text{length}})$

- [0054] 여기서,
- [0055] S는 상기 마이크로 상동 서열의 크기 및 상기 마이크로 상동 서열의 염기쌍 (base pairing) 에너지에 비례하는 마이크로 상동 인덱스 (Microhomology index)이고,
- [0056] Δ 는 두 마이크로 상동 서열 간의 5' 위치 사이의 거리 또는 3' 위치 사이의 거리 (결실 길이)이고,
- [0057] W_{length} 는 상기 마이크로 상동 서열 간의 거리에 대한 가중치임;
- [0058]
- [0059] 보다 구체적으로, 상기 S는 마이크로 상동 서열의 길이 및 이를 구성하는 염기 쌍 에너지에 비례하는 지수로서, 그 예로 하기 수학적 식 4로 계산될 수 있다.
- [0060] [수학적 식 4]
- [0061] 마이크로 상동 인덱스 (Microhomology index) = (마이크로 상동 서열의 G 및 C의 수)*2 + (마이크로 상동 서열의 A 및 T의 수)
- [0062] 여기서, G:C 쌍은 A:T 쌍에 비하여 안정한 점을 고려하여, GC 수에 대해서는 +2로, AT 수에 대해서는 +1로 지정하였으나, 이에 제한되는 것은 아니며, AT 수에 비해 GC 수에 가중치를 주는 다양한 방식으로 계산할 수 있다.
- [0063]
- [0064] 또한, 상기 식에서 W_{length} 는 두 개의 마이크로 상동 서열 간의 거리에 대한 가중치로서, 그 예로 20일 수 있다. 그러나, 이에 제한되지 않는다.
- [0065]
- [0066] 또한, 본 발명의 일 구현예에 따르면, 상기 iv) 단계는 마이크로 상동 서열 간의 거리, 즉 결실 길이가 3의 배수인 경우와 3의 배수가 아닌 경우를 나누어 패턴 스코어를 계산하는 단계로 수행될 수 있으나, 이에 제한되지 않는다.
- [0067] 여기서, 마이크로 상동 서열 간의 거리, 즉 결실 길이가 3의 배수인 경우, in-frame 결실이 야기될 것으로 판정될 수 있다. 반면, 결실 길이가 3의 배수가 아닌 경우에는, out-of-frame 결실이 야기될 것으로 판정될 수 있다.
- [0068] 또한, 상기 iv) 단계 수행 이전에 iii) 단계에서 얻은 정보 중 중복되는 정보를 제거하는 단계를 포함할 수 있으나, 이에 제한되는 것은 아니다.
- [0069] 상기 방법에서 v) 단계에는 iv) 단계에서 산출된 패턴 스코어를 바탕으로 마이크로 상동 스코어 (Microhomology score), Out-of-frame 스코어, 또는 둘 다를 계산하는 단계이다.
- [0070] 또한, 보다 구체적으로, 상기 마이크로 상동 스코어 및 Out-of-frame 스코어는 각각 하기 수학적 식 2 및 3으로 계산될 수 있다.
- [0071] [수학적 식 2]
- [0072] 마이크로 상동 스코어 = \sum 패턴 스코어
- [0073] 여기서, 상기 마이크로 상동 스코어는 얻어진 모든 마이크로 상동에 대한 패턴 스코어 값의 합임;
- [0074] [수학적 식 3]

- [0075] $\text{Out-of-frame 스코어} = \sum \text{out-of-frame 결실의 패턴 스코어} / \text{마이크로 상동 스코어} (\sum \text{패턴 스코어})$
- [0076] 여기서, $\sum \text{out-of-frame 결실의 패턴 스코어}$ 는 결실 길이가 3의 배수가 아닌 경우에 해당하는 마이크로 상동 서열의 패턴 스코어 값의 합임.
- [0077]
- [0078] 상기 단계에서 계산된, 마이크로 상동 스코어 및 Out-of-frame 스코어를 바탕으로 뉴클레아제 표적 서열에 대한 마이크로 상동-결부 결실 및 프레임 이동 돌연변이에 대한 빈도를 예측할 수 있다.
- [0079] 본 발명에 따른 방법은 컴퓨터 프로그램으로 구현되어, 유전자 녹아웃 효율이 높은 표적을 간편하게 선정하는데 이용될 수 있다. 본 발명의 방법을 구현할 수 있는 컴퓨터 프로그래밍 언어는 Python, C, C++, 자바(Java), 포트란(Fortran), 비주얼 베이직(Visual Basic) 등이 있으나 이에 제한되지 않는다. 상기 각 프로그램은 CD-ROM(compact disc read only memory), 하드 디스크, 자기 디스켓, 또는 그와 유사한 매체 또는 기구 등의 기록 매체로 저장될 수 있으며, 내부 또는 외부 네트워크 시스템에 연결될 수 있다. 예를 들면, 컴퓨터 시스템은 HTTP, HTTPS, 또는 XML 프로토콜을 이용하여 GenBank(<http://www.ncbi.nlm.nih.gov/nucleotide>)와 같은 서열 데이터베이스에 접속하여 표적 유전자 및 상기 유전자의 조절 영역의 핵산 서열을 검색할 수 있다.
- [0080] 이와 같은 본 발명의 방법을 사용함으로써, 뉴클레아제 표적 서열의 마이크로 상동-관련 결실 빈도를 효과적으로 예측함으로써, 배양된 세포, 식물 및 동물에 이르기까지 녹아웃하고자 하는 표적에 대한 적절한 표적 사이트 선정에 도움을 줄 수 있다. 나아가, 유전자 녹아웃 세포 클론 및 가축과 같은 동물뿐만 아니라, 뉴클레아제 매개 유전자 또는 세포 치료에서도 효율을 현저히 높일 수 있다.
- [0081] 또 하나의 양태로서, 본 발명은 뉴클레아제에 의한 Out-of-frame 결실 효율이 높은 서열의 선정을 위한 정보의 제공 방법을 제공한다.
- [0082] 구체적으로,
- [0083] (a) 뉴클레아제 표적 후보 서열을 제공하는 단계;
- [0084] (b) 주어진 뉴클레아제 표적 후보 서열에 존재하는 마이크로 상동 (microhomology)에 대한 정보를 수집하는 단계; 및
- [0085] (c) 상기 (b) 단계에서 수집된 마이크로 상동에 대한 정보를 바탕으로 마이크로 상동과 결부된 out-of-frame 형태의 결실의 빈도를 예측하는 단계를 포함 하는, 뉴클레아제에 의한 Out-of-frame 결실 효율이 높은 서열의 선정을 위한 정보의 제공 방법을 제공한다.
- [0086] 상기 (a) 내지 (c) 단계 및 각 용어에 대해서는 상기 기술된 바와 같다.
- [0087]
- [0088] 또 하나의 양태로서, 본 발명은 상기 방법의 단계를 수행하는, 컴퓨터 프로그램을 제공한다.
- [0089] 상기 방법, 각 단계 및 컴퓨터 프로그램에 대해서는 앞서 설명한 바와 같다.
- [0090] 또 하나의 양태로서, 본 발명은 상기 프로그램이 기록된 컴퓨터 판독 가능한 기록 매체를 제공한다.
- [0091] 상기 프로그램, 기록 매체 등에 대해서는 앞서 설명한 바와 같다.

발명의 효과

- [0092] 본 발명에 따른 마이크로 상동을 기반으로 유전자가위 기술에 의한 녹아웃 효율이 높은 타겟 선정 방법을 이용

할 경우 인-프레임 돌연변이가 생기는 확률이 낮은 표적 사이트를 선정할 수 있어 보다 용이하게 특정 유전자가 녹아웃된 돌연변이체를 제작할 수 있다. 따라서 상기와 같은 유전자가위 기술을 이용한 유전자 녹아웃 효율을 높이는 방법은 생명과학과 임상연구 분야에 유용하게 이용될 수 있을 것이다.

도면의 간단한 설명

[0093]

도 1a 내지 1e는 마이크로 상동(microhomology)과 결부된 뉴클레아제-유발 결실 패턴을 예측하는 것이다. (도 1a)는 뉴클레아제의 표적 사이트에서의 마이크로 상동을 통한 DNA 결합을 나타낸 것이고, (도 1b)는 마이크로 상동 염기서열-결부 DNA 수복으로 나타나는 인 실리코(*In silico*) 예측 결실 패턴을 나타낸 것이며, (도 1c)는 패턴 스코어를 딥 시퀀싱 실험으로 측정된 결실 패턴의 빈도와 비교한 것을 나타낸 것이다. 또한, (도 1d)는 마이크로 상동 스코어를 실험으로 측정된 마이크로 상동-결부 결실의 빈도와 비교한 것을 나타낸 것이며, (도 1e)는 out-of-frame 스코어를 TALENs 또는 RGENs로 형질감염된 세포에서 관찰된 프레임 이동 결실의 빈도와 비교한 것을 나타낸 것이다.

도 2a 내지 2d는 채점 시스템(scoring system)을 실험적으로 판별하는 것으로, (도 2a)는 BRCA1 유전자의 유력한 표적 사이트의 out-of-frame 스코어의 분포를 나타낸 것이며, (도 2b)는 out-of-frame indels(삽입 또는 결실)의 빈도를 높은 스코어(High)와 낮은 스코어(Low)에서 딥 시퀀싱으로 확인한 것을 나타낸 것이다. 점선은 out-of-frame 스코어의 가우스 분포의 정점의 값과 합치된다. 또한, (도 2c)는 out-of-frame 스코어와 상기 (도 2b)의 out-of-frame indels(삽입 또는 결실) 빈도의 연관성을 나타낸 것이며, (도 2d)는 out-of-frame 스코어와 68 RGENs로 유발된 프레임 이동 indels(삽입 또는 결실)(왼쪽) 또는 결실(오른쪽)의 빈도간의 연관성을 나타낸 것이다.

도 3은 TALENs 또는 RGENs로 유발된 돌연변이를 분석한 것으로, (a)는 HEK293T 세포의 10 TALENs로 유발된 돌연변이 또는 K562 세포의 10 RGENs로 유발된 돌연변이의 평균 빈도, (b)는 TALENs 또는 RGENs로 유발된 결실 및 삽입의 빈도, (c)는 TALENs 또는 RGENs로 유발된 마이크로 상동-결부 결실의 빈도를 나타낸 것이다.

도 4a 내지 4c는 결실 길이의 무게 인자(weight factor)를 평가한 것으로, (도 4a) TALENs 또는 (도 4b) RGENs로 수득한 딥 시퀀싱 데이터를 피팅(fitting)하여 계산된 결실 길이의 무게 인자를 싱글-지수 함수로 나타낸 것이고('선'으로 표시), (도 4c)는 TALENs 또는 RGENs의 평균 무게 인자를 나타낸 것이다.

도 5a 내지 5c는 마이크로 상동과 결부된 가상적 결실 패턴(hypothetical deletion pattern)에 스코어를 할당하는 소스 코드(source code)를 나타낸 것이다.

도 6a 및 6b는 패턴 스코어와 딥 시퀀싱을 사용하여 실험으로 측정된 패턴의 빈도를 비교한 것으로, 화살은 채점 시스템(scoring system)으로 올바르게 예측된 최다 결실 패턴을 나타낸 것이다(피어슨 상관 계수(Pearson correlation coefficient)를 도시함).

도 7은 BRCA1 유전자의 마이크로 상동 스코어의 분포를 나타낸 것으로, 마이크로 상동 스코어를 인간 BRCA1 유전자의 모든 RGENs 표적 사이트에 할당하였고, 마이크로 상동 스코어의 분포는 가우스 공식에 피팅하여 최고값 4026 및 폭 1916인 것을 나타낸 것이다.

도 8은 높은 out-of-frame 스코어 및 낮은 out-of-frame 스코어를 가지는 사이트를 나타낸 것으로, (a)는 MCM6 유전자에서 2개의 RGENs 표적 사이트가 29bp 간격으로 분리된 것을 나타낸 것이고, 2 사이트의 out-of-frame 스코어는 괄호 안에 나타내었다. (b)는 RGENs를 함유하는 플라스미드로 형질감염된 세포의 가장 흔한 결실 패턴을 나타낸 것으로, 마이크로 상동 염기서열은 밑줄로, 2개의 PAM 서열은 하이라이트로 나타낸 것이다.

도 9는 out-of-frame 스코어와 실험적인 데이터를 비교한 것으로, (a)는 TALENs 또는 RGENs를 통해서 생성된 돌연변이를 가지며 생존한 상태로 태어난 81마리 생쥐의 유전자형(genotype)을 분석한 것을 나타낸 것이고, (b)는 out-of-frame 결실의 빈도와 out-of-frame 스코어 간의 상관 관계를 나타낸 것이다 (Pearson correlation coefficient = 0.996).

도 10은, 유전자 녹아웃 효율이 높은 타겟 선정 시스템에 대한 실시예 흐름도를 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0094]

이하, 본 발명을 하기 예에 의해 상세히 설명한다. 다만, 하기 예는 본 발명을 예시하기 위한 것일 뿐, 하기 예

에 의해 본 발명의 범주가 제한되는 것은 아니다.

[0095] 실시예 1: 재료 및 방법

[0096] (1) 세포배양

[0097] K562 (ATCC, CCL-243) 세포는 10% FBS (fetal bovine serum), 100 단위/mL 페니실린 및 100 µg/mL 스트렙토마이신을 함유하는 RPMI1640 배지 (Gibco)에서 배양하였다. 2×10^6 K562세포에 Amaxa SF Cell line 4D-Nucleofector 키트 (Lonza)를 사용하여 20 µg Cas9-인코딩 플라스미드를 형질감염 (transfection)시켰다. 형질감염 24시간 후 상기 1×10^6 K562세포에 추가적으로 *in vitro* 전사된 crRNA 및 tracrRNA를 각각 60mg 및 120mg으로 형질감염시켰다. 상기 2차레 형질감염된 세포에서 48시간 후 게놈 DNA를 추출하였다.

[0098] 또한, HEK293T/17 (ATCC, CRL-11268) 및 HeLa (ATCC, CCL-2)는 10% FBS, 100 단위/mL 페니실린, 100 µg/mL 스트렙토마이신 및 0.1mM NEAA (non-essential amino acids)를 함유하는 DMEM 배지 (Gibco)에서 배양하였다. TALENs를 사용하여 돌연변이를 유발하기 위해서 2×10^5 HEK293T세포에 lipofectamine 2000 (Invitrogen, Carlsbad, CA)을 사용하여 TALENs-인코딩 플라스미드 (500ng)를 형질감염시켰다. 형질감염된 세포에서 72시간 후 게놈 DNA를 추출하였다. 1.6×10^4 HeLa세포에 Lipofectamine 2000 (Invitrogen)을 사용하여 Cas9 인코딩 플라스미드 (0.1 µg)와 sgRNA 발현 플라스미드 (0.1 µg)을 형질감염시켰다. 상기 형질감염된 세포는 72시간 후 세포용해 완충액 (0.005% SDS *Tritirachium album* Proteinase K(1:50; Sigma-Aldrich)으로 처리하여 게놈 DNA를 추출하였다.

[0099]

[0100] (2) TALEN-인코딩 플라스미드의 제조

[0101] TALENs는 표 1 및 2에 나타난 사이트를 표적화하도록 설계되었다. 상기 TALENs 인코딩 플라스미드는 윈스텝 골든 게이트 복제 시스템을 이용하여 제조하였다.

표 1

[0102]

뉴클레아제 (세포 종류)	유전자	이름	표적 사이트 (5'to 3')*	서열번호
TALEN (HEK293T)	APP	APP_1	TAGACCCCGCCACAGCAGC ctctgaagttgg ACAGCAAACCATTGCTTCA	1
	CD4	CD4_1	TGTCTCAGTGGAGCTCCAG gatagtggcacc TGGACATGCACTGTCTTGCA	2
	CREBBP	CREB_1	TGTCCAATGACCTGTCCCAG aagctgtatgcc ACCATGGAGAAGCACAAAGGA	3
	TP53	TP53_1	TACAACACTACATGTGTAACAG ttctctgcatggg CGGCATGAACCGGAGGCCCA	4
	CFTR	CFTR_1	TCGGAAGGCAGCCTATGTGA gatacttcaata GCTCAGCCTTCTTCTCTCA	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	TATGTATGTGAAGGCCATC ccatgttgact ACCTGGTGCCTCTAGTGAAA	9
	NFKB1	NFKB_2	TTGTCATTGCTGTTGTCCCT ctgctacgttcc TATTGTCATTAAGGTATCA	10
RGEN (K562)	C4BPB	C4BP_1	AATGACCACTACATCCTCAAGGG	11

<i>CCR5</i>	CCR5_1	TGACATCAATTATTATACATCGG	12
<i>DROSHA</i>	DROS_1	GATTGCCAATATGCTTCAGTGGG	13
<i>CCR5</i>	CCR5_2	CCTCCGCTCTACTACTGGTGT	14
<i>CCR5</i>	CCR5_3	CCTGCCTCCGCTCTACTACTGG	15
<i>CCR5</i>	CCR5_4	GAATCTAAAACTCTGCTTCGG	16
<i>CCR5</i>	CCR5_5	CCTAAAACTCTGCTTCGGTGC	17
<i>CCR5</i>	CCR5_6	AAATGAGAAGAAGAGGCACAGG	18
<i>AAVS1</i>	AAVS1_1	CTCCCTCCCAGGATCCTCTCTGG	19
<i>EMX1</i>	EMX1	GAGTCCGAGCAGAAGAAGAAGG	20

[0103] *TALEN 사이트는 왼쪽-반 부위 (left-half site, 대문자로 표시), 스페이서(소문자로 표시), 및 오른쪽-반 부위 (대문자로 표시)로 이루어져 있다. PAM 서열은 밑줄로 표시하였다.

표 2

뉴클레아제 (세포 종류)	유전자	이름	표적 사이트 (5' to 3')*	서열번호	
TALEN (HEK293T)	BRCA1	BRCA1_1	TCCAGCTGCTGCTCATACTA <u>ctgatactgctg</u> GGTATAATGCAATGGAAGAA	21	
	BRCA1	BRCA1_h	TCCTGAACATCTAAAAGATG <u>aagtttctatca</u> TCCAAAGTATGGGCTACAGA	22	
	CXCR4	CXCR4_1	TCTTCCTGCCACCATCTAC <u>tccatcatcttc</u> TTAACTGGCATTGTGGGCAA	23	
	CXCR4	CXCR4_h	TGGGTTGATTTTCAGCACCTA <u>cagtgtacagtc</u> TTGTATTAAGTTGTTAATAA	24	
	MCM6	MCM6_1	TTAGAAGTAATTTTAAAGGGC <u>tgaagctgtgga</u> ATCAGCTCAAGCTGGTGACA	25	
	MCM6	MCM6_h	TGGAATCAACTTGTATGAAA <u>ccttgtcaaat</u> GTACTCCACAAGTATGTACA	26	
	PHF8	PHF8_1	TACAGAAGGCCAAAAGAAG <u>aaatatatcaag</u> AAGAAGCCTTGCTGAAGGA	27	
	PHF8	PHF8_h	TACAGCTGCTTGCTCCGCC <u>tataccacagag</u> CACAGCTGGACATTATGGA	28	
	SLC18A2	SLC18_1	TCCAGTCATATCCGATAGGT <u>gaagatgaagaa</u> TCTGAAAGTGACTGAGATGA	29	
	SLC18A2	SLC18_h	TGTATAAAACAGTGTTTCCA <u>gtgacacaactc</u> ATCCAGAAGTGTCTTAGTCA	30	
TP53	TP53	TP53_1	TGTACCACCATCCACTACAA <u>ctacatgtgtaa</u> CAGTTCCTGCATGGGCGGCA	31	
	TP53	TP53_h	TTGTGAGCCACCACGTCCAG <u>ctggaagggtca</u> ACATCTTTTACATTCTGCAA	32	
	RGEN (K562)	APP	APP_1	AGAGGAGGAAGAAGTGGCTG <u>AGG</u>	33
		APP	APP_h	GCCACAGCAGCCTCTGAAGT <u>TGG</u>	34
		BRCA1	BRCA1_1	GCTCATACTACTGATACTGCT <u>G</u>	35
		BRCA1	BRCA1_h	ATTGACAGCTTCAACAGAAAG <u>G</u>	36
MCM6		MCM6_1	GCTAGGGACAGAAGTGT <u>TCTGG</u>	37	
MCM6		MCM6_h	CTCGTGGCCTGGAGCTG <u>GCTGG</u>	38	

[0105] *TALEN 사이트는 왼쪽-반 부위 (left-half site, 대문자로 표시), 스페이서(소문자로 표시), 및 오른쪽-반 부위 (대문자로 표시)로 이루어져 있다. PAM 서열은 밑줄로 표시하였다.

[0106] (3) Cas9-인코딩 플라스미드의 제조

[0107] Cas9-인코딩 플라스미드 및 sgRNA-인코딩 플라스미드를 제조하였다. 상기 Cas9-인코딩 플라스미드의 CMV 프로모

터 제어하에 발현되는 Cas9 융합 단백질은 C-말단에 NLS(Nuclear localization signal) 및 HA 에피토프를 함유하는 펩타이드 태그(NH₃-GGSGPPKRRKRVYPYDVPDYA-COOH, 서열번호 39)를 가진다.

[0108] (4) RNA 제조

[0109] K562 세포에서 사용된 RNA는 T7 RNA 중합효소(MEGAscript T7 키트(Ambion))를 사용하여 런-오프(runoff) 반응으로 인 비트로 (*in vitro*) 전사한 것을 사용하였다. sgRNA 및 crRNA의 주형은 상보적인 2개의 뉴클레오타이드(표 1 또는 표 3)를 어닐링 및 익스텐션하여 제작하였다. 상기 전사된 RNA는 순차적으로 페놀:클로로폼(phenol:chloroform), 클로로폼 및 에탄올 침전으로 추출한 후 분광분석법(spectrometry)으로 정량하였다.

[0110] (5) 타겟화된 딥 시퀀싱 (targeted deep sequencing)

[0111] 뉴클레아제 표적 사이트를 포함하는 게놈 DNA 분절 (segment)을 Phusion 중합효소(New England Biolabs)를 사용하여 증폭시켰다. 같은 양의 PCR 앰플리콘(amplicon)을 사용하여 Illumina MiSeq를 이용한 paired-end read (읽기) 시퀀싱을 수행하였다(총 읽기에서 0.005% 미만의 희귀한 시퀀싱 읽기는 제외시킴). 상기 시퀀싱 결과 중에서 RGENs 절단 사이트(cleavage site; PAM의 3bp upstream) 및 TALENs 표적 사이트(스페이서) 주변에 위치하는 indels(삽입 또는 결실)는 각각 RGENs 및 TALENs로 유발된 돌연변이로 간주하였다.

[0112] 실시예 2: 인간 세포에서 TALENs 및 RGENs에 의해 유발된 돌연변이 서열의 결정

[0113] 인간 세포에서 10개의 TALEN 및 10개의 RGEN에 의하여 유발된 돌연변이 서열을 딥 시퀀싱을 통하여 분석하였다. 그 결과, HEK293T 세포 및 K562 세포에서 각각 19.7±3.6% (평균 ± s.e.m) 및 47.0±5.9%의 빈도로 돌연변이를 유발하는 것으로 나타났다(도 3, 표 1 및 표 3).

[0114] 상기 돌연변이 유형에 있어서 결실 대 삽입을 비교한 결과, 결실이 우세하고 (TALENs: 98.7% 대 1.3%; RGENs: 75.1% 대 24.9%), 삽입은 마이크로 상동 염기서열과 무관하므로, 삽입을 제외하고 결실 데이터를 위주로 분석하였다. 집합체(aggregate)에서의 결실은 TALENs의 경우 44.3% 빈도로, TALENs의 경우 44.3% 빈도로 마이크로 상동 염기서열과 관련되어 있는 것으로 확인되었다(도 3 및 표 3).

[0115] 따라서, TALENs 및 RGENs로 유발된 각각의 43.7%(= 0.987 x 0.443) 및 39.6%(= 0.751 x 0.527) indels(삽입 또는 결실)는 모두 마이크로 상동 염기서열과 관련되어 있는 것으로 나타났다.

[0116] 상기와 같이 지정된 뉴클레아제 표적 사이트에서 마이크로 상동 염기서열-결부 결실을 예측할 수 있다. 극단적인 경우에는, 전부 또는 제로로 단백질을 코딩하는 유전자에 프레임 이동을 일으킬 수 있다. 이와는 대조적으로, 마이크로 상동 염기서열-독립적 indels(삽입 또는 결실)의 3분의 1은 인-프레임 돌연변이를 일으킨다. 상기 indels(삽입 또는 결실)의 ~60%가 평균적으로 마이크로 상동 염기서열-독립이라는 점을 감안한다면, 지정된 사이트에서의 인-프레임 돌연변이가 차지하는 비중은 20%(= 60%/3 + 0%) 내지 60%(= 60%/3 + 40%)인 것을 알 수 있으며 이는 2개의 극단적인 경우에 있어 3배 차이가 있는 것으로 나타났다. 대부분의 진핵 세포가 일배체(haploid)보다는 이배체(diploid)로 구성되어 있기 때문에 null 세포에서 2개의 out-of-frame 돌연변이를 가지는 비중은 표적 사이트에 따라서 16%(= 0.40 x 0.40)에서 64%(= 0.80 x 0.80) 사이의 범위를 가지는 것으로 확인되었다.

표 3

뉴클레오타제 (세포종류)	유전자	이름	서열 리드 횟수	삽입	결실	Out-of-frame 결실의 빈도 (%)	out-of-프레임 삽입 또는 결실(indel)의 빈도 (%)	마이크로 상동 결부 결실의 빈도 (%)	마이크로 상동 스코어 (Microhomology score)*	Out-of-프레임 스코어 ^b
TALEN (HEK293T)	<i>APP</i>	APP_1	58822	148	24260	74.18796373	74.22976073	45.08326	3930	73.61323155
	<i>CD4</i>	CD4_1	130890	221	15863	79.56250394	79.66923651	45.04633	3915	85.84929757
	<i>CREBBP</i>	CREB_1	146455	524	46455	72.3065332	72.41959173	48.77021	4184	48.11185468
	<i>TP53</i>	TP53_1	104451	216	13619	58.7561495	59.02421395	37.33461	2704	44.41568047
	<i>CFTR</i>	CFTR_1	133089	191	11835	57.82847486	58.21553301	40.79425	3171	48.53358562
	<i>CFTR</i>	CFTR_2	122477	90	9239	80.14936681	80.26583771	47.2129	3399	83.81877023
	<i>DROSHA</i>	DROS_1	218200	360	34204	61.34370249	61.23423215	42.91603	4195	46.79380215
	<i>DROSHA</i>	DROS_2	240203	1455	74503	69.29251171	69.37649754	39.50177	3400	81.05882353
	<i>NFKB1</i>	NFKB_1	107680	189	14017	57.95105943	57.90511052	44.29835	4111	43.29846753
	<i>NFKB1</i>	NFKB_2	235082	748	47387	80.92514825	80.69595928	52.7383	3642	93.49258649
RGEN (K562)	<i>CABPB</i>	C4BP_1	47856	21247	11768	38.978586	76.08662729	46.46924	2969	40.9902324
	<i>CCR5</i>	CCR5_1	200645	10727	94967	83.49216043	83.75877533	47.60201	3316	71.26055489
	<i>DROSHA</i>	DROS_1	251509	15723	106834	56.85549544	60.24217303	40.52596	4530	46.55629139
	<i>CCR5</i>	CCR5_2	76347	1723	26406	74.16496251	75.49148566	47.13929	3772	65.16436904
	<i>CCR5</i>	CCR5_3	73367	2511	10001	62.34376562	69.46131714	55.49345	5118	57.44431419
	<i>CCR5</i>	CCR5_4	69780	1325	17745	65.08312201	67.29417934	59.77289	4148	68.63548698
	<i>CCR5</i>	CCR5_5	99571	3256	29392	80.3041644	82.11529037	62.9491	4569	76.01225651
	<i>CCR5</i>	CCR5_6	106450	22712	25837	68.4754422	83.03363612	44.9402	3660	60.51912568
	<i>AAVS1</i>	AAVS1_1	43249	7812	18964	86.24762708	93.29997012	57.83959	5894	72.34476
	<i>EMX1</i>	EMX1	52945	16745	22858	47.30072622	69.47453476	64.47283	4756	50.75694

[0118] 상기 indel 서열 분석에 따르면 마이크로 상동-결부 결실의 빈도는 마이크로 상동의 크기 및 결실의 길이에 따라 달라지는 것으로 나타났다. 따라서, 마이크로 상동 서열의 크기가 증가할수록 결실 빈도가 증가하는 것으로 나타났다. 또한, 도 4에 나타난 바와 같이, 결실의 길이가 증가할수록, 결실 빈도는 기하급수적으로 감소하는 것으로 나타났다. 아울러, 도 1b에 나타난 바와 같이, 인간의 APP 유전자에 특이적인 TALENs 쌍에 의해 유발된 2가지 가장 흔한 결실은 각각 20bp 및 17bp로 분리된 5- 및 4-뉴클레오타이드 서열로, 표적 사이트에 근접하게 위치하는 것으로 확인되었다.

[0119] 실시예 3: 마이크로 상동-결부 결손의 예측 방법식

- [0120] 상기 결과를 토대로, 마이크로 상동 염기서열-결부 결실을 예측하기 위해 간단한 방정식(수학식)을 개발하여, 인 실리코(*in silico*) 상에서 마이크로 상동 염기서열과 최소한 2bp 결부된 뉴클레아제 표적 사이트의 결실 패턴을 예측하였고, 각 가상적 결실 패턴(hypothetical deletion pattern)에 대해서 스코어를 부여하였다.
- [0121] 상기 스코어는, Python 언어로 작성된 컴퓨터 프로그램(도 5a 내지 5c)을 이용하여, 도 1b에 나타난 바와 같이, 다음 수학식 5로 계산하였다.
- [0122] [수학식 5]
- [0123] 패턴 스코어 = $S \times \exp(-\Delta / 20)$,
- [0124] 여기서, S는 마이크로 상동 서열의 크기 및 염기 쌍 에너지에 대응되는 마이크로 상동 인덱스이고,
- [0125] Δ 는 염기 쌍(base pair, bp)의 결실 길이임.
- [0126] 상기 수학식 1에서의 G:C 쌍은 A:T 쌍 보다 더 안정하기 때문에, 마이크로 상동 서열에서 임의로 각 A:T 쌍에 +1을 부여하였고, 각 G:C 쌍에는 +2를 부여한 후 계산하여 마이크로 상동 염기서열 인덱스를 취득하였다 (Fonseca Guerra, C. *et al.*, Journal of the American Chemical Society, 122:4117-4128, 2000). 도 1c에서 나타난 바와 같이, 상기 수학식 1을 사용한 결과, 정확하게 TALENs 사이트에서 3가지 가장 빈번한 결실 패턴을 예측하였다. 상기 수학식 1을 포함하는 프로그램이 기록된 매체를 이용하여 기타 19개의 표적 사이트에 대한 패턴 스코어를 계산한 다음 딥 시퀀싱 데이터와 비교한 결과, 도 6a 및 6b에서 나타난 바와 같이, 상기 프로그램이 기록된 매체는 정확하게 5 TALENs 사이트와 8 RGENs 사이트에서 가장 빈번한 결실 패턴을 예측하였으며, 전반적으로, 상기 패턴 스코어는 딥 시퀀싱 데이터와 합치하는 것으로 나타났다. 또한, 상기 결과에서 피어슨 상관 계수(Pearson correlation coefficient)는 20 사이트에서 0.411~0.945 범위를 가지며 평균값은 0.727인 것으로 나타났다.
- [0127] **실시예 4: 채점 시스템의 유용성 검증**
- [0128] 마이크로 상동 매개 결실(microhomology-mediated deletions) 또는 out-of-frame 돌연변이를 형성하는 경향이 강한 뉴클레아제 표적 사이트를 선정하기 위하여 각각의 표적 사이트에 2 가지 스코어를 할당하였다.
- [0129] 첫 번째 스코어는 마이크로 상동 스코어로 표적 사이트에서의 가상적 결실 패턴(hypothetical deletion pattern)에 할당된 모든 스코어의 합계(\sum 패턴 스코어(pattern score))로 계산하였고, 두 번째 스코어는 out-of-frame 스코어로 다음 수학식 3를 사용하여 계산하였다.
- [0130] [수학식 3]
- [0131] Out-of-frame 스코어 = \sum Out-of-frame 결실의 패턴 스코어 / \sum 패턴 스코어
- [0132] 이때, 표적 사이트들 간의 간격은 +30bp였다. 그 결과, 마이크로 상동 스코어 및 out-of-frame 스코어는 각각의 마이크로 상동 염기서열-결부 결실 및 프레임 이동 돌연변이의 빈도를 예측하는 데 통계적으로 유의함을 확인할 수 있었다(돌연변이 각각의 피어슨 계수 = 0.635 및 0.797)(도 1d 및 도 1e). 상기 결과로 채점 시스템(scoring system)을 사용하여 표적 유전자의 적절한 조작(과표)용 사이트를 선정할 수 있음을 알 수 있었다.
- [0133] 본 발명에 따른 채점 시스템의 유용성을 검증하기 위하여, 9개의 유전자를 대상으로 높은 스코어와 낮은 스코어를 가지는 2개의 타겟 사이트를 각각의 유전자로부터 선택하였다. 우선, 인간 BRCA1 유전자(엑손과 인트론에 있는 9,494 사이트)의 모든 RGENs 타겟 사이트(5'-X₂₀NGG-3', X₂₀은 crRNA 또는 sgRNA 서열에 해당되며 NGG는 PAM(protospacer-adjacent motif)으로 Cas9에 의해 인식됨)를 찾은 다음 각 타겟 사이트에 마이크로 상동 스코어 및 out-of-frame 스코어를 할당하였다. 도 2a에 나타난 바와 같이, 상기 out-of-frame 스코어는 65.9에서

정점을 갖는 가우스 함수에 따라 분포되었는데, 그 이유는 3분의 2의 모든 마이크로 상동 염기서열-결부 결실은 프레임 이동 돌연변이로 귀착되기 때문이다. 상기 BRCA1의 엑손에서 임의로 2개의 타겟 사이트를 각각 하나씩 스코어 상위 20% 및 하위 20%에서 선택하였다. 동일하게, 다른 8개의 유전자에 대해서도 높은 스코어 및 낮은 스코어를 가지는 사이트를 선택하였다. 따라서, RGENs 또는 TALENs로 각각 6개 또는 12개의 사이트가 타겟화(targeted)되었다(표 2). 그 다음, 상기 뉴클라에제를 인코딩하는 플라스미드로 인간 세포를 형질감염하여 돌연변이를 유도한 다음, 표적 사이트를 포함하는 부위를 증폭시켰다. 그 다음, PCR 앰플리콘을 딥 시퀀싱하여 각 표적 사이트에 out-of-frame indel의 비율을 확인하였다(표 4).

표 4

뉴클레아제 (세포 종류)	유전자	이름	서열 리드 횟수	삽입	결실	Out-of-frame 결실의 빈도 (%)	out-of-프레임 삽입 또는 결실(indel)의 빈도 (%)	마이크로 상동 스코어 (Microhomology score)*	Out-of-프레임 스코어 ^b
TALEN (HEK293T)	BRCA1	BRCA1_l	77583	795	32519	39.10479085	39.61392158	4303	21.77551
	BRCA1	BRCA1_h	122533	871	62077	81.10301121	81.08088489	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	48.1871345	48.46955659	3236	31.33498
	TP53	TP53_h	176325	1210	28962	79.16705144	78.8308357	3769	85.35421
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	88.1639982	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

*Microhomology score = \sum pattern score.

^bOut-of-frame score = $\frac{\sum \text{pattern score of an out-of-frame deletion}}{\sum \text{pattern score}}$ ($\pm 30\text{bp}$ between target sites)

[0134]

[0135]

[0136]

그 결과, 도 2b에 나타난 바와 같이, 9개의 유전자 쌍에 있어서 높은 스코어 사이트는 낮은 스코어 사이트보다

out-of-frame indels(삽입 또는 결실)를 더 자주 생성하였고, 모든 9개의 유전자에서의 높은 스코어 사이트는 frame-shifting indels(삽입 또는 결실)를 예측 스코어의 평균값인 66% 이상의 빈도로 생성하는 것으로 확인되었다. 이와는 대조적으로, 상기 모든 9개의 유전자에서의 낮은 스코어 사이트는 평균보다 훨씬 낮은 빈도로 out-of-frame 돌연변이를 생성하는 것으로 나타났다. 도8에서 나타난 바와 같이, 2개의 RGENs는 MCM6 유전자에서 out-of-frame indels(삽입 또는 결실)를 36.2% 또는 74.8%의 빈도로 각각 낮은 스코어 또는 높은 스코어 사이트(각 사이트는 29bp로 분리)에서 유발하였으므로 타겟 사이트의 선택이 돌연변이 생성에 중요한 것으로 나타났다.

[0137] 상기 실험에 있어서, 높은 스코어 사이트와 낮은 스코어 사이트는 각각 79.3% 및 42.5%의 빈도를 가지는 frame-shifting indels(삽입 또는 결실)를 평균적으로 생성하였다(스튜던트 t-검정, $p < 0.001$). 또한, 이배체(diploid) 세포 및 유기체(organism)에서 null 클론을 얻을 수 있는 확률은 각각 62.8%(0.793×0.793) 및 18.1%($= 0.425 \times 0.425$)로, 상기 2개의 극단적인 경우의 예상치인 64% 및 16%와 매우 유사하였다. 따라서, 도 2c에서 나타난 바와 같이, out-of-frame 스코어는 frame-shifting indels(삽입 또는 결실)의 빈도를 예측하는데 적합하였다(피어슨 계수 = 0.934).

[0138] 채점 시스템의 유용성을 추가로 확인하기 위해서 여러 종류의 유전자를 타겟으로 68개 RGENs를 이용하여 실시예 1의 방법으로 형질감염된, RGENs를 함유하는 HeLa 세포에서 검증하였다(표 5: 68 RGENs로 유발된 HeLa 세포의 돌연변이 분석).

표 5

[0139]

유전자	표적 사이트 (5' to 3')	서열 리드 횟수	삽입	결실	Out-of-frame 결실의 빈도 (%)	out-of-프레임 삽입 또는 결실 (indel)의 빈도 (%)	마이크로 상동 스코어 (Microhomology score) ^a	Out-of-프레임 스코어 ^b
ABL1	TGGGGCTGGATAATGGAGC GTGG (SEQ ID NO: 40)	3777	630	849	89.8704	93.712	5895	67.6844 7837
ACK	CGGTCCAACAACGATCCCA GAGG (SEQ ID NO: 41)	2374	306	1112	74.1007	79.2666	4429	61.2102 0546
ALK	CTGTGACCACGGGACGGTG CTGG (SEQ ID NO: 42)	4753	905	2248	66.1922	74.3102	5617	66.2275 2359
ARG	TCCATCTCGCTCAGGTACG AGGG (SEQ ID NO: 43)	4316	985	2188	80.8044	86.0384	4220	69.4312 7962
AXL	GTCCCGTGTCCGAAAGCTG CAGG (SEQ ID NO: 44)	3514	494	1870	61.6043	68.5702	4729	55.2548 1074
BLK	ACTACCCGCTATGAATGA TCGG (SEQ ID NO: 45)	4121	1286	1280	81.4844	90.0624	4684	56.8531 1699
BRK	CCCAGAGGCCACATACTT GGGG (SEQ ID NO: 46)	3380	913	1229	55.9805	74.2297	5984	61.1631 016
CK4	ACATGCCGCTATTGAGCC ACGG (SEQ ID NO: 47)	3946	133	794	55.9194	60.1942	4259	62.1507 3961
CSK	CTGACCGACCCCTAGACCG CAGG (SEQ ID NO: 48)	4102	1053	1715	82.7405	88.7283	5058	64.8477 6592
CTK	GCGGAAACACGGGACCAAG TCGG (SEQ ID NO: 49)	4469	376	1571	78.9306	81.1505	6340	69.9526 8139

DD R2	CCCCAGTGCTCGGTTTGTC ACGG (SEQ ID NO: 50)	6486	1082	3531	84.3104	87.5569	5379	63.3203 1976
EG FR	CAAAGCTGTATTTGCCCTC GGGG (SEQ ID NO: 51)	4302	194	688	67.0058	73.2426	3892	57.3484 0699
Ep hA 1	GCTCCAATTGGATCTACCG CGGG (SEQ ID NO: 52)	3762	317	2322	70.801	73.7779	4049	67.6463 3243
Ep hA 10	TGGACCGCGCAGGTCTCC ATGG (SEQ ID NO: 53)	3575	754	774	71.3178	85.0785	5892	64.6978 9545
Ep hA 2	AGGCTCCGAGTAGCGACA CTGG (SEQ ID NO: 54)	3700	696	727	77.7166	88.2642	5328	73.4046 5465
Ep hA 3	TTGTCGACCAGGTTTCTAC AAGG (SEQ ID NO: 55)	2132	608	636	87.1069	92.0418	3497	69.4881 3269
Ep hA 4	AACACCGAGATCCGGGATG TAGG (SEQ ID NO: 56)	5136	287	2520	85.2381	85.1087	4003	68.9982 5131
Ep hA 5	ACTGCAGCGCGAAGGGGA GTGG (SEQ ID NO: 57)	4830	109	1800	67.2778	67.7842	6062	62.2731 7717
Ep hA 6	TCTCTCAATACGAATCTT GAGG (SEQ ID NO: 58)	3660	344	1357	52.5424	59.3768	4342	63.7954 8595
Ep hA 7	CACCTGGTATGTTCTGATC GGGG (SEQ ID NO: 59)	6125	1850	2738	89.2988	92.6548	4648	74.4406 1962
Ep hB 1	CACATGCATCCCCAACGCA GAGG (SEQ ID NO: 60)	3688	361	2105	71.6865	74.2092	4395	61.5927 19
Ep hB 2	GGCTACGGACCAAGTTTAT CCGG (SEQ ID NO: 61)	3553	49	537	68.9013	70.9898	3974	59.3356 8193
Ep hB 4	GCAGAATATTCGGACAAAC ACGG (SEQ ID NO: 62)	4113	1337	1722	90.0697	93.9523	4455	77.0819 3042
Ep hB 6	CTTACCCCTTTACTACCGT CAGG (SEQ ID NO: 63)	4867	472	2010	89.7512	90.5318	4798	67.2780 3251
FE R	AGACTGGGAATTACGGTTA CTGG (SEQ ID NO: 64)	4619	172	2246	67.4978	67.9487	4468	61.0116 3832
FE S	GGAGGCCGAGCTTCGTCTA CTGG (SEQ ID NO: 65)	3287	75	756	32.8042	38.7485	4584	48.5820 2443
FG FR 1	CTCTGCATGGTTGACCGTT CTGG (SEQ ID NO: 66)	4070	210	1386	83.4776	83.7719	4649	67.8425 4678
FG FR 3	CGGCAACTACCTGCGTC GTGG (SEQ ID NO: 67)	2250	299	1171	65.585	70.9524	4392	48.1329 6903
FG FR 4	AACTCCCATAGTGGGTCGA GAGG (SEQ ID NO: 68)	6126	204	659	62.3672	70.2202	4744	57.2512 6476
FG R	GCAGCTGTACGCCGTGGTG TCGG (SEQ ID NO: 69)	4216	175	1686	45.255	49.2746	5234	36.3584 2568

FM S	ATCTACTTGATCGAGGTTG AGGG (SEQ ID NO: 70)	6805	467	2273	53.5416	60.9489	4919	48.3431 5918
FR K	CTGGTCAGTTTGGCGAAGT ATGG (SEQ ID NO: 71)	4682	537	699	81.9742	89.4013	4712	72.2410 8659
FY N	GGGACCTTGCCTACGAGAG GAGG (SEQ ID NO: 72)	4055	130	1897	66.5788	67.8836	4443	66.9367 5445
HC K	TGTCGCCCGCTTACTCT CTGG (SEQ ID NO: 73)	4822	200	420	86.6667	89.5161	3736	72.8854 3897
HE R2 /E rb B2	AGCTGGCGCCGAATGTATA CCGG (SEQ ID NO: 74)	4921	121	1935	76.1757	77.0914	5021	69.9462 2585
IG F1 R	TCAGTACGCCGTTTACGTC AAGG (SEQ ID NO: 75)	4857	1117	2543	65.0806	74.7268	3991	55.1490 8544
IN SR	GAGAATTGCTCTGTCATCG AAGG (SEQ ID NO: 76)	5838	924	920	84.8913	91.5944	4280	67.5233 6449
IT K	AAGCGGACTTTAAAGTTCG AGGG (SEQ ID NO: 77)	5075	125	472	80.5085	84.0871	4851	78.5198 9281
JA K2	AGCAACAGAGCCTATCGGC ATGG (SEQ ID NO: 78)	4060	254	1473	67.2098	70.3532	4379	66.3165 1062
JA K3	CTGGAAAGTCGAGAAGGG CTGG (SEQ ID NO: 79)	3349	102	574	86.2369	86.9822	4551	74.2913 6454
KD R	TCCAGGTTTCCTGTGATCG TGGG (SEQ ID NO: 80)	5604	988	1684	61.1045	75	3825	63.3464 0523
KI T	TATTCTCATTGTTTCATC CAGG (SEQ ID NO: 81)	5426	428	1633	55.2358	61.8147	5110	56.5362 0352
LC K	GAGCCTTCGTAGTAACCA GTGG (SEQ ID NO: 82)	3159	141	680	82.9412	83.8002	4884	73.4234 2342
LM R1	GCCACCCGTCGACGTCCC TGGG (SEQ ID NO: 83)	3363	236	1810	78.5083	80.2053	8541	61.9716 6608
LM R2	GCTCAGGAGCGTTGAACCT GAGG (SEQ ID NO: 84)	4756	1648	1807	68.9541	83.3864	4369	58.4115 3582
LT K	TGGCTCAAGATACTAGGC GGGG (SEQ ID NO: 85)	4131	172	1195	82.3431	80.9802	5454	65.5298 8632
ME R	CTATCCCGGGACCTTTTC CAGG (SEQ ID NO: 86)	2890	135	1320	81.3636	82.6804	5269	58.9485 6709
MU SK	GCATAGCTACCAATAAGCA TGGG (SEQ ID NO: 87)	4871	154	2709	65.2639	66.2592	4309	54.4209 7935
PD GF Ra	CAGCCTAAGACCAGGAACG CCGG (SEQ ID NO: 88)	4452	353	2708	84.8227	85.7563	5043	71.3067 6185

PD GF Rb	AGGGAACGTAGTTATCGTA AGGG (SEQ ID NO: 89)	3996	149	2407	55.7541	57.903	4091	53.9965 7785
PY K2	GGTCCTGAATCGTATTCTT GGGG (SEQ ID NO: 90)	4180	695	1995	77.594	82.3792	3720	57.3118 2796
RE T	TGCTGGGTGATGCGGCCGG TGGG (SEQ ID NO: 91)	3179	305	1027	69.2308	75.0751	5776	63.7811 6343
RO N	GTCATCGGGCCGTTATGG TGGG (SEQ ID NO: 92)	3350	1133	1326	78.9593	88.2066	6432	62.1890 5473
RO R1	GCCATAGATGGTGGACCGA AAGG (SEQ ID NO: 93)	5172	571	2748	82.2416	84.9654	6204	57.6241 1348
RO S	TGAGGTGCACTAATAGAGG GTGG (SEQ ID NO: 94)	4098	503	1663	44.979	56.5559	3834	53.5732 916
RY K	TATTGCCTTACATGAATTG GGGG (SEQ ID NO: 95)	6079	753	2584	67.8406	74.1984	4018	67.8695 8686
SR C	GTCTGACTTCGACAACGCC AAGG (SEQ ID NO: 96)	4141	232	1700	35.0588	41.2526	4157	44.8400 2887
SR M	CCACACTCCGAATTCGCC TTGG (SEQ ID NO: 97)	1423	73	722	75.2078	77.1069	4392	73.9754 0984
SY K	GGTGATGTTGCCGAAAAAG AAGG (SEQ ID NO: 98)	3825	368	1474	57.9376	65.5809	4424	51.3788 4268
TI E1	CGCCTGTGGGACGGGACAC GGGG (SEQ ID NO: 99)	2050	437	657	64.5358	77.5137	9164	63.7494 5439
TI E2	CAGAGTTCATATTCTGTCC GAGG (SEQ ID NO: 100)	5063	1238	2267	68.8134	78.9444	4027	60.4420 1639
TN K1	GCAGTAGGTTGCGCGTAGC GAGG (SEQ ID NO: 101)	3497	1307	725	69.931	89.2224	7094	65.2100 3665
TR KB	GCCGTGGTACTCCGTGTGA TTGG (SEQ ID NO: 102)	4525	1080	1973	62.3923	74.8772	3748	68.7299 8933
TR KC	CATCAGCGTTGATGCAGTA GAGG (SEQ ID NO: 103)	5151	83	876	48.0594	50.9906	5474	54.7497 2598
TX K	GTTGTTTACCAGCCACAGC TGGG (SEQ ID NO: 104)	5371	1954	1682	66.4685	83.8284	4931	66.9843 8451
TY K2	GAACCGGCTGTGTACCGTT GTGG (SEQ ID NO: 105)	4569	87	466	86.0515	86.9801	5638	75.8957 077
TY RO 3	GGCCACTAGCGTTGCTG CTGG (SEQ ID NO: 106)	4466	345	2254	60.9583	65.0635	4665	58.1779 2069
YE S	TCAGGTCTGTATTAATGG CTGG (SEQ ID NO: 107)	5584	1157	1364	80.9384	88.8933	4727	62.8305 4792

^aMicrohomology score = \sum pattern score.

^bOut-of-frame score = $\frac{\sum \text{pattern score of an out-of-frame deletion}}{\sum \text{pattern score}}$. ($\pm 35\text{bp}$ between target sites)

[0140]

[0141]

도 2d에 나타난 바와 같이, out-of-frame 스코어는 frame-shifting indels(삽입 또는 결실) 또는 결실의 빈도와 잘 합치하였다(피어슨 계수 = 0.717 또는 0.732). 상기 out-of-frame indels(삽입 또는 결실)의 빈도 범위는 38.7%~94.0%이었다. 이배체 인간 세포에서 null 클론을 얻을 수 있는 확률은 15.0%(= 0.387 x 0.387)~88.4%로 극단적인 경우에는 5.9배 차이가 나는 것으로 나타났다. HeLa 세포를 포함한 대부분의 암세포주는 다배체(multi-ploid > 3n)이므로 더 높은 스코어 사이트를 선택하는 것이 중요하다. 상기 채점 시스템은 TALENs에서 훨씬 나은 결과를 얻을 수 있었다. 그 이유는 RGENs보다 TALENs가 마이크로 상동 염기서열 독립 삽입(microhomology-independent insertions)을 더 낮은 빈도로 유발하기 때문이다.

[0142]

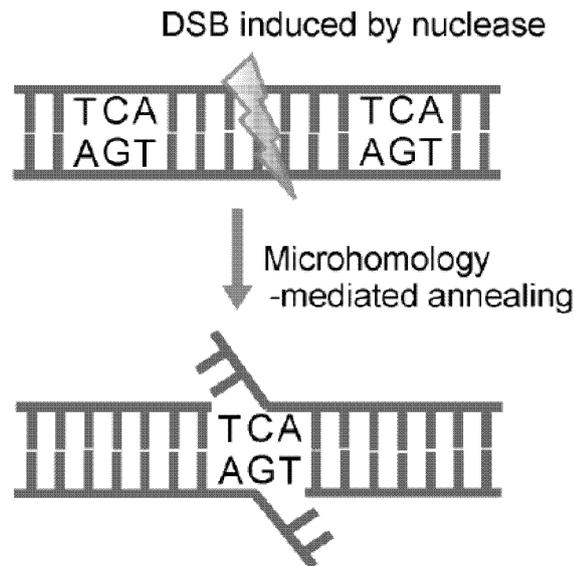
TALENs 또는 RGENs를 통해 제조된 돌연변이 81 마리의 생쥐(Sung, Y.H. et al., Genome research, 24:125-131, 2014; Sung, Y.H. et al., Nature biotechnology, 31:23-24, 2013)를 대상으로 유전자형(genotype)을 분석한 결과 out-of-frame 결실은 out-of-frame 스코어(피어슨 계수 = 0.996)와 잘 합치되는 것으로 나타났다(도 9a 및 b). (Pearson correlation coefficient=0.996))

[0143]

이상의 설명으로부터, 본 발명이 속하는 기술분야의 당업자는 본 발명이 그 기술적 사상이나 필수적 특징을 변경하지 않고서 다른 구체적인 형태로 실시될 수 있다는 것을 이해할 수 있을 것이다. 이와 관련하여, 이상에서 기술한 실시 예들은 모든 면에서 예시적인 것이며 한정적인 것이 아닌 것으로서 이해해야만 한다. 본 발명의 범위는 상기 상세한 설명보다는 후술하는 특허 청구범위의 의미 및 범위 그리고 그 등가 개념으로부터 도출되는 모든 변경 또는 변형된 형태가 본 발명의 범위에 포함되는 것으로 해석되어야 한다.

도면

도면1a

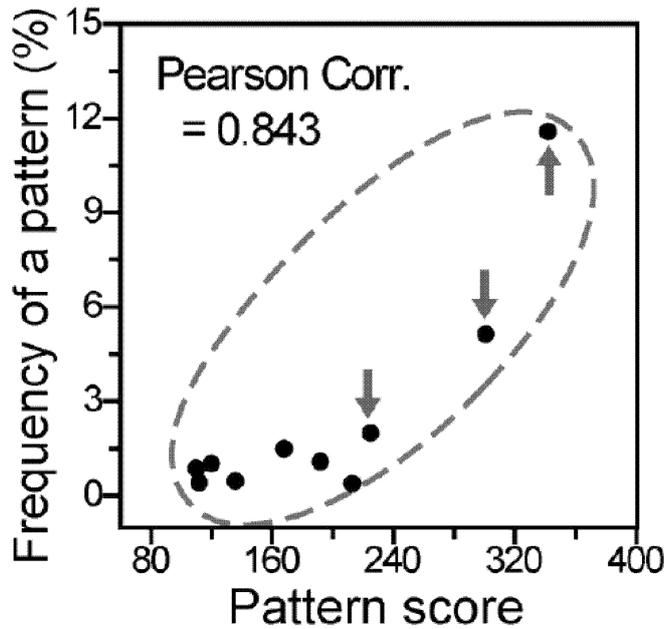


도면1b

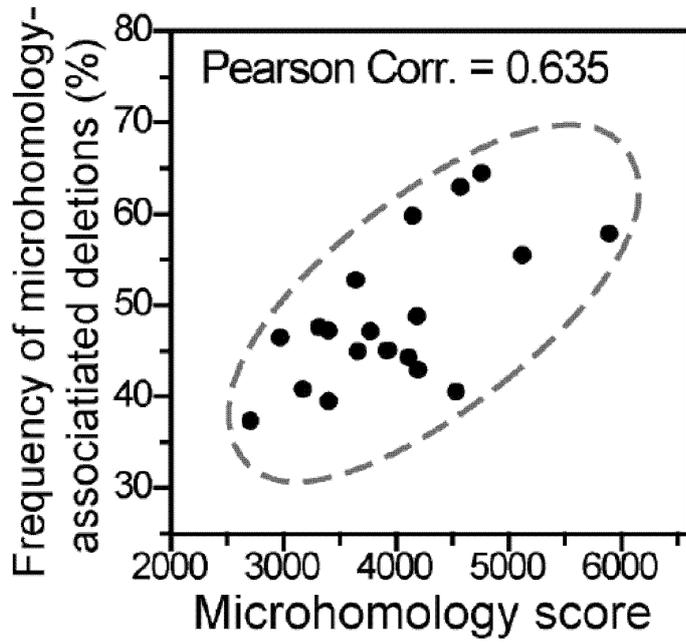
	Δ Deletion length	↓ Microhomology index (S)	↓ Pattern score	In-frame
CTAGACCCCCGCCACAGCAGCCTCTGAAGTTGGACAGCAAACCATTGCTTCAC	-	-	-	-
CTAGACCCCCGCCACAGC-----AAAACCATTGCTTCAC	20	8	342	no
CTAGACCCCCGCCACAGCAGC-----AAAACCATTGCTTCAC	17	7	301	no
CTAGACCCCCGCCACAGCAGCCTG-----GACAGCAAACCATTGCTTCAC	6	3	225	yes
CTAGACCCCCGCCACAGCAGCCTGA-----CAGCAAACCATTGCTTCAC	7	3	213	no
CTAGACCCCCGCCACAGCAG-----TTGGACAGCAAACCATTGCTTCAC	9	3	192	yes
CTAGACCCCCGCCACAG-----TTGGACAGCAAACCATTGCTTCAC	12	3	168	yes
CTAGACCCCCGCCACAGCAGC-----ATTGCTTCAC	22	4	136	no
CTAGAC-----AGCAAACCATTGCTTCAC	29	5	120	no
CTAGACCCCCGC-----AAAACCATTGCTTCAC	26	4	112	no
CTAGACCCCCGCA-----TTGCTTCAC	31	5	110	no

$\exp\left(-\frac{\Delta}{20}\right) \times S = \text{Pattern score}$

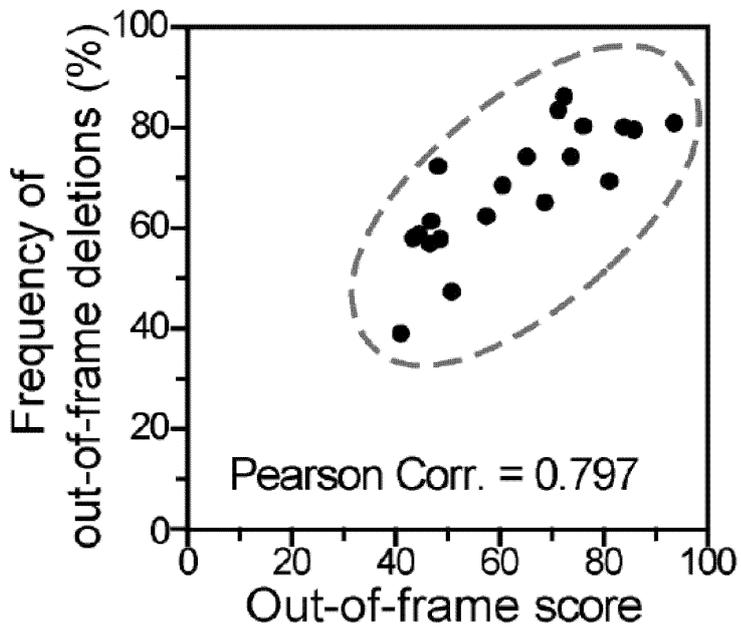
도면1c



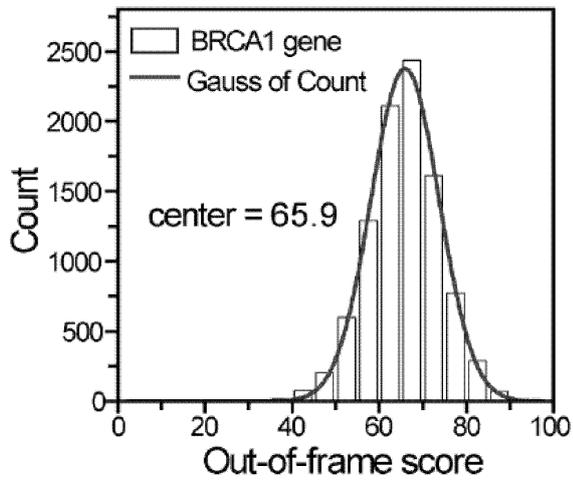
도면1d



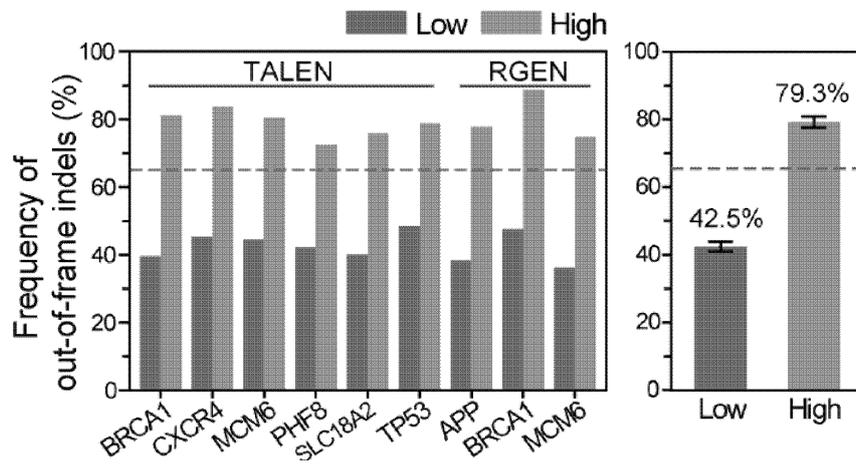
도면1e



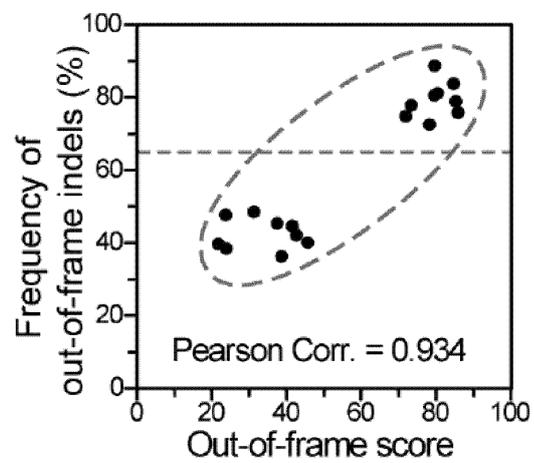
도면2a



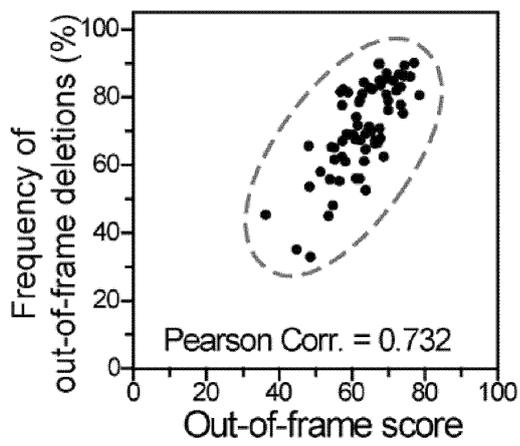
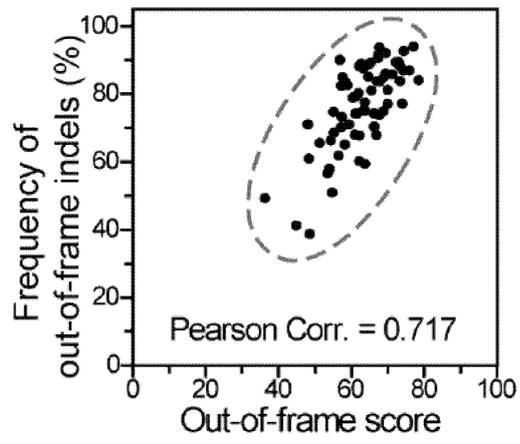
도면2b



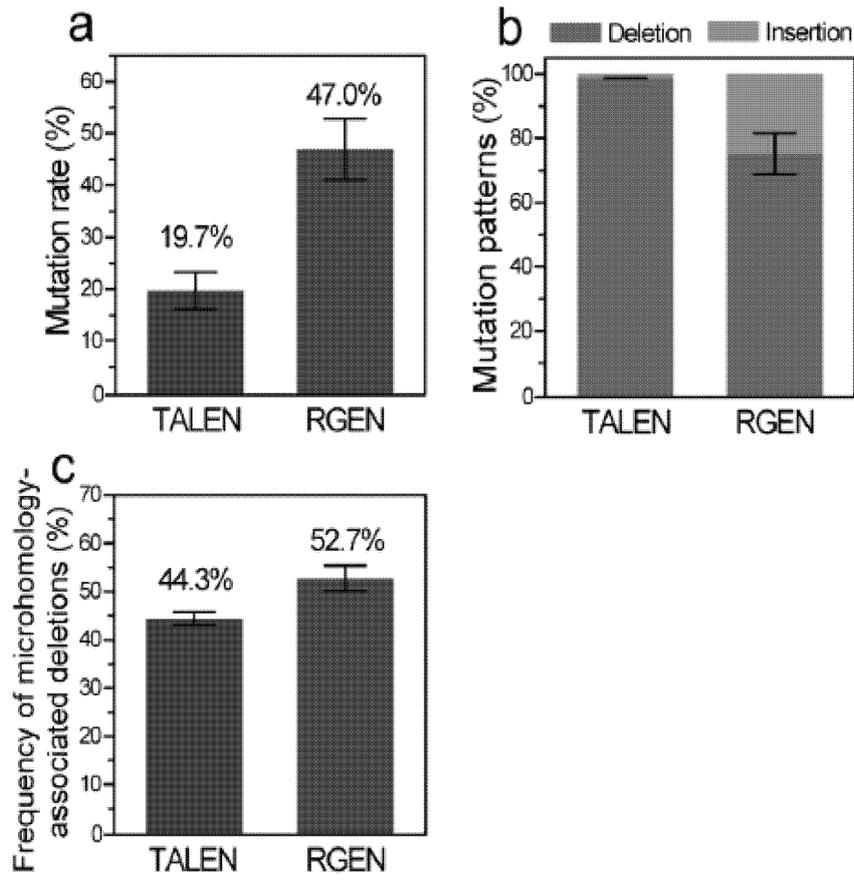
도면2c



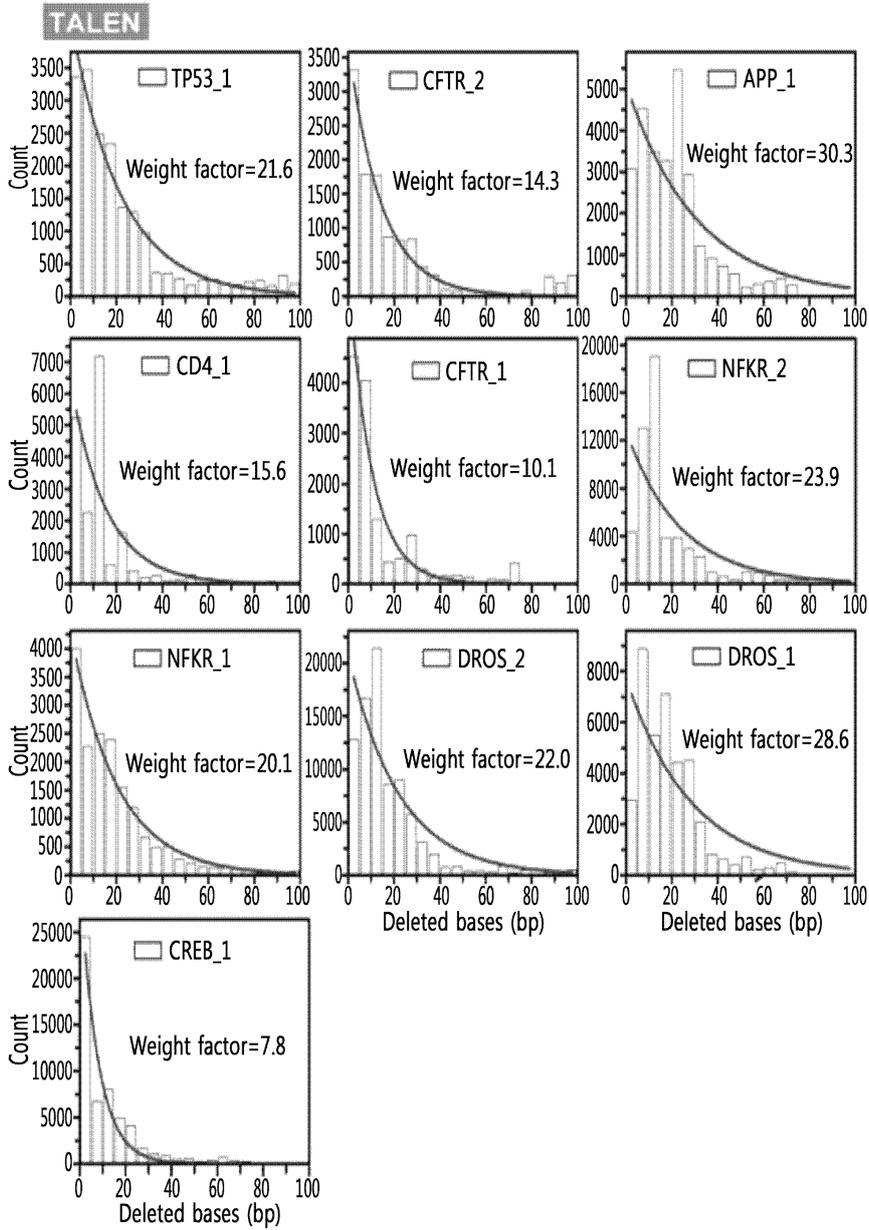
도면2d



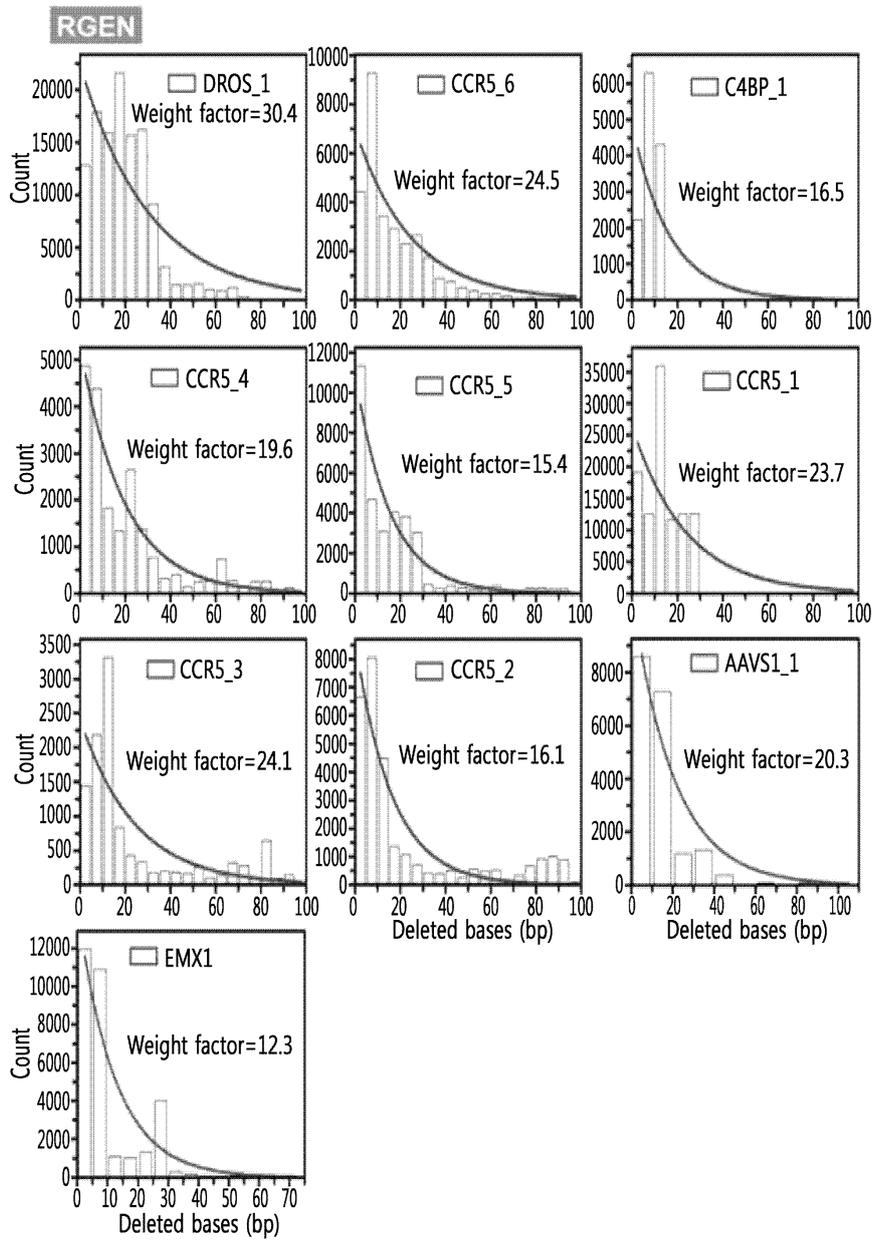
도면3



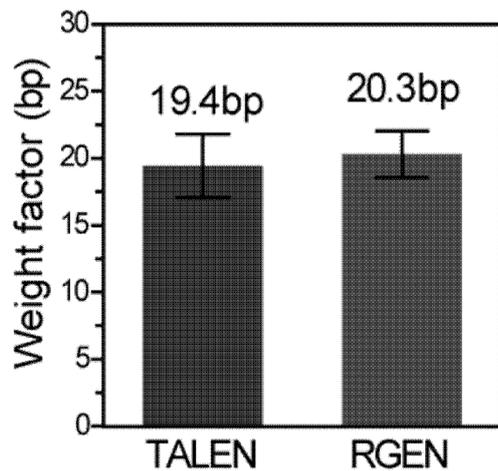
도면4a



도면4b



도면4c



도면5a

```

#!/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()

```

도면5b

```
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])
```

도면5c

```

        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))

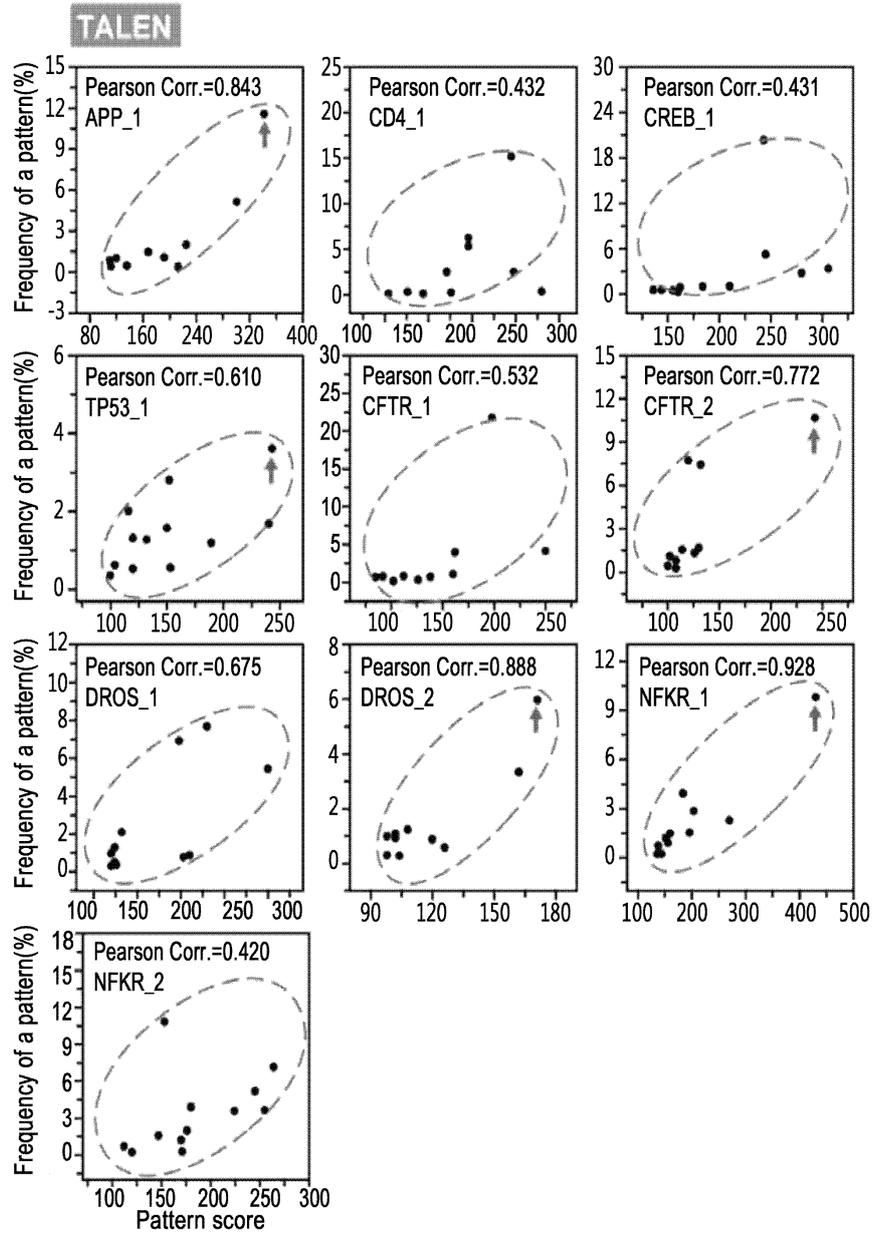
    f2.write(seq[0:left_end]+'-
'+length+seq[right_end:]+'\t'+scrap+'\t'+str(length)+'\t'+str(100*length_factor*((len(scrap)-
num_GC)+(num_GC*2))+'\n')
    sum_score_3+=score_3
    sum_score_not_3+=score_not_3

    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()

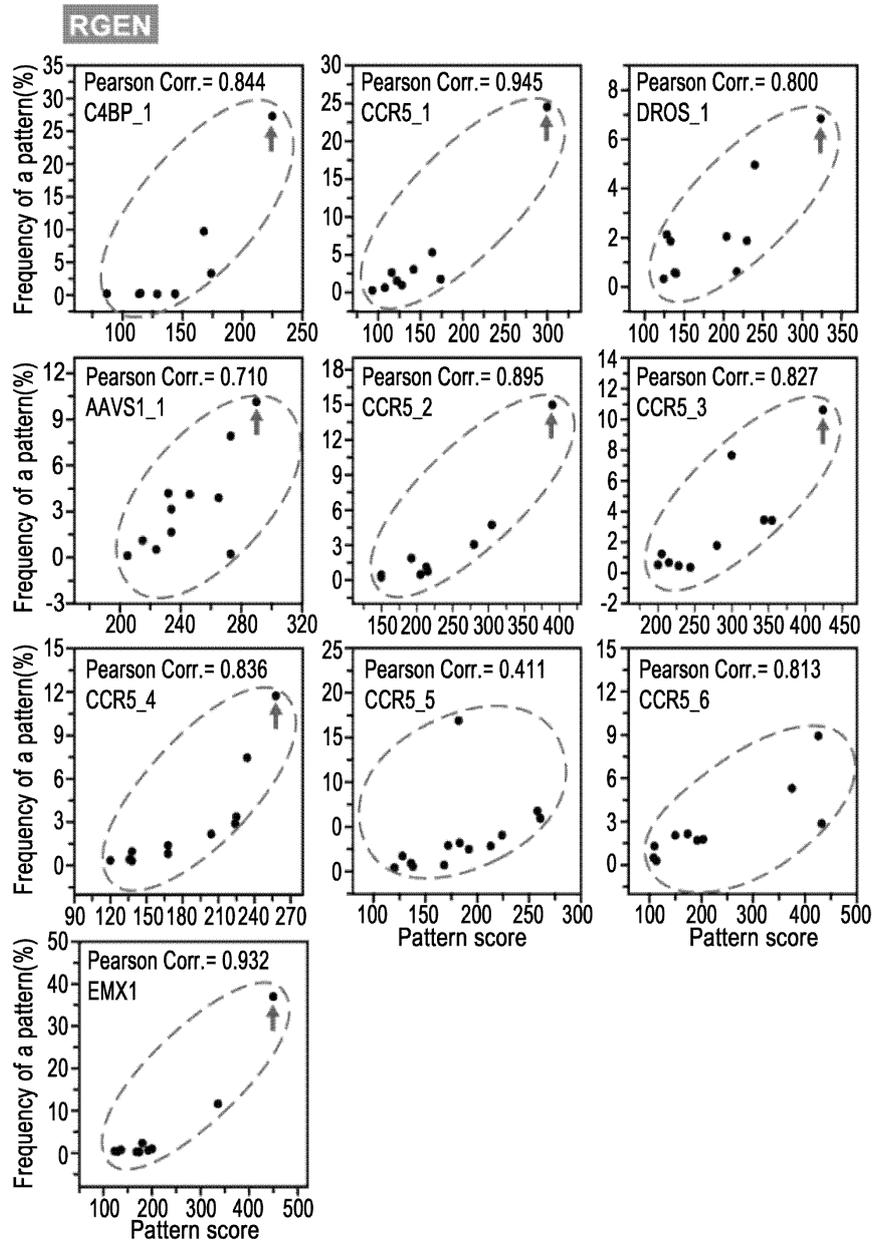
```

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

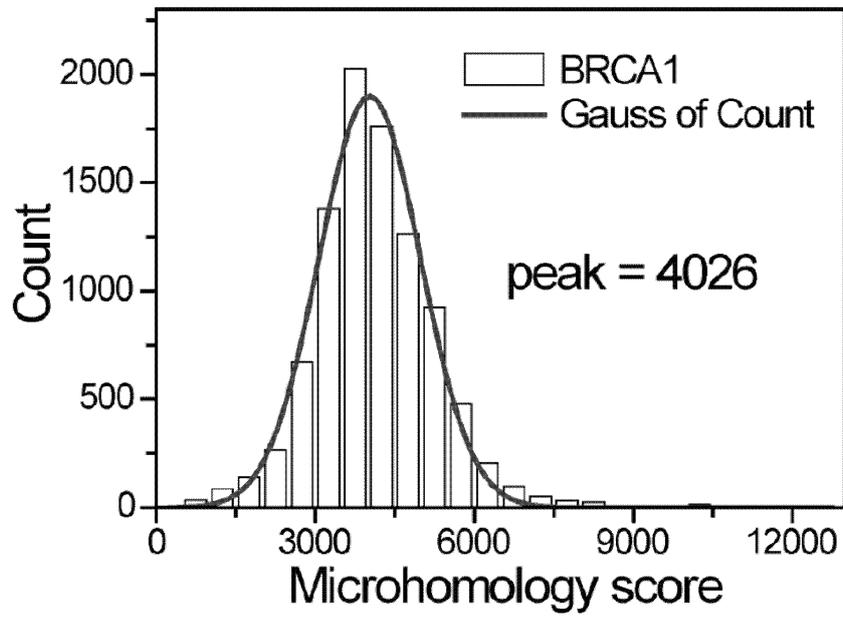
도면6a



도면6b



도면7



도면8

a MCM6_h (71.9) MCM6_l (38.7)
 5' - CACAGCACCTCGTGGCCTGGAGCCTGGCTGGAGCTCTGCTAGGGACAGAAGTGTTCCTGGAAAGTATGCTTCCAGGATT - 3
 3' - GTGTCGTGGAGCACCGGACCTCGGACCGACCTCGAGACGATCCCTGTCTTCACAAAGACCTTCCACTACGAAGGTCCTAA - 5

b

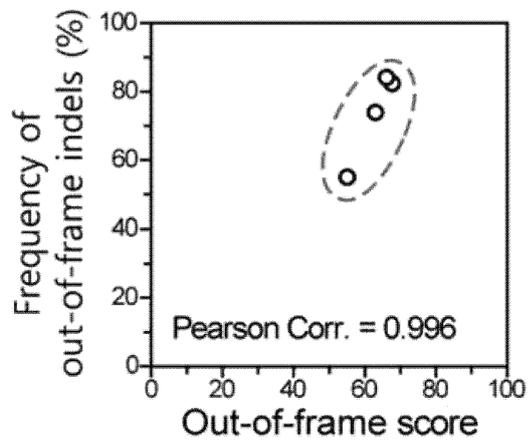
	deletion length	in-frame	frequency
MCM6_h CAGCACCTCGTGGCCTGGAGCCTGGCTGGAGCTCTGCTAGGGACAGAAGTGT CAGCACCTCGTGGCCTGGAGC-----TCTGCTAGGGACAGAAGTGT	- 11	- no	- 18.9%
MCM6_l CTGGAGCTCTGCTAGGGACAGAAGTGTTCCTGGAAAGTATGCTTCCAGGATT CTGGAGCTCTGCTAGGGACAGAAGTGT-----ATGCTTCCAGGATT	- 12	- yes	- 48.5%

도면9

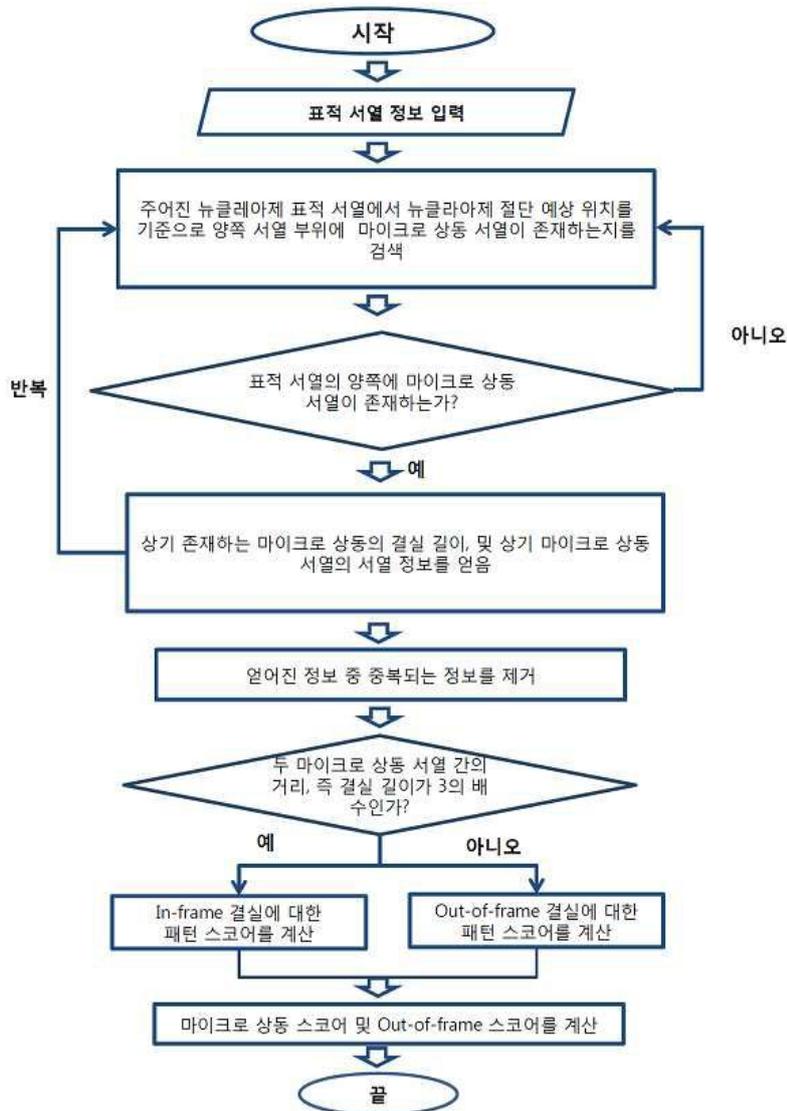
a

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



도면10



서열 목록

- <110> INSTITUTE FOR BASIC SCIENCE
- <120> A method of selecting a nuclease target sequence for gene knockout based on microhomology
- <130> KPA150441-KR
- <150> US 61/983,988
- <151> 2014-04-24
- <150> KR 10-2014-0101133
- <151> 2014-08-06
- <160> 107
- <170> Kopatent In 2.0
- <210> 1

<211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> APP target site
 <400> 1
 tagacccccg ccacagcagc ctctgaagtt ggacagcaaa accattgctt ca 52
 <
 210> 2
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> CD4 target site
 <400> 2
 tgtctcagct ggagctccag gatagtgga cctggacatg cactgtcttg ca 52
 <210> 3
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> TP53 target site
 <400> 3
 tgtccaatga cctgtcccag aagctgtatg ccaccatgga gaagcacaag ga 52
 <210> 4
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> TP53 target site
 <400> 4
 tacaactaca tgtgtaacag ttctgcatg ggcggcatga accggaggcc ca 52
 <210> 5
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> CFTR target site
 <400> 5

tcggaaggca gcctatgtga gataactcaa tagctcagcc ttcttcttct ca 52
 <210> 6
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> CFTR target site
 <400> 6
 tctcttactg ggaagaatca tagcttccta tgacccggat aacaaggagg aa 52
 <210> 7
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> DROSHA target site
 <400> 7
 tgaggaggag attgccaata tgcttcagtg ggaggagctg gaggggcaga aa 52
 <210> 8
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> DROSHA target site
 <400> 8
 tgaaggatac agaatgact gtgaatcaac ccatatcatc aaggagctga ta 52
 <210> 9
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> NFkB1 target site
 <400> 9
 tatgtatgtg aaggcccatc ccatggtgga ctacctggtg cctctagtga aa 52
 <210> 10
 <211> 52
 <212> DNA
 <213> Artificial Sequence

<220><223> NFKB1 target site
 <400> 10
 ttgtcattgc tgttgtccct ctgctacgtt cctattgtca ttaaaggtat ca 52
 <210> 11
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> C4BPB target site
 <400> 11
 aatgaccact acatcctcaa ggg 23
 <210> 12
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> CCR5 target site
 <400>
 > 12
 tgacatcaat tattatacat cgg 23
 <210> 13
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> DROSHA target site
 <400> 13
 gattgccaat atgcttcagt ggg 23
 <210> 14
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> CCR5 target site
 <400> 14
 cctccgctct actcactggg gtt 23
 <210> 15
 <211> 23

<212> DNA
 <213> Artificial Sequence
 <220><223> CCR5 target site
 <400> 15
 cctgcctccg ctctactcac tgg 23
 <210> 16
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> CCR5 target site
 <400> 16
 gaatcctaaa aactctgctt cgg 23
 <210> 17
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> CCR5 target site
 <400> 17
 cctaaaaact ctgcttcggt gtc 23

 <210> 18
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> CCR5 target site
 <400> 18
 aaatgagaag aagaggcaca ggg 23
 <210> 19
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> AAVS1 target site
 <400> 19
 ctccctccca ggatcctctc tgg 23
 <210> 20

<211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> EMX1 target site
 <400>
 20
 gagtccgagc agaagaagaa ggg 23
 <210> 21
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> BRCA1 target site
 <400> 21
 tccagctgct gctcactacta ctgatactgc tgggtataat gcaatggaag aa 52
 <210> 22
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> BRCA1 target site
 <400> 22
 tcctgaacat ctaaagatg aagtttctat catccaaagt atgggctaca ga 52
 <210> 23
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> CXCR4 target site
 <400> 23
 tcttctgcc caccatctac tccatcatct tcttaactgg cattgtgggc aa 52
 <210> 24
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> CXCR4 target site
 <400> 24

tgggttgatt tcagcaccta cagtgtacag tcttgatta agttgtaat aa 52

<210> 25

<211> 52

<212> DNA

<213> Artificial Sequence

<220><223> MCM6 target site

<400> 25

ttagaagtaa ttttaagggc tgaagctgtg gaatcagctc aagctggtga ca 52

<210> 26

<211> 52

<212> DNA

<213> Artificial Sequence

<220><223> MCM6 target site

<400> 26

tggaatcaac ttgatgaaa ccttgcaaa atgtactcca caagtatga ca 52

<210> 27

<211> 52

<212> DNA

<213> Artificial Sequence

<220><223> PHF8 target site

<400> 27

tacagaagc ccaaaagaag aaatatatca agaagaagcc tttgctgaag ga 52

<210> 28

<211> 52

<212> DNA

<213> Artificial Sequence

<220><223> PHF8 target site

<400> 28

tacagcctgc ttgctccgcc tataccacag agcacagcct ggacattatg ga 52

<210> 29

<211> 52

<212> DNA

<213> Artificial Sequence

<220><223> SLC18A2 target site
 <400> 29
 tccagtcata tccgataggt gaagatgaag aatctgaaag tgactgagat ga 52
 <210> 30
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> SLC18A2 target site
 <400> 30
 tgtataaac agtgtttcca gtgacacaac tcatccagaa ctgtcttagt ca 52
 <210> 31
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> TP53 target site
 <400> 31
 tgtaccacca tccactacaa ctacatgtgt aacagttcct gcatgggcgg ca 52
 <210> 32
 <211> 52
 <212> DNA
 <213> Artificial Sequence
 <220><223> TP53 target site
 <400> 32
 ttgtgagcca ccacgtccag ctggaagggt caacatcttt tacattctgc aa 52
 <210> 33
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> APP target site
 <400> 33
 agaggaggaa gaagtggctg agg 23
 <210> 34
 <211> 23

<212> DNA
 <213> Artificial Sequence
 <220><223> APP target site
 <400> 34
 gccacagcag cctctgaagt tgg 23
 <210> 35
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> BRCA1 target site
 <400> 35
 gctcatacta ctgataactgc tgg 23
 <210> 36
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> BRCA1 target site
 <400>
 36
 attgacagct tcaacagaaa ggg 23
 <210> 37
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> MCM6 target site
 <400> 37
 gctagggaca gaagtgtttc tgg 23
 <210> 38
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> MCM6 target site
 <400> 38
 ctcgtggcct ggagcctggc tgg 23
 <210> 39

<211> 21

<212> PRT

<213> Artificial Sequence

<220><223> peptide tag

<400> 39

Gly Gly Ser Gly Pro Pro Lys Lys Lys Arg Lys Val Tyr Pro Tyr Asp
 1 5 10 15

Val Pro Asp Tyr Ala
 20

<210> 40

<211> 23

<212> DNA

<213> Artificial Sequence

<220><223> target site

<400> 40

tggggctgga taatggagcg tgg 23

<210> 41

<211> 23

<212> DNA

<213> Artificial Sequence

<220><223> target site

<400> 41

cggccaaca acgatcccag agg 23

<210> 42

<211> 23

<212> DNA

<213> Artificial Sequence

<220><223> target site

<400> 42

ctgtgaccac gggacggtgc tgg 23

<210> 43

<211> 23

<212> DNA

<213> Artificial Sequence
 <220><223> target site
 <400> 43
 tccatctcgc tcaggtacga ggg 23
 <210>
 > 44
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 44
 gtcccgtgtc ggaaagctgc agg 23
 <210> 45
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 45
 actacaccgc tatgaatgat cgg 23
 <210> 46
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 46
 cccagaggcc cacatacttg ggg 23
 <210> 47
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 47
 acatgccgct attgagcca cgg 23
 <210> 48

<211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 48
 ctgaccgacc cctagaccgc agg 23
 <210> 49
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 49
 gcggaaacac gggaccaagt cgg 23

 <210> 50
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 50
 cccagtgct cggtttgtca cgg 23
 <210> 51
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 51
 caaagctgta ttgcccctcg ggg 23
 <210> 52
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 52

gctccaattg gatctaccgc ggg 23

<210> 53
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 53

tggaccggcg caggtctcca tgg 23

<210> 54
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 54

aggctccgag tagcgcacac tgg 23

<210> 55
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 55

ttgtcgacca ggtttctaca agg 23

<210> 56
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 56

aacaccgaga tccgggatgt agg 23

<210> 57
 <211> 23
 <212> DNA
 <213> Artificial Sequence

<220><223> target site
 <400> 57
 actgcagcgc cgaaggggag tgg 23
 <210> 58
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 58
 tctctcaata cgaattcttg agg 23

 <210> 59
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 59
 cacctggtat gttcgatcg ggg 23
 <210> 60
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 60
 cacatgcatc cccaacgcag agg 23
 <210> 61
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 61
 ggctacggac caagtttacc cgg 23

 <210> 62
 <211> 23

<212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 62
 gcagaatatt cggacaaaca cgg 23
 <210> 63
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 63
 cttcacctt tactaccgtc agg 23
 <210> 64
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 64
 agactgggaa ttacggttac tgg 23

 <210> 65
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 65
 ggaggccgag cttcgtctac tgg 23
 <210> 66
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 66
 ctctgcatgg ttgaccgttc tgg 23
 <210> 67

<211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 67
 cggcaactac acctgcgtcg tgg 23

<210> 68
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 68
 aactccata gtgggtcgag agg 23

<210> 69
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 69
 gcagctgtac gccgtggtgt cgg 23

<210> 70
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 70
 atctacttga tcgaggttga ggg 23

<210> 71
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 71

ctggtcagtt tggcgaagta tgg	23
<210> 72	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	
<400> 72	
gggaccttgc gtacgagagg agg	23
<210> 73	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	
<400> 73	
tgtcgcccgc gttgactctc tgg	23
<210> 74	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	
<400> 74	
agctggcgcc gaatgtatac cgg	23
<210> 75	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	
<400> 75	
tcagtacgcc gtttactca agg	23
<210> 76	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	

<400> 76	
gagaattgct ctgtcatcga agg	23
<210> 77	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	
<400> 77	
aagcggactt taaagttcga ggg	23
<210> 78	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	
<400> 78	
agcaacagag cctatcggca tgg	23
<210> 79	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	
<400> 79	
ctggaaagtc gcagaagggc tgg	23
<210> 80	
<211> 23	
<212> DNA	
<213> Artificial Sequence	
<220><223> target site	
<400> 80	
tccagtttc ctgtgatcgt ggg	23
<210> 81	
<211> 23	
<212> DNA	

<213> Artificial Sequence
 <220><223> target site
 <400> 81
 tatttcatt cgtttcatcc agg 23
 <210> 82
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 82
 gagccttcgt aggtaaccag tgg 23

 <210> 83
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 83
 gccacccgtc gacgtcccct ggg 23
 <210> 84
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 84
 gctcaggagc gttgaacttg agg 23
 <210> 85
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 85
 tggctccaag atactaggcg ggg 23

 <210> 86

<211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 86
 ctattcccg gaccttttcc agg 23
 <210> 87
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 87
 gcatagctac caataagcat ggg 23
 <210> 88
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 88
 cagcctaaga ccaggaacgc cgg 23

 <210> 89
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 89
 agggaacgta gttatcgtaa ggg 23
 <210> 90
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 90
 ggtcctgaat cgtattcttg ggg 23

<210> 91
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 91
 tgctgggtga tgcggccggt ggg 23

<210> 92
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 92
 gtcacgggc cggttatggt ggg 23

<210> 93
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 93
 gccatagatg gtggaccgaa agg 23

<210> 94
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 94
 tgaggtgcac taatagagg tgg 23

<210> 95
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site

<400> 95
tattgcctta catgaattgg ggg 23
<210> 96
<211> 23
<212> DNA
<213> Artificial Sequence
<220><223> target site
<400> 96
gtctgacttc gacaacgcca agg 23
<210> 97
<211> 23
<212> DNA
<213> Artificial Sequence
<220><223> target site
<400> 97
ccacactccg aattcgcctt tgg 23
<210> 98
<211> 23
<212> DNA
<213> Artificial Sequence
<220><223> target site
<400> 98
ggtgatgttg ccgaaaaaga agg 23
<210> 99
<211> 23
<212> DNA
<213> Artificial Sequence
<220><223> target site
<400> 99
cgctgtggg acgggacacg ggg 23
<210> 100
<211> 23
<212> DNA
<213> Artificial Sequence

<220><223> target site
 <400> 100
 cagagttcat attctgtccg agg 23

<210> 101
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 101
 gcagtaggtt gcgctagcg agg 23

<210> 102
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 102
 gccgtggtac tccgtgtgat tgg 23

<210> 103
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 103
 catcagcgtt gatgcagtag agg 23

<210> 104
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 104
 gttgtttacc agccacagct ggg 23

<210> 105
 <211> 23

<212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 105
 gaaccggctg tgtaccgttg tgg 23
 <210> 106
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 106
 ggccacacta gcgttgetgc tgg 23

 <210> 107
 <211> 23
 <212> DNA
 <213> Artificial Sequence
 <220><223> target site
 <400> 107
 tcaggtctgt atttaatggc tgg 23