(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2021-0018131
(43) 공개일자 2021년02월17일

(51) 국제특허분류(Int. Cl.)
C12N 15/85 (2006.01) *A61K 48/00* (2006.01)
C12N 15/10 (2017.01) *C12N 15/113* (2010.01)
C12N 15/90 (2006.01) *C12N 9/22* (2006.01)
(52) CPC특허분류
C12N 15/85 (2013.01)
A61K 48/005 (2013.01)
(21) 출원번호 10-2020-0098119
(22) 출원일자 2020년08월05일
심사청구일자 2020년08월05일
(30) 우선권주장
1020190097643 2019년08월09일 대한민국(KR)

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김형범
서울특별시 마포구 토정로18길 11, 107동 1702호 (현석동, 래미안웰스트림)
송명재
서울특별시 마포구 마포대로 195, 108동 1503호(아현동, 마포 래미안 푸르지오)
(뒷면에 계속)
(74) 대리인
리앤목특허법인

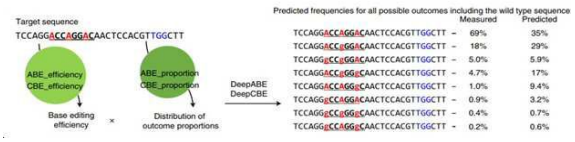
전체 청구항 수 : 총 21 항

(54) 발명의 명칭 **염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템**

(57) 요 약

본 발명은 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템, 및 상기 시스템을 이용하여 염기교정 유전자가위의 효율 및 결과를 예측하는 방법에 관한 것이다. 일 양상에 따른 예측 시스템을 사용하면, 유전자가위를 일일이 제작하여 검증할 필요 없이 간단한 방법으로 효율 및 정확성의 예측이 가능하여 안전한 교정이 가능한 유전자가위를 선별할 수 있다. 나아가, 병원성/유사병원성 인간 점돌연변이 질환 중 염기교정 유전자가위로 질환을 만들거나 교정할 수 있는 경우들의 효율 및 결과 빈도의 예측이 가능하여 염기교정 유전자가위의 대상 질환을 선별할 수 있다.

대 표 도 - 도4



(52) CPC특허분류
C12N 15/102 (2013.01)
C12N 15/113 (2013.01)
C12N 15/907 (2013.01)
C12N 9/22 (2013.01)
C12N 2310/20 (2017.05)
(72) 발명자
김희권
서울특별시 서대문구 연희로10가길 47, 303호(연희동)
이성태
대전광역시 유성구 배울2로 24, 311동 102호(관평동, 대덕테크노밸리3단지아파트)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711109258
과제번호 2017R1A2B3004198
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 개인기초연구(과기정통부)(R&D)
연구과제명 크리스퍼 유전자가위의 활성에 영향을 미치는 인자 규명 및 대량산출(high-throughput)방법을 이용한 유전학 연구 기초 기술 개발
기 여 율 35/100
과제수행기관명 연세대학교
연구기간 2020.03.01 ~ 2021.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711105621
과제번호 2017M3A9B4062403
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 바이오.의료기술개발(R&D)
연구과제명 생체 내 유전자 교정을 통한 근육 및 안 질환 치료 기술 개발
기 여 율 30/100
과제수행기관명 연세대학교
연구기간 2020.01.01 ~ 2020.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1465030234
과제번호 HI17C0676000020
부처명 보건복지부
과제관리(전문)기관명 한국보건산업진흥원
연구사업명 질환극복기술개발(R&D)
연구과제명 효율적인 생체 내 유전자 수술 방법 개발을 통한 유전성 간질환 치료법 발굴
기 여 율 25/100
과제수행기관명 연세대학교 산학협력단
연구기간 2020.01.01 ~ 2020.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1711108917
과제번호 2018R1A5A2025079
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 집단연구지원(R&D)
연구과제명 만성난치질환 시스템의학 연구센터
기 여 율 10/100
과제수행기관명 연세대학교
연구기간 2020.03.01 ~ 2021.02.28

명 세 서

청구범위

청구항 1

염기교정 유전자가위의 표적 서열을 입력 받는 표적 서열 입력부; 및

상기 표적 서열 입력부에서 입력 받은 표적 서열을 효율 예측 모델 및 교정결과 예측 모델에 각각 적용하여 염기교정 유전자가위의 효율 및 교정결과 스코어를 획득하고, 상기 효율 스코어와 교정결과 스코어를 곱하여 염기교정 유전자가위의 효율 및 결과를 동시에 예측하는 결과 예측부를 포함하는 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 2

청구항 1에 있어서,

상기 효율 예측 모델은

염기교정 유전자가위의 활성 데이터를 정보 입력부를 통해 입력 받는 단계; 및

상기 정보 입력부에서 입력 받은 데이터를 이용하여 컨볼루션 신경망(convolutional neural network: CNN)을 기반으로 한 딥러닝을 수행하여 효율 예측 모델을 생성하는 단계를 통해 생성되는 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 3

청구항 1에 있어서,

상기 교정결과 예측 모델은

염기교정 유전자가위의 교정결과 데이터를 입력 받는 정보 입력부를 통해 입력 받는 단계; 및

상기 정보 입력부에서 입력 받은 데이터를 이용하여 컨볼루션 신경망(convolutional neural network: CNN)을 기반으로 한 딥러닝을 수행하여 교정결과 예측 모델을 생성하는 단계를 통해 생성되는 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 4

청구항 1에 있어서,

상기 염기교정 유전자가위는 아데닌 염기교정 유전자가위(Adenine Base Editor: ABE) 또는 시토신 염기교정 유전자가위(Cytosine Base Editor: CBE)인 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 5

청구항 2에 있어서,

상기 염기교정 유전자가위의 활성 데이터는 염기교정 유전자가위가 목적하는 표적 뉴클레오티드 주변의 서열 컨텍스트(context)가 고려된 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 6

청구항 2에 있어서,

상기 염기교정 유전자가위의 활성 데이터는

가이드 RNA를 코딩하는 염기서열 및 상기 가이드 RNA가 목적하는 표적 서열을 포함하는 올리고뉴클레오티드를 포함하는 세포 라이브러리에 염기교정 유전자가위를 도입하는 단계;

상기 염기교정 유전자가위가 도입된 세포 라이브러리로부터 분리한 DNA를 이용하여 딥 시퀀싱을 수행하는 단계; 및

상기 딥 시퀀싱으로부터 수득한 서열 데이터로부터 염기교정 범위 내 표적 뉴클레오티드 전환 여부를 검출하는 단계를 통해 수득되는 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 7

청구항 3에 있어서,

상기 염기교정 유전자가위의 교정결과 데이터는

가이드 RNA를 코딩하는 염기서열 및 상기 가이드 RNA가 목적하는 표적 서열을 포함하는 올리고뉴클레오티드를 포함하는 세포 라이브러리에 염기교정 유전자가위를 도입하는 단계;

상기 염기교정 유전자가위가 도입된 세포 라이브러리로부터 분리한 DNA를 이용하여 딥 시퀀싱을 수행하는 단계;

상기 딥 시퀀싱으로부터 수득한 서열 데이터로부터 염기교정 범위 내 표적 뉴클리오티드 전환 빈도를 검출하는 단계를 통해 수득되는 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 8

청구항 1에 있어서,

상기 효율 스코어는 표적 서열의 각 위치에 대하여 하기 [수학식 1]을 이용하여 산출되는 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

[수학식 1]

$$\text{염기 편집 효율 (\%)} = \frac{\text{넓은염기교정 범위(표적 서열의 위치3 내지10)에서 의도된표적 뉴클레오타드 전환을 포함하는 모든 서열의 총 리드(read)}}{\text{총 리드(read)}} \times 100$$

청구항 9

청구항 1에 있어서,

상기 교정결과 스코어는 하기 [수학식 2]을 이용하여 산출되는 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

[수학식 2]

$$\text{염기 편집결과 빈도} = \frac{\text{특정 염기 - 편집된 결과 서열의 리드(read)}}{\text{넓은염기교정 범위(표적 서열의 위치3 내지10)에서의도된표적 뉴클레오타드 전환을 포함하는 모든 서열의 총 리드(read)}}$$

청구항 10

청구항 1에 있어서,

상기 표적 서열은 24 내지 26개의 뉴클레오티드로 구성된 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 11

청구항 1에 있어서,

상기 표적 서열은 PAM 서열 및 프로토스페이서 서열을 포함하는 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 12

청구항 1에 있어서,

상기 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템은 결과 예측부에서 예측된 염기교정 유전자가위의 효율 및 결과를 출력하는 출력부를 추가로 포함하는 것인, 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템.

청구항 13

염기교정 유전자가위의 표적 서열을 설계하는 단계; 및

상기 설계된 표적 서열을 청구항 1에 따른 염기교정 효율 및 결과 예측 시스템에 적용하는 단계를 포함하는 염기교정 유전자가위의 염기교정 효율 및 결과 예측 방법.

청구항 14

인간 점돌연변이 데이터를 수득하는 단계;

상기 인간 점돌연변이 데이터 중에서 병원성 또는 유사병원성 점돌연변이에 해당하는 데이터를 1차로 선별하는 단계;

상기 1차로 선별된 데이터 중에서 점돌연변이가 정상 염기 아데닌(A)이 비정상 염기 구아닌(G)으로 바뀌어 발생하는 경우; 정상 염기 구아닌(G)이 비정상 염기 아데닌(A)으로 바뀌어 발생하는 경우; 정상 염기 시토신(C)이 비정상 염기 티민T)으로 바뀌어 발생하는 경우; 또는 정상 염기 티민(T)이 비정상 염기 시토신(C)으로 바뀌어 발생하는 경우에 해당하는 데이터를 2차로 선별하는 단계;

상기 2차로 선별된 데이터 중에서 점돌연변이가 프로토스페이서 영역의 5' 말단으로부터 3 내지 10 bp 위치에 존재하는 데이터를 3차로 선별하는 단계;

및

상기 3차로 선별된 데이터를 청구항 1에 따른 염기교정 효율 및 결과 예측 시스템에 적용하는 단계를 포함하는 염기교정 유전자가위를 사용할 수 있는 인간 점돌연변이 관련 질환에 대한 정보를 제공하는 방법.

청구항 15

청구항 12에 있어서,

상기 인간 점돌연변이 관련 질환은 어셔 증후군(Usher syndrome), 종양괴사인자 수용체 관련 주기적 증후군(TNF receptor-associated periodic syndrome: TRAPS), 마판 증후군(marfan syndrome), 제3형 청년기 발병 당뇨병(Type 3 form of Maturity-Onset Diabetes of the Young: MODY3), 선천성 비진행성 야맹증(Congenital

stationary night blindness type 1F), 가족성 고콜레스테롤혈증(Familial hypercholesterolemia), 선천근육무력증후군(congenital myasthenic syndrome: CMS), 린치증후군(Lynch syndrome) 등이 확인되었고, CBE의 경우 로이-디에츠 증후군(Loeys-Dietz syndrome: LDS), 망막색소변성증(retinitis pigmentosa), 렙틴 결핍 또는 장애(Leptin deficiency 또는 dysfunction), 가족성 고콜레스테롤혈증(Familial hypercholesterolemia), 상염색체 열성 청각장애(autosomal recessive deafness), 콜레스테롤 모노옥시다제 결핍(cholesterol monooxygenase (side-chain-cleaving) deficiency) 및 진행성 근간대성간질(progressive myoclonus epilepsy)로 이루어진 군으로부터 선택되는 어느 하나인 것인, 염기교정 유전자가위를 사용할 수 있는 인간 점돌연변이 관련 질환에 대한 정보를 제공하는 방법.

청구항 16

청구항 13 내지 15에 따른 방법을 컴퓨터로 실행하기 위한 프로그램이 기록된 컴퓨터 판독가능 기록매체.

청구항 17

염기교정 유전자가위를 세포에 도입하는 단계를 포함하는 세포의 유전체에서 표적 뉴클레오티드를 편집하는 방법으로서,

상기 염기교정 유전자가위는 (i) RNA-가이드 뉴클레아제 또는 이를 코딩하는 유전자, (ii) 탈아미노효소 또는 이를 코딩하는 유전자, 및 (iii) 표적 서열과 혼성화 할 수 있는 가이드 RNA 또는 이를 코딩하는 유전자를 포함하고,

상기 표적 서열은 PAM 서열, 프로토스페이서 서열, 및 가이드 RNA에 상보적인 서열을 포함하고,

상기 가이드 RNA에 상보적인 서열은 5'-TAC-3', 5'-TAT-3', 5'-TAG-3', 5'-GAT-3', 5'-CAC-3', 5'-GAC-3', 5'-CAT-3', 5'-TAA-3', 5'-CAG-3', 5'-GAG-3', 5'-AAC-3', 5'-CAA-3', 5'-GAA-3', 5'-AAT-3', 5'-AAG-3', 5'-AAA-3', 5'-TCC-3', 5'-TCG-3', 5'-TCT-3', 5'-TCA-3', 5'-CCC-3', 5'-CCT-3', 5'-ACC-3', 5'-CCA-3', 5'-CCG-3', 5'-ACG-3', 5'-ACT-3', 5'-ACA-3', 5'-GCC-3', 5'-GCT-3', 5'-GCG-3', 및 5'-GCA-3'으로 이루어진 군으로부터 선택되는 서열을 포함하고,

상기 탈아미노효소는 표적 서열에서 아데닌 또는 시토신을 탈아미노화하는 것을 특징으로 하는 세포의 유전체에서 표적 뉴클레오티드를 편집하는 방법.

청구항 18

청구항 17에 있어서,

상기 RNA-가이드 뉴클레아제는 SpCas9, nCas9, 및 dCas9로 이루어진 군으로부터 선택되는 것인, 세포의 유전체에서 표적 뉴클레오티드를 편집하는 방법.

청구항 19

청구항 17에 있어서,

상기 표적 뉴클레오티드는 프로토스페이서 영역의 5' 말단으로부터 3 내지 10 bp 위치에 존재하는 것인, 세포의 유전체에서 표적 뉴클레오티드를 편집하는 방법.

청구항 20

염기교정 유전자가위를 포함하는 인간 점돌연변이 관련 질환의 예방 또는 치료용 약학적 조성물로서,

상기 염기교정 유전자가위는 (i) RNA-가이드 뉴클레아제 또는 이를 코딩하는 유전자, (ii) 탈아미노효소 또는

이를 코딩하는 유전자, 및 (iii) 표적 서열과 혼성화 할 수 있는 가이드 RNA 또는 이를 코딩하는 유전자를 포함하고,

상기 표적 서열은 PAM 서열, 프로토스페이서 서열, 및 가이드 RNA에 상보적인 서열을 포함하고,

상기 가이드 RNA에 상보적인 서열은 5'-TAC-3', 5'-TAT-3', 5'-TAG-3', 5'-GAT-3', 5'-CAC-3', 5'-GAC-3', 5'-CAT-3', 5'-TAA-3', 5'-CAG-3', 5'-GAG-3', 5'-AAC-3', 5'-CAA-3', 5'-GAA-3', 5'-AAT-3', 5'-AAG-3', 5'-AAA-3', 5'-TCC-3', 5'-TCG-3', 5'-TCT-3', 5'-TCA-3', 5'-CCC-3', 5'-CCT-3', 5'-ACC-3', 5'-CCA-3', 5'-CCG-3', 5'-ACG-3', 5'-ACT-3', 5'-ACA-3', 5'-GCC-3', 5'-GCT-3', 5'-GCG-3', 및 5'-GCA-3'으로 이루어진 군으로부터 선택되는 서열을 포함하고,

상기 탈아미노효소는 표적 서열에서 아데닌 또는 시토신을 탈아미노화하는 것을 특징으로 하는 인간 점돌연변이 관련 질환의 예방 또는 치료용 약학적 조성물.

청구항 21

청구항 20에 있어서,

상기 인간 점돌연변이 관련 질환은 어셔 증후군(Usher syndrome), 종양괴사인자 수용체 관련 주기적 증후군(TNF receptor-associated periodic syndrome: TRAPS), 마판 증후군(marfan syndrome), 제3형 청년기 발병 당뇨병(Type 3 form of Maturity-Onset Diabetes of the Young: MODY3), 선천성 비진행성 야맹증(Congenital stationary night blindness type 1F), 가족성 고콜레스테롤혈증(Familial hypercholesterolemia), 선천근육무력증후군(congenital myasthenic syndrome: CMS), 린치증후군(Lynch syndrome) 등이 확인되었고, CBE의 경우 로이-디에츠 증후군(Loeys-Dietz syndrome: LDS), 망막색소변성증(retinitis pigmentosa), 렙틴 결핍 또는 장애(Leptin deficiency 또는 dysfunction), 가족성 고콜레스테롤혈증(Familial hypercholesterolemia), 상염색체 열성 청각장애(autosomal recessive deafness), 콜레스테롤 모노옥시다제 결핍(cholesterol monooxygenase (side-chain-cleaving) deficiency) 및 진행성 근간대성간질(progressive myoclonus epilepsy)로 이루어진 군으로부터 선택되는 어느 하나인 것인, 인간 점돌연변이 관련 질환의 예방 또는 치료용 약학적 조성물.

발명의 설명

기 술 분 야

[0001] 본 발명은 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템, 및 상기 시스템을 이용하여 염기교정 유전자가위의 효율 및 결과를 예측하는 방법에 관한 것이다.

배 경 기 술

[0002] 점돌연변이는 인간에서 병원성(pathogenic) 또는 유사병원성(likely pathogenic) 돌연변이의 절반 이상을 차지하는 가장 일반적인 형태의 병원성 돌연변이나, 이 빈도는 짧은 판독 시퀀싱의(short-read sequencing) 광범위한 사용으로 인해 편향될 수 있다. 정상 세포 및 유기체에서 이러한 병원성 점돌연변이의 생성은 관련 질환 모델의 발달로 이어질 수 있다. 반대로, 돌연변이를 가진 세포 및 유기체에서 병원성 점돌연변이의 교정은 이러한 점돌연변이의 영향에 대한 연구에 동질유전자형(isogenic) 대조군을 제공할 수 있다. 또한, 이러한 병원성 점돌연변이의 교정은 관련 질환에 대한 치료적 모달리티(modality)가 될 수 있다. 병원성 점돌연변이의 생성 및 교정 모두에 있어서, 염기교정 유전자가위(base editors)는 이중가닥 절단(double-strand break)을 생성하거나 공여자 DNA 주형을 요구하지 않고 표적화된 방식으로 하나의 염기쌍을 다른 염기쌍으로 직접 전환할 수 있어 매우 매력적인 유전체(genome) 편집 도구이다. 아데닌 염기교정 유전자가위(Adenine base editors: ABEs)는 A, T 염기쌍을 G, C 염기쌍으로 전환할 수 있고, 시토신 염기교정 유전자가위(cytosine base editors: CBEs)는 G, C 염기쌍을 A, T 염기쌍으로 전환할 수 있다.

[0003] 그러나, i) 염기 편집 효율이 낮고/낮거나 ii) 염기 편집의 결과로 원치 않는 동시 돌연변이(concurrent mutations)가 발생하는 경우, 특히 편집가능한 윈도우(editable window) - 즉 염기교정 범위에 다수의 표적 뉴클레오티드가 있는 경우, 이러한 염기교정 유전자가위로 유도된 질환 모델 및 병원성 돌연변이의 교정은 어려울 수 있다. 따라서, 염기 편집의 효율 및 결과는 종종 질환모델 생성 전이나 생성되는 동안, 병원성 돌연변이의 치료적 교정 도중 실험적으로 측정된다. 그러나, 이와 같은 실험적 평가는 단일가닥 가이드 RNA(single-guide

RNA: sgRNA)의 제조, sgRNA와 함께 ABE 또는 CBE의 전달, 이들 성분을 함유하는 세포의 수확, 표적 서열의 PCR 증폭, 및 이어지는 시퀀싱을 포함하는 시간 소모적인 다단계 과정이다. 또한, 염기 편집을 효율적인 고-처리량 스크리닝(high-throughput screening)을 위한 도구로서 사용하려면 각각의 표적 시퀀스에서 염기 편집의 효율 및 결과를 알아야 한다. 그러나, 수천 개의 표적 서열이 연구되는 경우, 이와 같이 개별 부위 각각에서의 효율을 평가하는 기존의 평가 방법은 실용적인 접근법이라 할 수 없다. 또한, 환자로부터 유래된 세포가 관련 돌연변이를 포함하는 경우에는 이와 같은 평가를 실행할 수 없다. 즉, 염기교정 유전자가위가 만들 수 있는 다양한 염기 교정결과물들의 빈도를 예측하는 방법은 현재까지 전무하다.

발명의 내용

해결하려는 과제

[0004] 이에, 본 발명자들은 염기교정 유전자가위의 염기교정 효율과 위치별 염기 편집 빈도를 통해 교정결과를 예측할 수 있는 *in silico* 방법을 개발하고자 노력하였다. 그 결과, 아데닌 염기교정 유전자가위 및 시토신 염기교정 유전자가위에 대하여 각각 13,000여개 및 14,000여개의 표적 서열에서 이의 효율 및 교정결과 빈도 데이터를 생산하고, 염기교정 유전자가위의 효율에 유의한 영향을 미치는 표적 염기 주변의 서열 컨텍스트를 탐색하여, 상기 대규모 데이터에 근거한 딥러닝(Deep learning) 방법을 통해 염기교정 유전자가위의 효율과 정확성을 동시에 예측할 수 있는 시스템을 개발하고, 나아가 인간 점돌연변이 질환에 대한 염기교정 유전자가위의 효율을 예측하여 염기교정 유전자가위로 만들 수 있는 질환 및 교정가능한 질환을 선별할 수 있음을 확인하여 본 발명을 완성하였다.

[0005] 일 양상은

[0006] 염기교정 유전자가위의 표적 서열을 입력 받는 표적 서열 입력부; 및

[0007] 상기 표적 서열 입력부에서 입력 받은 표적 서열을 효율 예측 모델 및 교정결과 예측 모델에 각각 적용하여 염기교정 유전자가위의 효율 및 교정결과 스코어를 획득하고, 상기 효율 스코어와 교정결과 스코어를 곱하여 염기교정 유전자가위의 효율 및 결과를 동시에 예측하는 결과 예측부를 포함하는 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템을 제공하는 것이다.

[0008] 다른 양상은

[0009] 염기교정 유전자가위의 표적 서열을 설계하는 단계; 및

[0010] 상기 설계된 표적 서열을 일 양상에 따른 염기교정 효율 및 결과 예측 시스템에 적용하는 단계; 를 포함하는 염기교정 유전자가위의 염기교정 효율 및 결과 예측 방법을 제공하는 것이다.

[0011] 또 다른 양상은

[0012] 인간 점돌연변이 데이터를 수득하는 단계;

[0013] 상기 인간 점돌연변이 데이터로부터 점돌연변이가 정상 염기 아데닌(A)이 비정상 염기 구아닌(G)으로 바뀌어 발생하는 경우; 정상 염기 구아닌(G)이 비정상 염기 아데닌(A)으로 바뀌어 발생하는 경우; 정상 염기 시토신(C)이 비정상 염기 티민T)으로 바뀌어 발생하는 경우; 또는 정상 염기 티민(T)이 비정상 염기 시토신(C)으로 바뀌어 발생하는 경우에 해당하는 데이터를 1차로 선별하는 단계;

[0014] 상기 1차로 선별된 데이터 중에서 점돌연변이가 프로토스페이서 영역의 5' 말단으로부터 3 내지 10 bp 위치에 존재하는 데이터를 2차로 선별하는 단계;

[0015] 상기 2차로 선별된 데이터 중에서 병원성 또는 유사병원성 점돌연변이에 해당하는 데이터를 3차로 선별하는 단계; 및

[0016] 상기 3차로 선별된 데이터를 일 양상에 따른 염기교정 효율 및 결과 예측 시스템에 적용하는 단계를 포함하는 염기교정 유전자가위를 사용할 수 있는 인간 점돌연변이 관련 질환에 대한 정보를 제공하는 방법을 제공하는 것이다.

[0017] 또 다른 양상은 상기 방법을 컴퓨터로 실행하기 위한 프로그램이 기록된 컴퓨터 판독가능 기록매체를 제공하는 것이다.

[0018] 또 다른 양상은

[0019] 염기교정 유전자가위를 세포에 도입하는 단계; 를 포함하는 세포의 유전체에서 표적 뉴클레오티드를 편집하는 방법으로서,

[0020] 상기 염기교정 유전자가위는 (i) RNA-가이드 뉴클레아제 또는 이를 코딩하는 유전자, (ii) 탈아미노효소 또는 이를 코딩하는 유전자, 및 (iii) 표적 서열과 혼성화 할 수 있는 가이드 RNA 또는 이를 코딩하는 유전자를 포함하고,

[0021] 상기 표적 서열은 PAM 서열, 프로토스페이서 서열, 및 가이드 RNA에 상보적인 서열을 포함하고,

[0022] 상기 가이드 RNA에 상보적인 서열은 5'-TAC-3', 5'-TAT-3', 5'-TAG-3', 5'-GAT-3', 5'-CAC-3', 5'-GAC-3', 5'-CAT-3', 5'-TAA-3', 5'-CAG-3', 5'-GAG-3', 5'-AAC-3', 5'-CAA-3', 5'-GAA-3', 5'-AAT-3', 5'-AAG-3', 5'-AAA-3', 5'-TCC-3', 5'-TCG-3', 5'-TCT-3', 5'-TCA-3', 5'-CCC-3', 5'-CCT-3', 5'-ACC-3', 5'-CCA-3', 5'-CCG-3', 5'-ACG-3', 5'-ACT-3', 5'-ACA-3', 5'-GCC-3', 5'-GCT-3', 5'-GCG-3', 및 5'-GCA-3'으로 이루어진 군으로부터 선택되는 서열을 포함하고,

[0023] 상기 탈아미노효소는 표적 서열에서 아데닌 또는 시토신을 탈아미노화하는 것을 특징으로 하는 세포의 유전체에서 표적 뉴클레오티드를 편집하는 방법을 제공하는 것이다.

[0024] 또 다른 양상은

[0025] 염기교정 유전자가위를 포함하는 인간 점돌연변이 관련 질환의 예방 또는 치료용 약학적 조성물로서,

[0026] 상기 염기교정 유전자가위는 (i) RNA-가이드 뉴클레아제 또는 이를 코딩하는 유전자, (ii) 탈아미노효소 또는 이를 코딩하는 유전자, 및 (iii) 표적 서열과 혼성화 할 수 있는 가이드 RNA 또는 이를 코딩하는 유전자를 포함하고,

[0027] 상기 표적 서열은 PAM 서열, 프로토스페이서 서열, 및 가이드 RNA에 상보적인 서열을 포함하고,

[0028] 상기 가이드 RNA에 상보적인 서열은 5'-TAC-3', 5'-TAT-3', 5'-TAG-3', 5'-GAT-3', 5'-CAC-3', 5'-GAC-3', 5'-CAT-3', 5'-TAA-3', 5'-CAG-3', 5'-GAG-3', 5'-AAC-3', 5'-CAA-3', 5'-GAA-3', 5'-AAT-3', 5'-AAG-3', 5'-AAA-3', 5'-TCC-3', 5'-TCG-3', 5'-TCT-3', 5'-TCA-3', 5'-CCC-3', 5'-CCT-3', 5'-ACC-3', 5'-CCA-3', 5'-CCG-3', 5'-ACG-3', 5'-ACT-3', 5'-ACA-3', 5'-GCC-3', 5'-GCT-3', 5'-GCG-3', 및 5'-GCA-3'으로 이루어진 군으로부터 선택되는 서열을 포함하고,

[0029] 상기 탈아미노효소는 표적 서열에서 아데닌 또는 시토신을 탈아미노화하는 것을 특징으로 하는 인간 점돌연변이 관련 질환의 예방 또는 치료용 약학적 조성물을 제공하는 것이다.

[0030] 본 출원의 다른 목적 및 이점은 첨부한 청구범위 및 도면과 함께 하기의 상세한 설명에 의해 보다 명확해질 것이다. 본 명세서에 기재되지 않은 내용은 본 출원의 기술 분야 또는 유사한 기술 분야 내 숙련된 자이면 충분히 인식하고 유추할 수 있는 것이므로 그 설명을 생략한다.

과제의 해결 수단

[0031] 본 출원에서 개시된 각각의 설명 및 실시형태는 각각의 다른 설명 및 실시형태에도 적용될 수 있다. 즉, 본 출원에서 개시된 다양한 요소들의 모든 조합이 본 출원의 범주에 속한다. 또한, 하기 기술된 구체적인 서술에 의하여 본 출원의 범주가 제한된다고 볼 수 없다.

[0032] 일 양상은

[0033] 염기교정 유전자가위의 표적 서열을 입력 받는 표적 서열 입력부; 및

[0034] 상기 표적 서열 입력부에서 입력 받은 표적 서열을 효율 예측 모델 및 교정결과 예측 모델에 각각 적용하여 염기교정 유전자가위의 효율 및 교정결과 스코어를 획득하고, 상기 효율 스코어와 교정결과 스코어를 곱하여 염기교정 유전자가위의 효율 및 결과를 동시에 예측하는 결과 예측부를 포함하는 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템을 제공한다.

[0035] 본원에서, 용어 "염기교정 유전자가위(Base editor)"는 4세대 유전자가위 기술이라고 불리는 크리스퍼 유전자가위에서 유래된 새로운 타입의 유전자 가위이다. 염기교정 유전자가위는 DNA 두 가닥 모두를 자르는 기존 3세대 유전자가위와 다르게, 단일 염기를 교체하는 방식으로 작동한다. 염기교정 유전자가위는 DNA 한쪽 가닥을 자르는 Nickase Cas9(nCas9)와 아데닌 또는 시토신을 분해하는 탈아미노효소로 구성되어 있으며, 구체적으로

CRISPR/Cas9의 이중가닥 DNA 절단기능을 제거한 dCas9("dead" Cas9) 또는 nCas9에 아데닌 탈아미노효소(Adenine deaminase)를 결합하여 아데닌(A)을 구아닌(G)으로 교체할 수 있는 아데닌 염기교정 유전자가위(Adenine Base Editor: ABE)와 시토신 탈아미노효소(cytosine deaminase)를 결합하여 DNA 서열 중 시토신(C)만 찾아 티민(T)으로 교체할 수 있는 시토신 염기교정 유전자가위(Cytosine Base Editor: CBE)가 있다. 예를 들어, CBE의 경우 nCas9 또는 dCas9로 잘려진 DNA 한 가닥에서 탈아미노효소가 시토신(C)을 우라실(U)로 교체하면, 우라실(U)로 바뀐 염기는 DNA 복구 과정에 의해 티민(T)이 되는 원리로 작동한다. 염기교정 유전자가위를 이용하면 특정 서열을 교정하거나 교체하여 유전자를 결손시키거나 원하는 형질로 전환할 수 있다.

[0036] 본 발명자들은 고-처리량(high-throughput) 실험을 통해 프로토스페이서 영역의 5' 말단으로부터 3 내지 10 bp 위치(20 bp 위치는 PAM 서열(5'-NGG'-3')의 바로 상류에 자리함)에서 적어도 하나의 표적 아데닌을 포함하는 13,504개, 적어도 하나의 표적 시토신을 포함하는 14,157개의 표적 서열에 대해 염기교정 유전자가위의 활성 확인 및 염기 교정결과에 대한 대규모 데이터를 확보하고, 컨볼루션 신경망을 사용한 딥러닝으로 구축한 효율 예측 모델 및 교정결과 예측 모델의 2가지 모델을 결합하여 염기교정 효율 및 염기교정 유전자가위가 만들 수 있는 모든 염기 편집 결과물들에 대한 예측 수행이 가능한 DeepABE 및 DeepCBE 예측 시스템을 개발하고(DeepBaseEditor), 상기 예측 시스템의 정확성 검증을 통해 염기교정 유전자가위의 효율 및 교정결과를 동시에 예측할 수 있음을 확인하였다.

[0037] 본 발명자들은 상기 구축된 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템이 전통적인 기계 학습(machine learning) 기반 알고리즘에 비해 우수한 성능을 가지는 것을 확인하였다.

[0038] 본원에서 용어, "가이드 RNA (guide RNA)"는 표적 DNA 특이적인 RNA를 의미하며, 표적 서열과 전부 또는 일부 상보적으로 결합하여 염기교정 유전자가위의 아데닌 탈아미노효소 또는 시토신 탈아미노효소가 표적 서열 중 각각 아데닌(A)을 찾아 구아닌(G)으로, 시토신(C)을 찾아 티민(T)으로 교체할 수 있다.

[0039] 통상적으로 가이드 RNA는 두 개의 RNA, 즉, crRNA (CRISPR RNA) 및 tracrRNA (trans-activating crRNA)를 구성요소로 포함하는 이중 RNA (dual RNA); 또는 표적 DNA 내 서열과 전부 또는 일부 상보적인 서열을 포함하는 제1 부위 및 RNA-가이드 뉴클레아제와 상호작용하는 서열을 포함하는 제2 부위를 포함하는 형태를 말하나, 염기교정 유전자가위의 RNA-가이드 뉴클레아제가 표적 서열에서 활성을 가질 수 있는 형태라면 제한 없이 본 발명의 범위에 포함될 수 있다.

[0040] 또한, 상기 가이드 RNA는 RNA-가이드 뉴클레아제가 부착되는 것을 돕는 스캐폴드(scaffold) 서열을 포함할 수 있다.

[0041] 본원에서 용어, "표적 서열" 또는 "타겟 서열"은 염기교정 유전자가위가 표적으로 할 것으로 예상되는 염기서열을 의미한다. 구체적으로, 염기교정 유전자가위가 가이드 RNA를 통해 표적으로 할 것으로 예상되는 서열로서, 염기교정 유전자가위가 활성을 나타내는 것으로 알려진 서열일 수 있고, 또는 본 발명의 시스템을 이용하는 당업자가 분석하고자 하는 서열을 임의로 설계한 서열일 수도 있으나, 염기교정 유전자가위가 활성을 갖거나, 또는 가질 것으로 예상되어 분석하고자 하는 서열이라면, 본 발명의 범주에 제한 없이 포함될 수 있다.

[0042] 본원에서, 염기교정 유전자가위의 활성 데이터는, 가이드 RNA를 코딩하는 염기서열 및 상기 가이드 RNA가 목적하는 표적 서열을 포함하는 올리고뉴클레오티드를 포함하는 세포 라이브러리에 염기교정 유전자가위를 도입함으로써 수득될 수 있으나, 이에 제한되지 않는다.

[0043] 본원에서, 용어 "RNA-가이드 뉴클레아제"는 목적하는 유전체 상의 특정 위치를 인식하여 절단할 수 있는 뉴클레아제로서, 특히 가이드 RNA에 의해 표적 특이성을 갖는 뉴클레아제를 말한다. 상기 RNA-가이드 뉴클레아제는 이에 제한되는 것은 아니나, Cas9 (CRISPR-Associated Protein 9) 및 Cpf1 등이 포함될 수 있다.

[0044] 본원에서, "Cas9 단백질"은 CRISPR/Cas9 시스템의 주요 단백질 구성 요소로, crRNA(CRISPR RNA) 및 tracrRNA(trans-activating crRNA)와 복합체를 형성하여 활성화된 엔도뉴클레아제(endonuclease) 또는 니카아제(nickase)를 형성한다.

[0045] Cas9 단백질 또는 유전자 정보는 NCBI(National Center for Biotechnology Information)의 GenBank와 같은 공지의 데이터 베이스에서 얻을 수 있으나, 가이드 RNA와 함께 표적 특이적 뉴클레아제 활성을 가질 수 있는 것이라면 모두 본 발명의 범위에 포함될 수 있다. 또한, Cas9 단백질은 단백질 전달 도메인(protein transduction domain)과 연결될 수 있다. 상기 단백질 전달 도메인은 폴리아르기닌 또는 HIV 유래의 TAT 단백질일 수 있으나, 이에 제한되지 않는다. 나아가, 상기 Cas9 단백질은 그 목적에 따라 당업자에 의해 추가적인 도메인이 적절하게

연결될 수 있다.

[0046] 상기 Cas9 단백질은 야생형 Cas9 뿐만 아니라, 불활성화된 Cas9 (dCas9), 또는 Cas9 니케이즈(nickase)와 같은 Cas9의 변이체를 모두 포함할 수 있다. 상기 불활성화된 Cas9은 dCas9에 FokI 뉴클레아제 도메인을 연결한 RFN (RNA-guided FokI Nuclease), 또는 dCas9에 전사활성인자 (transcription activator) 또는 억제자 도메인 (repressor domain)을 연결한 것일 수 있고, 상기 Cas9 니케이즈는 D10A Cas9 또는 H840A Cas9일 수 있으나, 이에 제한되는 것은 아니다.

[0047] 상기 Cas9 단백질은 그 유래에도 제한되지 않는다. 예컨대 상기 Cas9 단백질은 스트렙토코커스 피요제네스 (Streptococcus pyogenes), 프란시셀라 노비시다 (Francisella novicida), 스트렙토코커스 써모필러스 (Streptococcus thermophilus), 레지오넬라 뉴모필라 (Legionella pneumophila), 리스테리아 이노큐아 (Listeria innocua), 또는 스트렙토코커스 뮤탄스 (Streptococcus mutans) 유래일 수 있다.

[0048] 본원에서는 Cas9을 바이러스 벡터에서 발현시키기 위해 Cas9의 일부를 발현할 수 있는 벡터를 제작하였다. 즉, Cas9 단백질을 바이러스 벡터에서 패키징이 가능한 크기로 나누어 각각의 벡터에서 발현시키고자 하였다. 상기와 같은 방식으로 제작된 Cas9 단백질을 split-Cas9이라 하며, split-Cas9은 기존에 크기가 커서 바이러스 벡터 등을 통해 패키징 되지 않던 Cas9 단백질을 패키징 가능한 크기로 나누어서 이들 각각을 벡터를 통해 발현시키더라도 세포 내에서 그 기능을 잃지 않음을 특징으로 한다.

[0049] 본원에서 Cas9 단백질은 바람직하게는 dCas9, nCas9 및 SpCas9으로 이루어진 군으로부터 선택된 어느 하나일 수 있다. 일 구체예에서는 인테인(intein)-매개된 split-Cas9-기반의 ABE 및 CBE를 사용하였다.

[0050] 일 구체예에서, 상기 효율 예측 모델은

[0051] 염기교정 유전자가위의 활성 데이터를 정보 입력부를 통해 입력 받는 단계; 및

[0052] 상기 정보 입력부에서 입력 받은 데이터를 이용하여 컨볼루션 신경망(convolutional neural network: CNN)을 기반으로 한 딥러닝을 수행하여 효율 예측 모델을 생성하는 단계를 통해 생성되는 것일 수 있다.

[0053] 일 구체예에서, 상기 교정결과 예측 모델은

[0054] 염기교정 유전자가위의 교정결과 데이터를 입력 받는 정보 입력부를 통해 입력 받는 단계; 및

[0055] 상기 정보 입력부에서 입력 받은 데이터를 이용하여 컨볼루션 신경망(convolutional neural network: CNN)을 기반으로 한 딥러닝을 수행하여 교정결과 예측 모델을 생성하는 단계를 통해 생성되는 것일 수 있다.

[0056] 본원에서 용어, 염기교정 유전자가위의 "활성"은 염기교정 유전자가위에 의해 단일 염기가 교체되는 활성, 즉 표적 서열에서 RNA-가이드 뉴클레아제, 구체적으로 Cas9이 유전자를 절단하고, 탈아미노효소가 아데닌(A)을 구아닌(G)으로, 또는 시토신(C)을 티민(T)으로 전환하는 활성을 의미한다. 본원에서 용어, "활성 데이터"는 특정 표적 서열과 상기 염기교정 유전자가위의 관계를 추출 및 학습할 수 있는 데이터에 해당하며, 본 발명의 시스템은 상기 활성 데이터를 이용하여 효율 예측 모델을 생성할 수 있다.

[0057] 구체적으로, 상기 염기교정 유전자가위의 활성 데이터는 표적 서열의 염기를 서열 분석하여 얻을 수 있다. 예컨대, 딥 시퀀싱 (deep sequencing), 또는 RNAseq을 수행하여 이에 따른 데이터를 수득할 수 있으나, 편집된 염기의 검출을 통한 염기교정 유전자가위의 활성 데이터를 얻을 수 있다면, 특정 방법에 제한되지 않는다.

[0058] 염기교정 유전자가위가 표적 서열에서 나타내는 활성을 나타낼 수 있다면, 데이터의 형태, 종류, 크기 등은 제한되지 않는다.

[0059] 염기교정 유전자가위의 활성 데이터는 기존의 공지된 활성 데이터일 수도 있고, 또는, 당업자가 적절히 채택할 수 있는 임의의 방법으로 직접 수득한 활성 데이터일 수 있으며, 본 발명의 목적상, 염기교정 유전자가위의 활성을 예측할 수 있는 활성 예측 모델을 생성할 수 있는 데이터라면, 데이터가 수득되는 방법은 제한되지 않는다.

[0060] 일 구체예에 있어서, 상기 염기교정 유전자가위의 활성 데이터는 염기교정 유전자가위가 목적하는 표적 뉴클레오티드 주변의 서열 컨텍스트(context)가 고려된 것일 수 있다.

[0061] 본원에서 용어, "서열 컨텍스트(context)"란 염기교정 유전자가위가 목적하는 표적 뉴클레오티드 주변의 서열 정보를 의미한다.

[0062] 일 구체예에 있어서, 상기 염기교정 유전자가위의 활성 데이터는

[0063] 가이드 RNA를 코딩하는 염기서열 및 상기 가이드 RNA가 목적하는 표적 서열을 포함하는 올리고뉴클레오티드를 포함하는 세포 라이브러리에 염기교정 유전자가위를 도입하는 단계;

[0064] 상기 염기교정 유전자가위가 도입된 세포 라이브러리로부터 분리한 DNA를 이용하여 딥 시퀀싱을 수행하는 단계; 및

[0065] 상기 딥 시퀀싱으로부터 수득한 서열 데이터로부터 염기교정 범위 내 표적 뉴클레오티드 전환 여부를 검출하는 단계를 통해 수득할 수 있다.

[0066] 본원에서 용어, "염기교정 범위" 또는 "편집가능한 윈도우(editable window)"는 타겟 서열에서의 염기교정 유전자가위가 활성을 나타내는 염기교정 범위를 의미한다. 일 구체예에서, 상기 "염기교정 범위"는 가이드 RNA가 표적하는 프로토스페이서 내 20개의 위치 중 5'에서 3' 방향으로 위치 3 내지 10 bp 사이의 범위를 의미하며, 이를 넓은 편집가능한 윈도우라고 지칭한다. 일 구체예에서, 좁은 편집가능한 윈도우의 경우, 가이드 RNA가 표적하는 프로토스페이서 내 20개의 위치 중 5'에서 3' 방향으로 위치 4 내지 8 bp 사이의 범위를 의미한다.

[0067] 본원에서 용어, 염기교정 유전자가위의 "교정결과"또는 "편집 결과"는 타겟 서열에 대한 염기교정 유전자가위 활성의 결과로 만들어지는 편집 산물(product)을 의미한다. 한편, 염기교정 범위(편집가능한 윈도우) 내에 편집가능한 타겟 뉴클레오티드가 다수 개 존재하는 경우 원하지 않는 염기가 편집될 수 있으며, 본원에서 용어, "교정결과 빈도" 또는 "염기 편집 빈도" 또는 "편집 결과 빈도"는 염기교정 유전자가위의 활성의 결과 만들어지는 각 결과물의 빈도를 의미한다. 인간 점돌연변이 관련 질환 가운데 상당수는 염기교정 범위 내에 동일 염기가 다수 개 자리하고 있어, 염기교정 유전자가위를 안전하게 사용하기 위해서는 위치별 편집 빈도를 미리 예측하는 것이 중요하다.

[0068] 일 구체예에 있어서, 상기 염기교정 유전자가위의 교정결과 데이터는

[0069] 가이드 RNA를 코딩하는 염기서열 및 상기 가이드 RNA가 목적하는 표적 서열을 포함하는 올리고뉴클레오티드를 포함하는 세포 라이브러리에 염기교정 유전자가위를 도입하는 단계;

[0070] 상기 염기교정 유전자가위가 도입된 세포 라이브러리로부터 분리한 DNA를 이용하여 딥 시퀀싱을 수행하는 단계;

[0071] 상기 딥 시퀀싱으로부터 수득한 서열 데이터로부터 염기교정 범위 내 표적 뉴클리오티드 전환 빈도를 검출하는 단계를 통해 수득될 수 있다.

[0072] 상기 염기교정 유전자가위의 교정결과 데이터는 기존의 공지된 데이터일 수도 있고, 또는, 당업자가 적절히 채택할 수 있는 임의의 방법으로 직접 수득한 활성 데이터일 수 있으며, 본 발명의 목적상, 염기교정 유전자가위의 교정결과를 예측할 수 있는 교정결과 예측 모델을 생성할 수 있는 데이터라면, 데이터가 수득되는 방법은 제한되지 않는다.

[0073] 본원에서 효율 예측 모델 및 교정결과 예측 모델 각각은, 염기교정 유전자가위의 활성 데이터 또는 교정결과 데이터가 저장된 공지된 데이터베이스를 이용하고, 상기 데이터베이스로부터 입력 받은 대규모 데이터를 이용하여 딥러닝 기술을 통해 생성될 수 있다. 즉, 염기교정 유전자가위의 활성 데이터 또는 교정결과 데이터는 직접 측정하여 수득한 것 외에, 공지된 데이터베이스, 문헌 등에서 수득하거나, 상기 데이터베이스 또는 문헌으로부터 수득한 데이터를 2차로 가공하여 수득할 수 있으며, 표적 서열과 염기편집 효율 또는 염기편집 빈도 간의 관계를 추출하고, 상기 추출된 특징을 조합하여 임의의 표적 서열에 대한 염기교정 유전자가위의 효율 및 편집결과를 예측할 수 있는 데이터라면 제한 없이 사용할 수 있다.

[0074] 본원에서 용어, "올리고뉴클레오티드 (oligonucleotide)"는 수 내지 수백 개의 뉴클레오티드가 포스포다이에스터 결합으로 연결된 물질을 말하며, 본 발명의 목적상 상기 올리고뉴클레오티드는 이중나선 DNA일 수 있다. 본원에서 사용되는 상기 올리고뉴클레오티드는 상기 올리고뉴클레오티드는 10 내지 300 bp, 바람직하게는 50 내지 200 bp, 보다 바람직하게는 100 내지 180 bp의 길이를 가질 수 있으나, 이에 제한되는 것은 아니고, 분석 목적 등에 따라 당업자에 의해 적절히 조절될 수 있다.

[0075] 본원에서 상기 올리고뉴클레오티드는 가이드 RNA 코딩 염기서열 및 표적 서열을 포함한다. 또한, 상기 올리고뉴클레오티드는 PCR 증폭될 수 있도록 프라이머가 결합될 수 있는 추가의 서열을 포함할 수 있다.

[0076] 상기 표적 서열은 10 내지 100 bp, 바람직하게는 20 내지 50 bp, 보다 더 바람직하게는 20 내지 30 bp, 가장 바

람직하게는 24 내지 26 bp의 길이를 가질 수 있으나, 특별히 이에 제한되는 것은 아니다.

[0077] 또한, 상기 가이드 RNA 코딩 서열은 10 내지 100 bp, 바람직하게는 15 내지 50 bp, 보다 바람직하게는 20 내지 30 bp의 길이를 가질 수 있으나, 특별히 이에 제한되는 것은 아니다.

[0078] 상기 올리고뉴클레오티드는 바코드 서열을 더 포함할 수 있다.

[0079] 상기 바코드 서열은 각 올리고뉴클레오티드를 식별하도록 하기 위한 뉴클레오티드 서열을 의미한다. 본원에서 상기 바코드 서열은 2 이상의 반복 뉴클레오티드 (AA, TT, CC, GG)를 포함하지 않는 것일 수 있으나, 각 올리고뉴클레오티드를 식별하도록 설계된 것이라면 특별히 이에 제한되는 것은 아니다. 복수의 올리고뉴클레오티드들에 있어, 상기 바코드 서열은 각 올리고뉴클레오티드가 식별될 수 있도록 적어도 2 개의 염기가 다르도록 설계된 것일 수 있다. 상기 바코드 서열은 5 내지 50 bp의 길이를 가질 수 있으나, 특별히 이에 제한되지 않는다.

[0080] 상기 올리고뉴클레오티드는 세포에 도입되어 염색체 내에 통합 (integration)되는 것일 수 있다.

[0081] 본원에서 용어, "라이브러리"는 특성이 다른 동종의 물질이 2종 이상 포함된 집단(pool or population)을 의미한다. 따라서, 올리고뉴클레오티드 라이브러리는 염기서열이 다른 2종 이상의 올리고뉴클레오티드, 예컨대 가이드 RNA, 및/또는 표적 서열이 다른 2종의 올리고뉴클레오티드를 포함하는 집단일 수 있고, 세포 라이브러리는 특성이 다른 2종 이상의 세포, 구체적으로 본 발명의 목적상 각각의 세포가 포함하는 올리고뉴클레오티드가 다른, 예컨대 도입된 가이드 RNA, 및/또는 표적 서열, 또는 종류가 다른 세포들의 집단일 수 있다.

[0082] 본원에서 용어, "벡터"는 상기 올리고뉴클레오티드를 세포 내에 전달할 수 있도록 하는 매개체, 예컨대 유전적 작제물을 의미하는 것으로, 본원에서 벡터는 각각의 가이드 RNA 코딩 염기서열 및 표적 염기서열을 포함하는 올리고뉴클레오티드를 포함할 수 있다. 상기 벡터는 바이러스 벡터 또는 플라스미드 벡터일 수 있고, 바이러스 벡터는 구체적으로 렌티 바이러스 벡터 또는 레트로바이러스 벡터 등이 사용될 수 있으나, 이에 제한되는 것은 아니고 당업자는 본 발명의 목적을 달성할 수 있는 한 공지된 벡터를 자유롭게 사용할 수 있다.

[0083] 구체적으로, 상기 벡터는 개체의 세포 내에 존재하는 경우 삽입물, 즉 올리고뉴클레오티드가 발현될 수 있도록 삽입물에 작동가능하게 연결된 필수적인 조절 요소를 포함할 수 있다.

[0084] 상기 벡터는 표준적인 재조합 DNA 기술을 이용하여 제조 및 정제될 수 있다. 상기 벡터의 종류는 원핵세포 및 진핵세포 등 목적하는 세포에서 작용할 수 있도록 하는 한, 특별히 한정되지 않는다. 벡터는 프로모터, 개시코돈, 및 종결코돈 터미네이터를 포함할 수 있다. 그 외에 시그널 펩타이드를 코드하는 DNA, 및/또는 인핸서 서열, 및/또는 원하는 유전자의 5'측 및 3'측의 비번역 영역, 및/또는 선택마커 영역, 및/또는 복제가능단위 등을 적절하게 포함할 수도 있다.

[0085] 상기 벡터를 라이브러리를 제조하기 위한 세포에 전달하는 방법은 당업계에 공지된 다양한 방법을 이용하여 달성될 수 있다. 예컨대, 칼슘 포스페이트-DNA 공침전법, DEAE-덱스트란-매개 트랜스펙션법, 폴리브렌-매개 형질감염법, 전기충격법, 미세주사법, 리포좀 융합법, 리포펙타민 및 원형질체 융합법 등의 당 분야에 공지된 여러 방법에 의해 수행될 수 있다. 또한, 바이러스 벡터를 이용하는 경우, 감염(infection)을 수단으로 하여 바이러스 입자를 사용하여 목적물, 즉 벡터를 세포 내로 전달시킬 수 있다. 아울러, 유전자 밤바드먼트 등에 의해 벡터를 세포 내로 도입할 수 있다.

[0086] 상기 도입된 벡터는 세포 내에서 벡터 자체로 존재하거나, 염색체 내에 통합될 수 있으나, 특별히 이에 제한되는 것은 아니다.

[0087] 본원에서 제조된 세포 라이브러리는 가이드 RNA-표적 서열을 포함하는 올리고뉴클레오티드가 도입된 세포 집단을 말한다. 이때 각각의 세포들은 벡터, 구체적으로 바이러스의 종류 및/또는 수가 다르게 도입된 것일 수 있다.

[0088] 상기 벡터가 도입될 수 있는 세포의 종류는, 벡터의 종류 및/또는 목적하는 세포의 종류에 따라 적절하게 당업자가 선택할 수 있으나, 그 예로, 대장균, 스트렙토미세스, 살모넬라 티피뮤리움 등의 박테리아 세포; 효모 세포; 피치아 파스토리스 등의 균류세포; 드로조필라, 스포도프테라 Sf9 세포 등의 곤충 세포; CHO(중국 햄스터 난소 세포, chinese hamster ovary cells), SP2/0(마우스 골수종), 인간 림프아구(human lymphoblastoid), COS, NSO(마우스 골수종), 293T, 보우 멜라노마 세포, HT-1080, BHK(베이비 햄스터 신장세포, baby hamster kidney cells), HEK(인간 배아신장 세포, human embryonic kidney cells), PERC.6(인간망막세포) 등의 동물 세포; 또는 식물 세포가 될 수 있다.

[0089] 본 발명의 용어, "정보 입력부"는 상술한 염기교정 유전자가위의 활성 데이터 또는 교정결과 데이터를 입력 받는 구성 요소로서, 상기 정보 입력부는 일 구체예에 따른 예측 시스템의 사용자로부터 직접 염기교정 유전자가위에 관한 데이터를 입력 받거나, 또는 미리 저장된 데이터를 입력 받는 것일 수 있으나, 이에 제한되지 않는다.

[0090] 본 발명의 시스템은 미리 수득한 염기교정 유전자가위에 관한 데이터 또는 공지된 염기교정 유전자가위에 관한 데이터가 저장된 저장부를 추가로 포함할 수 있으나, 이에 제한되지 않는다. 상기 저장부를 포함할 경우, 본 발명 시스템의 정보 입력부는 상기 저장부로부터 설정된 크기 또는 범위의 데이터를 입력 받아, 염기교정 유전자가위의 활성 또는 교정결과를 예측하는데 이용할 수 있다.

[0091] 일 구체예에서, "효율 예측 모델" 및 "교정결과 예측 모델"은 상기 정보 입력부를 통해 입력된 염기교정 유전자가위에 관한 데이터를 이용하여, 표적 서열 및 염기 편집 결과, 상기 염기 편집 결과의 빈도 간의 관계를 추출하고 조합하여, 표적 서열과 염기교정 유전자가위 간의 관계를 학습할 수 있는 예측 모델을 생성하는 단계를 통해 생성될 수 있다. 상기 효율 예측 모델 및 교정결과 예측 모델은 학습된 정보를 기반으로 딥러닝 기술을 이용하여 생성되고, 일 구체예에 따른 예측 시스템의 사용자는 상기 예측 모델을 통해 염기교정 유전자가위의 효율 및 교정결과를 예측할 수 있다.

[0092] 구체적으로, 본원의 예측 모델은 컨볼루션 신경망 (convolutional neural network, CNN)을 기반으로 하여 표적 서열 및 염기 편집 결과, 상기 염기 편집 결과의 빈도 간의 관계를 학습하는 딥-러닝을 수행하는 것일 수 있으나, 이에 제한되지 않는다.

[0093] 본원에서 용어, "딥러닝 (Deep Learning)"은 컴퓨터가 사람처럼 생각하고 배울 수 있도록 하는 인공지능(AI) 기술로서, 인공신경망 이론을 기반으로 복잡한 비선형 문제를 기계가 스스로 학습해결 할 수 있도록 하는 기술이다. 상기 딥러닝 기술을 이용하여, 사람이 모든 판단 기준을 정해주지 않아도 컴퓨터가 스스로 인지, 추론, · 판단할 수 있게 되고, 음성 이미지 인식과 사진 분석 등에 광범위하게 활용하는 것이 가능하다.

[0094] 즉, 딥러닝(deep learning)은 여러 비선형 변환기법의 조합을 통해 높은 수준의 추상화(abstractions, 다량의 데이터나 복잡한 자료들 속에서 핵심적인 내용 또는 기능을 요약하는 작업)를 시도하는 기계학습(machine learning) 알고리즘의 집합으로 정의될 수 있다.

[0095] 본원에서 용어, "컨볼루션 신경망(convolutional neural networks: CNN)"은 제공된 정보의 일부를 표현하는 특징(feature)을 추출하고, 정보의 계층화를 통해 일반화를 이루어 내는 기술을 의미한다.

[0096] 본 발명자들은, 유전자가위 활성 대량측정법을 이용하여 대량의 유전자가위의 효율 및 교정결과 데이터를 생산하고, 강력한 컨볼루션 신경망(convolutional neural networks: CNNs)을 사용하는 딥러닝 프레임워크를 기반으로 실제 실험 결과 값과 인공지능이 제시한 예측 값의 상관관계가 0.69~0.79에 수렴하는 높은 신뢰도를 보이는 활성 예측 모델 및 그 상관관계가 0.91~0.93에 도달하는 높은 신뢰도를 보이는 교정결과 예측 모델을 개발하고, 상기 두 모델을 결합하여 DeepABE 및 DeepCBE로 명명되는 예측 시스템을 개발하였다. 나아가, 생물학적 복제시료 및 인간 유도만능줄기 세포에서 그 정확성을 검증하였다.

[0097] 본원에서 용어, "결과 예측부"는 상술한 방법으로 구축된 효율 예측 모델 및 교정결과 예측 모델에 표적 서열 입력부를 통해 입력된 표적 서열을 적용하여, 염기교정 유전자가위의 염기교정 효율 및 결과를 예측하는 구성이다. 일 구체예에서, 결과 예측부는 표적 서열 정보로부터 염기교정 유전자가위의 염기교정 효율 및 결과를 예측할 수 있으나, 예측의 정확성을 높이기 위한 요인, 예를 들어, 표적 뉴클레오티드 주변의 서열 컨텍스트(context) 또는 염색질 접근성을 추가로 고려할 수 있다.

[0098] 구체적으로, 상기 결과 예측부는 미리 설정된 방법에 의해 염기교정 유전자가위에 의한 표적 서열의 염기 편집 여부 또는 염기 편집 빈도를 예측하는 것일 수 있으나, 이에 제한되지 않는다. 상기 결과 예측부는 염기 편집 여부 또는 염기 편집 빈도 외에도 다른 염기교정 유전자가위의 활성을 예측할 수 있는 지표라면, 그 종류나 형태, 예측 방법에 관계없이 염기교정 유전자가위의 활성을 예측하기 위해 이용할 수 있다.

[0099] 일 구체예에서, 상기 결과 예측부는 표적 서열 입력부에서 입력 받은 표적 서열을 효율 예측 모델 및 교정결과 예측 모델에 각각 적용하여 염기교정 유전자가위의 효율 및 교정결과 스코어를 획득하고, 상기 효율 스코어와 교정결과 스코어를 곱하여 염기교정 유전자가위의 염기교정 효율 및 결과를 예측할 수 있다.

[0100] 상기 효율 스코어는 표적 서열의 각 위치에 대하여 하기 [수학식 1]을 이용하여 산출될 수 있다.

[0101] [수학식 1]

$$\text{염기 편집 효율 (\%)} = \frac{\text{넓은염기교정 범위(표적 서열의 위치3 내지10)에서 의도된 표적 뉴클레오타드 전환을 포함하는 모든 서열의 총 리드(read)}}{\text{총 리드(read)}} \times 100$$

[0102]

[0103] 또한, 상기 교정결과 스코어는 하기 [수학식 2]을 이용하여 산출될 수 있다.

[0104] [수학식 2]

$$\text{염기 편집결과 빈도} = \frac{\text{특정 염기 - 편집된 결과 서열의 리드(read)}}{\text{넓은염기교정 범위(표적 서열의 위치3 내지10)에서의도된 표적 뉴클레오타드 전환을 포함하는 모든 서열의 총 리드(read)}}$$

[0105]

[0106] 일 양상에 따른 예측 시스템은 표적 뉴클레오티드 주변의 서열 컨텍스트 또는 표적 서열의 염색질 접근성 정보를 이용하여 상기 결과 예측부에서 예측된 염기교정 유전자가위의 활성을 최적화(fine-tuning)하는 미세 조정부를 추가로 포함할 수 있다.

[0107] 본 발명의 "미세 조정부"는 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템의 정확성을 높이기 위해, 입력된 표적 서열의 서열 정보뿐만 아니라, 표적 뉴클레오티드 주변의 서열 컨텍스트 또는 염기교정 유전자가위의 표적 서열에 대한 염색질 접근성까지 고려하여 효율 예측 모델에서 예측된 염기교정 유전자가위의 활성을 최적화하는 구성을 의미한다.

[0108] 상기 염색질 접근성 정보는 공지된 데이터 베이스, 문헌 등에서 수득하거나, 또는 직접 측정할 수 있으며, 구체적으로 타겟 서열의 DNase I에 대한 민감성으로부터 계산되는 것일 수 있으나, 이에 제한되는 것은 아니다.

[0109] 일 구체예에서, 상기 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템은 결과 예측부에서 예측된 염기교정 유전자가위의 효율 및 결과를 출력하는 출력부를 추가로 포함할 수 있다.

[0110] 상기 출력부가 출력하는 염기교정 유전자가위의 염기교정 효율 및 결과에 대한 정보는 출력되는 신호의 형태나 종류는 제한되지 않는다.

[0111] 다른 양상은 염기교정 유전자가위의 표적 서열을 설계하는 단계; 및

[0112] 상기 설계된 표적 서열을 일 양상에 따른 염기교정 효율 및 결과 예측 시스템에 적용하는 단계를 포함하는 염기교정 유전자가위의 염기교정 효율 및 결과 예측 방법을 제공한다.

[0113] 일 양상에 따른 방법에 따르면, 실제 실험 결과값과 예측 모델이 제시한 예측 값의 상관관계가 0.50 내지 0.95에 수렴하는 예측 모델을 통해 염기 편집 효율 및 편집 결과의 빈도를 예측하여, 안전한 교정이 가능한 유전자가위를 선별하고, 상기 유전자가위로 질환 모델을 만들거나 교정할 수 있는 질환 정보를 제공할 수 있다.

[0114] 다른 양상은

[0115] 인간 점돌연변이 데이터를 수득하는 단계;

[0116] 상기 인간 점돌연변이 데이터로부터 점돌연변이가 정상 염기 아데닌(A)이 비정상 염기 구아닌(G)으로 바뀌어 발생하는 경우; 정상 염기 구아닌(G)이 비정상 염기 아데닌(A)으로 바뀌어 발생하는 경우; 정상 염기 시토신(C)이 비정상 염기 티민T)으로 바뀌어 발생하는 경우; 또는 정상 염기 티민(T)이 비정상 염기 시토신(C)으로 바뀌어 발생하는 경우에 해당하는 데이터를 1차로 선별하는 단계;

[0117] 상기 1차로 선별된 데이터 중에서 점돌연변이가 프로토스페이서 영역의 5' 말단으로부터 3 내지 10 bp 위치에 존재하는 데이터를 2차로 선별하는 단계;

[0118] 상기 2차로 선별된 데이터 중에서 병원성 또는 유사병원성 점돌연변이에 해당하는 데이터를 3차로 선별하는 단계; 및

[0119] 상기 3차로 선별된 데이터를 일 양상에 따른 염기교정 효율 및 결과 예측 시스템에 적용하는 단계를 포함하는 염기교정 유전자가위를 사용할 수 있는 인간 점돌연변이 관련 질환에 대한 정보를 제공하는 방법을 제공한다.

[0120] 상기 인간 점돌연변이 관련 질환은 점돌연변이가 프로토스페이서의 5' 말단으로부터 3 내지 10 bp 범위 내에 존

재하고(본원에서, "염기교정 범위"), 및/또는 프로토스페이서의 하류(downstream)에 PAM 서열이 있는 경우에 있어서, 상기 염기교정 범위 내 정상 염기(A 또는 C)가 비정상 염기(G 또는 T)로 바뀌어서 발생하거나; 또는 상기 염기교정범위 내 정상 염기(G 또는 T)가 비정상 염기(A 또는 C)로 바뀌어서 발생하는 질환이면 제한없이 포함될 수 있다.

[0121] 일 구체예에 있어서, 상기 인간 점돌연변이 관련 질환은 어셔 증후군(Usher syndrome), 종양괴사인자 수용체 관련 주기적 증후군(TNF receptor-associated periodic syndrome: TRAPS), 마판 증후군(marfan syndrome), 제3형 청년기 발병 당뇨병(Type 3 form of Maturity-Onset Diabetes of the Young: MODY3), 선천성 비진행성 야맹증(Congenital stationary night blindness type 1F), 가족성 고콜레스테롤혈증(Familial hypercholesterolemia), 선천근육무력증후군(congenital myasthenic syndrome: CMS), 린치증후군(Lynch syndrome) 등이 확인되었고, CBE의 경우 로이-디에츠 증후군(Loeys-Dietz syndrome: LDS), 망막색소변성증(retinitis pigmentosa), 렙틴 결핍 또는 장애(Leptin deficiency 또는 dysfunction), 가족성 고콜레스테롤혈증(Familial hypercholesterolemia), 상염색체 열성 청각장애(autosomal recessive deafness), 콜레스테롤 모노옥시다제 결핍(cholesterol monooxygenase (side-chain-cleaving) deficiency) 및 진행성 근간대성간질(progressive myoclonus epilepsy)로 이루어진 군으로부터 선택되는 어느 하나일 수 있으나, 이에 제한되는 것은 아니다.

[0122] 다른 양상은 상기 방법을 컴퓨터로 실행하기 위한 프로그램이 기록된 컴퓨터 판독가능 기록매체를 제공한다.

[0123] 상기 프로그램은 일 양상의 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템 또는 염기교정 유전자가위의 염기교정 효율 및 결과 예측 방법을 컴퓨터 프로그래밍 언어로 구현한 것일 수 있으며, 염기교정 유전자가위의 염기교정 효율 및 결과를 예측하는데 이용될 수 있다.

[0124] 본 발명의 프로그램을 구현할 수 있는 컴퓨터 프로그래밍 언어는 Python, C, C++, 자바(Java), 포트란(Fortran), 비쥬얼 베이직(Visual Basic) 등이 있으나 이에 제한되지 않는다. 상기 프로그램은 USB 메모리, CDROM(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 핵산서열을 검색할 수 있다.

[0125] 상기 프로그램은 온라인 또는 오프라인으로 제공될 수 있으며, 컴퓨터로 구현되는 전자기기와 결합되어 염기교정 유전자가위의 염기교정 효율 및 결과 예측 시스템을 실행시키기 위해 기록매체에 저장된 컴퓨터 프로그램의 형태로 제공될 수 있다.

[0126] 다른 양상은 염기교정 유전자가위를 세포에 도입하는 단계를 포함하는 세포의 유전체에서 표적 뉴클레오티드를 편집하는 방법으로서,

[0127] 상기 염기교정 유전자가위는 (i) RNA-가이드 뉴클레아제 또는 이를 코딩하는 유전자, (ii) 탈아미노효소 또는 이를 코딩하는 유전자, 및 (iii) 표적 서열과 혼성화 할 수 있는 가이드 RNA 또는 이를 코딩하는 유전자를 포함하고,

[0128] 상기 표적 서열은 PAM 서열, 프로토스페이서 서열, 및 가이드 RNA에 상보적인 서열을 포함하고,

[0129] 상기 가이드 RNA에 상보적인 서열은 5'-TAC-3', 5'-TAT-3', 5'-TAG-3', 5'-GAT-3', 5'-CAC-3', 5'-GAC-3', 5'-CAT-3', 5'-TAA-3', 5'-CAG-3', 5'-GAG-3', 5'-AAC-3', 5'-CAA-3', 5'-GAA-3', 5'-AAT-3', 5'-AAG-3', 5'-AAA-3', 5'-TCC-3', 5'-TCG-3', 5'-TCT-3', 5'-TCA-3', 5'-CCC-3', 5'-CCT-3', 5'-ACC-3', 5'-CCA-3', 5'-CCG-3', 5'-ACG-3', 5'-ACT-3', 5'-ACA-3', 5'-GCC-3', 5'-GCT-3', 5'-GCG-3', 및 5'-GCA-3'으로 이루어진 군으로부터 선택되는 서열을 포함하고,

[0130] 상기 탈아미노효소는 표적 서열에서 아데닌 또는 시토신을 탈아미노화하는 것을 특징으로 하는 세포의 유전체에서 표적 뉴클레오티드를 편집하는 방법을 제공한다.

[0131] 일 구체예에서, 상기 RNA-가이드 뉴클레아제는 SpCas9, nCas9, 및 dCas9로 이루어진 군으로부터 선택되는 것일 수 있다.

[0132] 일 구체예에서, 상기 표적 뉴클레오티드는 프로토스페이서 영역의 5' 말단으로부터 3 내지 10 bp 위치에 존재하는 것일 수 있다.

[0133] 다른 양상은 염기교정 유전자가위를 포함하는 인간 점돌연변이 관련 질환의 예방 또는 치료용 약학적 조성물로서,

[0134] 상기 염기교정 유전자가위는 (i) RNA-가이드 뉴클레아제 또는 이를 코딩하는 유전자, (ii) 탈아미노효소 또는 이를 코딩하는 유전자, 및 (iii) 표적 서열과 혼성화 할 수 있는 가이드 RNA 또는 이를 코딩하는 유전자를 포함하고,

[0135] 상기 표적 서열은 PAM 서열, 프로토스페이서 서열, 및 가이드 RNA에 상보적인 서열을 포함하고,

[0136] 상기 가이드 RNA에 상보적인 서열은 5'-TAC-3', 5'-TAT-3', 5'-TAG-3', 5'-GAT-3', 5'-CAC-3', 5'-GAC-3', 5'-CAT-3', 5'-TAA-3', 5'-CAG-3', 5'-GAG-3', 5'-AAC-3', 5'-CAA-3', 5'-GAA-3', 5'-AAT-3', 5'-AAG-3', 5'-AAA-3', 5'-TCC-3', 5'-TCG-3', 5'-TCT-3', 5'-TCA-3', 5'-CCC-3', 5'-CCT-3', 5'-ACC-3', 5'-CCA-3', 5'-CCG-3', 5'-ACG-3', 5'-ACT-3', 5'-ACA-3', 5'-GCC-3', 5'-GCT-3', 5'-GCG-3', 및 5'-GCA-3'으로 이루어진 군으로부터 선택되는 서열을 포함하고,

[0137] 상기 탈아미노효소는 표적 서열에서 아데닌 또는 시토신을 탈아미노화하는 것을 특징으로 하는 인간 점돌연변이 관련 질환의 예방 또는 치료용 약학적 조성물을 제공한다.

발명의 효과

[0138] 일 양상에 따른 예측 시스템을 사용하면, 유전자가위를 일일이 제작하여 검증할 필요 없이 간단한 방법으로 효율 및 정확성의 예측이 가능하여 안전한 교정이 가능한 유전자가위를 선별할 수 있다. 나아가, 병원성/유사병원성 인간 점돌연변이 질환 중 염기교정 유전자가위로 질환을 만들거나 교정할 수 있는 경우들의 효율 및 결과 빈도의 예측이 가능하여 염기교정 유전자가위의 대상 질환을 선별할 수 있다.

도면의 간단한 설명

[0139] 도 1은 대규모 활성 데이터에 기반하여 아데닌 및 시토신 염기교정 유전자가위의 특성을 분석한 도이다.

(a, b) 프로토스페이서(protospacer) 내 편집가능한 (a) 아데닌 또는 (b) 시토신의 위치와 통합된 표적 서열(integrated target sequences)에서 고-처리량 방식으로 측정된 염기 편집 빈도 사이의 관계. 통합된 표적 서열 내 프로토스페이서 영역의 위치 1 내지 20에서 측정된 염기 편집 빈도를 나타내었다. 위치 20은 PAM(NGG)의 바로 상류에 자리한다. 분석된 표적 서열의 수(n)는 다음과 같다: (a) ABE의 경우 n = 2,427 내지 2,898; (b) CBE의 경우 n = 2,847 내지 3,858. 박스에서 상단, 중앙 및 하단선은 각각 25 번째, 50 번째 및 75 번째 백분위수를 나타낸다. 수염(whiskers)은 각각 1 번째, 99 번째 백분위수 값을 나타낸다.

(c, d) 동일한 통합된 표적 서열에 대해 SpCas9-유도된 인델(indel) 빈도 및 (c) ABE- 또는 (d) CBE- 유도된 염기 전환 효율이 결정되었다. 염기 편집 효율에 대한 프로토스페이서 영역 내 염기 위치의 효과를 배제하기 위해, 동일한 프로토스페이서 영역 내 위치에 해당하는 염기를 갖는 표적 서열만을 비교하였다. 히트(heat) 색상은 육각형 빈(bin) 내 표적 서열의 수를 나타낸다. 분석된 표적 서열의 수(n)는 다음과 같다: (c) ABE의 경우 n = 2,172 내지 2,307; (d) CBE에 대해 n = 2,746 내지 2,964. 스피어만(Spearman) 상관계수(R) 및 피어슨(Pearson) 상관계수(r)를 표시하였다.

(e, f) 프로토스페이서 영역의 4' 말단으로부터 4 내지 8 bp 위치에서 (e) ABE- 및 (f) CBE- 지정된 염기 편집 빈도에 대한 표적 염기(빨간색) 주위의 서열 컨텍스트(context)의 효과. 표적 염기 전환 빈도는 각 위치에서 가장 높은 중앙값 편집 빈도를 나타내는 서열 모티프의 중앙값 빈도로 정규화되어, 상대 빈도(relative frequency)가 산출되었다. 분석된 표적 모티프의 수(n)는 다음과 같다: (e) ABE의 경우 n = 383 내지 1,413; (f) CBE의 경우 n = 498 내지 1,110. 박스에서 상단, 중앙 및 하단선은 각각 25 번째, 50 번째 및 75 번째 백분위수를 나타낸다. 수염은 각각 1 번째, 99 번째 백분위수 값을 나타낸다. 표적 염기 전환 빈도에서 통계적으로 유의미한 차이가 없는 컨텍스트 서열의 부분집합(subset)은 a, b, c, ... 및 h와 같은 문자로 표시하였다 ($P<0.05$; one-way ANOVA 및 Tukey의 사후 검정에 의해 결정됨).

도 2는 각 위치에서 염기 편집에 대한 표적 아데닌(좌측) 또는 시토신(우측)주위의 서열 컨텍스트의 효과를 나타낸 도이다. 표적 염기는 빨간색으로 표시하였고, 상대적 염기 편집 빈도는 각 위치에서 가장 높은 중앙값 편집 빈도를 나타내는 서열 모티프의 중앙값 빈도로 정규화되었다. 분석된 표적 서열의 수(n)는 다음과 같다:

위치 4에서의 ABE: n = 159 (TAC), 106 (TAT), 99 (TAG), 126 (GAT), 143 (GAC), 176 (CAT), 219 (CAC), n =

78 (TAA), n = 169 (GAG), n = 267 (CAG), n = 146 (AAC), n = 144 (GAA), n = 158 (CAA), n = 123 (AAT); n = 166 (AAG), n = 154 (AAA);

위치 5에서의 ABE: n = 153 (TAC), 97 (TAT), 70 (TAG), 155 (GAT), 161 (GAC), 205 (CAT), 252 (CAC), 2 (TAA), 165 (GAG), 268 (CAG), 140 (AAC), 181 (GAA), 189 (CAA), 123 (AAT), 129 (AAG), 142 (AAA);

위치 6에서의 ABE: n = 186 (TAC), 103 (TAT), 117 (TAG), 163 (GAT), 188 (GAC), 168 (CAT), 226 (CAC), 80 (TAA), 194 (GAG), 306 (CAG), 135 (AAC), 165 (GAA), 117 (CAA), 130 (AAT), 176 (AAG), 143 (AAA);

위치 7에서의 ABE: n = 168 (TAC), 108 (TAT), 119 (TAG), 127 (GAT), 155 (GAC), 169 (CAT), 235 (CAC), 78 (TAA), 174 (GAG), 289 (CAG), 125 (AAC), 178 (GAA), 134 (CAA), 97 (AAT), 162 (AAG), 121 (AAA);

위치 8에서의 ABE: n = 170 (TAC), 105 (TAT), 76 (TAG), 148 (GAT), 169 (GAC), 221 (CAT), 240 (CAC), (TAA), 165 (GAG), 282 (CAG), 118 (AAC), 194 (GAA), 170 (CAA), 125 (AAT), 134 (AAG), 134 (AAA);

위치 4에서의 CBE: n = 186 (TCC), 101 (TCG), 176 (TCT), 211 (TCA), 230 (CCC), 219 (CCT), 183 (ACC), 243 (CCA), 183 (CCG), 115 (ACG), 153 (ACT), 177 (ACA), 203 (GCC), 220 (GCT), 170 (GCG), 194 (GCA);

위치 5에서의 CBE: n = 173 (TCC), 93 (TCG), 173 (TCT), 173 (TCA), 182 (CCC), 231 (CCT), 181 (ACC), 219 (CCA), 170 (CCG), 91 (ACG), 178 (ACT), 157 (ACA), 222 (GCC), 223 (GCT), 148 (GCG), 84 (GCA);

위치 6에서의 CBE: n = 198 (TCC), 108 (TCG), 157 (TCT), 174 (TCA), 192 (CCC), 195 (CCT), 193 (ACC), 195 (CCA), 176 (CCG), 104 (ACG), 152 (ACT), 161 (ACA), 228 (GCC), 159 (GCT), 154 (GCG), 200 (GCA);

위치 7에서의 CBE: n = 195 (TCC), 125 (TCG), 171 (TCT), 232 (TCA), 180 (CCC), 208 (CCT), 198 (ACC), 248 (CCA), 175 (CCG), 118 (ACG), 165 (ACT), 165 (ACA), 213 (GCC), 213 (GCT), 158 (GCG), 173 (GCA);

위치 8에서의 CBE: n = 163 (TCC), 93 (TCG), 152 (TCT), 199 (TCA), 177 (CCC), 233 (CCT), 211 (ACC), 205 (CCA), 171 (CCG), 70 (ACG), 126 (ACT), 193 (ACA), 227 (GCC), 225 (GCT), 172 (GCG), 200 (GCA).

박스에서 상단, 중앙 및 하단선은 각각 25 번째, 50 번째 및 75 번째 백분위수를 나타낸다. 수염은 각각 1 번째, 99 번째 백분위수 값을 나타낸다. 염기 편집 및 인델 빈도에서 통계적으로 유의미한 차이가 없는 컨텍스트 서열의 부분집합(subset)은 a, b, c, ... 및 j와 같은 문자로 표시하였다(P<0.05; one-way ANOVA 및 Tukey의 사후 검정에 의해 결정됨).

도 3은 프로토스페이서 영역의 5' 말단으로부터 4 내지 8 bp 위치에서 SpCas9-유도된 인델 빈도에 대한 표적 염기(빨간색) 주위의 서열 컨텍스트의 효과를 나타낸 도이다. (a) 및 (b)의 표적 서열은 각각 도 1의 (e, f)와 동일하다. 인델 빈도는 각 위치에서 가장 높은 중앙값 편집 빈도를 나타내는 서열 모티프의 중앙값 빈도로 정규화되었다. 3개 뉴클레오티드 모티프 당 분석된 표적 서열의 수(n)는 다음과 같다:

(a) n = 807 (TAC), 478 (TAT), 464 (TAG), 685 (GAT), 1,142 (GAC), 787 (CAT), 903 (CAC), 350 (TAA), 832 (GAG), 1,379 (CAG), 638 (AAC), 823 (GAA), 738 (CAA), 558 (AAT), 735 (AAG), 620 (AAA);

(b) n = 915 (TCC), 520 (TCG), 829 (TCT), 989 (TCA), 961 (CCC), 1,086 (CCT), 966 (ACC), 1,110 (CCA), 875 (CCG), 498 (ACG), 774 (ACT), 853 (ACA), 1,093 (GCC), 1,040 (GCT), 802 (GCG), 951 (GCA).

박스에서 상단, 중앙 및 하단선은 각각 25 번째, 50 번째 및 75 번째 백분위수를 나타낸다. 수염은 각각 1 번째, 99 번째 백분위수 값을 나타낸다. 염기 편집 및 인델 빈도에서 통계적으로 유의미한 차이가 없는 컨텍스트 서열의 부분집합(subset)은 a, b, c, ... 및 f와 같은 문자로 표시하였다(P<0.05; one-way ANOVA 및 Tukey의 사후 검정에 의해 결정됨).

도 4는 염기교정 유전자가위의 염기교정 효율 및 가능한 교정결과들의 빈도를 예측하는 예측 모델의 개략도이다. 편집 가능한 윈도우(볼드 및 밑줄) 내의 세 개의 아데닌(빨간색으로 표시)을 예시로서 나타내었고, ABE-편집된 결과에 대한 계산상 예측 및 실험적으로 측정된 빈도를 나타내었다. 프로토스페이서 인접 모티프(protospacer adjacent motif: PAM)은 파란색으로 표시하였고 염기 편집된 뉴클레오티드는 소문자로 표시하였다.

도 5는 ABE_efficiency, CBE_efficiency, ABE_proportion 및 CBE_proportion모델의 개발에 있어서, 교차검증을 사용하여 히든 레이어(hidden layer)의 수와 인풋 서열의 길이를 결정한 과정을 나타낸 도이다. 히트(heat) 맵은 10배 교차 검증(n=10)의 평균 (a) Spearman 상관 계수, (b) 쿨백-라이블러 발산 값을 나타낸 것이다.

도 6은 주어진 표적 서열에서 aBe- 및 CBe-유도된 염기 전환의 효율 및 결과를 예측하는 예측 모델의 개발 및 평가를 나타낸 도이다.

(a) 통합(integrated) 및 내인성(endogenous) 부위에서 ABE_efficiency, ABE_proportion 및 DeepABE의 성능 평가. 분석된 표적 서열의 수, 교정결과 빈도 및 각 결과의 효율(n)은 각각 다음과 같다: 통합된 위치의 경우 n = 438, n = 1,976 및 n = 2,124; 내인성 위치의 경우 n = 94, n = 435 및 n = 462. Spearman 상관 계수(R) 및 Pearson 상관 계수(r)를 표시하였다.

(b) 통합 및 내인성 위치에서 CBE_efficiency, CBE_proportion 및 DeepCBE의 성능 평가. 분석된 표적 서열의 수, 교정결과 빈도 및 각 결과의 효율(n)은 각각 다음과 같다: 통합된 부위의 경우 n = 482, n = 2,978 및 n = 3,107; n = 10; 내인성 부위의 경우 n = 522 및 n = 553.

(c) 동일 및 무작위 표적 쌍에 대해, 내인성 부위에서 (ABE_proportion 또는 CBE_proportion에 의해) 예측된 염기 편집 교정결과 빈도 vs. 측정된 염기 편집 교정결과 빈도 간의 대칭적(symmetrized) KL 발산 값(Kullback-Leibler divergence value)을 사용한 ABE_proportion 및 CBE_proportion의 성능 평가. 통합 부위(HT_ABE_Test (도면의 HT_ABE에 해당) 또는 HT_CBE_Test (도면의 HT_CBE에 해당)) vs. 내인성 부위(Endo_ABE_HEK293T 또는 Endo_CBE_HEK293T)에서 측정된 염기 편집 교정결과 빈도 간의 KL 발산 값을 참조 비교로서 나타내었다. 분석된 표적 서열의 수(n)는 다음과 같다: (좌측부터 우측으로) n = 59, n = 62, n = 269, n = 62, n = 52, n = 65, n = 290 및 n = 65.

(d) DeepABE / CBE 및 DeepABE / CBE-CA(염색질 접근성)의 성능 비교. 각 점은 측정된 인델 빈도와 예측된 활성 간의 Spearman 상관 계수를 나타낸다. 총 10 회에 걸친(n = 2 X 5) 미세조정(fine-tuning) 및 후속 테스트 결과를 나타내었다(NS: 현저하지 않음).

도 7은 DeepABE 및 DeepCBE 개발의 개요를 나타낸 도이다.

도 8 및 도 9는 HEK293T (a-c), HCT116 (d) 및 U2OS (e, f) 세포의 내인성 부위에서 ABE_efficiency, ABE_proportion 및 DeepABE의 성능 평가 결과를 나타낸 도이다. 분석된 표적 서열의 수, 결과 및 각 결과의 효율(n)은 각각 다음과 같다: n = 94; n = 435; HEK293T 세포, replicate 1의 경우 n = 462; n = 87; n = 353; HEK293T, replicate 2의 경우 n = 379; n = 75; n = 316; HEK293T, replicate 3의 경우 n = 337; n = 41; n = 213; HCT116 세포의 경우 n = 244; n = 24; n = 100; U2OS 세포, replicate 1의 경우 n = 124; n = 25; n = 91; U2OS 세포, replicate 2의 경우 n = 116. Spearman 상관 계수(R) 및 Pearson 상관 계수(r)를 표시하였다.

도 10 및 도 11은 HEK293T (a-c), HCT116 (d) 및 U2OS (e, f) 세포의 내인성 부위에서 CBE_efficiency, CBE_proportion 및 DeepCBE의 성능 평가 결과를 나타낸 도이다. 분석된 표적 서열의 수, 결과 및 각 결과의 효율(n)은 각각 다음과 같다: n = 102; n = 522; HEK293T, replicate 1의 경우 n = 553; n = 95; n = 531; HEK293T, replicate 2의 경우 n = 559; n = 83; n = 413; HEK293T, replicate 3의 경우 n = 436; n = 36; n = 193; HCT116 세포의 경우 n = 203; n = 28; n = 149; U2OS 세포, replicate 1의 경우 n = 170; n = 23; n = 136; U2OS 세포, replicate 2의 경우 n = 159. Spearman 상관 계수(R) 및 Pearson 상관 계수(r)를 표시하였다.

도 12는 ABE 및 CBE-유도된 모델링 및 질환-관련 인간 점돌연변이의 교정에 대한 예측 결과. 상기 병원성 또는 유사병원성 점돌연변이는 적절한 거리에서 PAM(NGG)과 관련되며, 원칙적으로 야생형(wild-type) 서열로부터 생성되거나 위치 3 내지 10의 편집가능한 윈도우를 사용하는 ABE 또는 CBE에 의해 야생형의 서열로 전환된 것일 수 있다.

(a) 원칙적으로 ABE(녹색) 또는 CBE(주황색)를 사용하여 생성될 수 있는 질환-관련 점돌연변이의 수.

(b) 원칙적으로 ABE(녹색) 또는 CBE(주황색)를 사용하여 교정할 수 있는 질환-관련 점돌연변이의 수.

(c, d) 인간 iPSC에서 질환-관련 점돌연변이의 모델링에 대한 (c) ABE_proportion 및 CBE_proportion 및 (d) DeepABE 및 DeepCBE의 성능 평가. 모델링은 ABE 또는 CBE에 의해 정상 인간 iPSC에서 병원성/유사병원성 돌연변이를 도입함으로써 수행되었다. Spearman 상관 계수(R) 및 Pearson 상관 계수(r)를 표시하였다. (c)에서 결과의 수는 n = 465 (ABE의 경우) 및 767 (CBE의 경우) (d)에서 병원성/유사병원성 돌연변이 부위의 수는 n = 31 (ABE의 경우) 및 49 (CBE의 경우)이다.

도 13은 ABE- 및 CBE-유도된 모델링 및 질환-관련 인간 점돌연변이 교정결과를 예측한 도이다.

(a) ABE- 및 CBE-유도된 모델링 및 질환-관련 인간 점돌연변이 교정에 대한 *in silico* 실험 결과를 나타낸 분포

도이다. 파이 차트(pie chart)에 효율 ≥ 5% (빨간색) 또는 <5% (파란색)으로 생성 또는 교정될 수 있는 병원성 및 유사병원성 점돌연변이의 수를 나타내었다. 염기교정 범위 내 단일 A 또는 C를 갖는 점돌연변이를 연한 빨간색 또는 연한 파란색으로 나타내었고, 염기교정 범위 내 2개 이상 A 또는 C를 가진 점돌연변이를 짙은 빨간색 또는 짙은 파란색으로 나타내었다. 각 파이의 영역은 상응하는 점돌연변이의 수에 비례한다.

구체적으로, (a)의 좌측 차트는 염기교정 유전자가위로 만들 수 있는 인간질환의 수를, (a)의 우측 상단 차트는 염기교정 유전자가위로 추가 변이 없이 사용 가능한 인간 질환의 수를 나타낸 것으로, 상기 파이 차트에서 염기교정 유전자가위의 효율은 본원의 "효율 스코어"와 "교정결과 스코어"를 곱하여 구한 값으로 평가하였다. (a)의 우측 하단 차트의 경우, 염기교정 유전자가위로 위험하지 않은 추가 변이를 동반하여 사용 가능한 인간 질환의 수를 나타낸 것으로, 본원의 "효율 스코어"와 "교정결과 스코어"를 곱한 뒤 위험한 변이가 없는 것으로 예측한 교정결과를 더하여 산출하였다.

(b) 염기교정 유전자가위로 만들 수 있는 교정결과의 예시를 나타낸 표이다.

발명을 실시하기 위한 구체적인 내용

[0140] 이하 일 양상을 실시예 및 실험예를 통하여 보다 상세하게 설명한다. 그러나 이들 실시예 및 실험예는 일 양상을 예시적으로 설명하기 위한 것으로 일 양상의 범위가 이들 실시예 및 실험예에 한정되는 것은 아니며, 일 양상의 실시예 및 실험예는 당업계에서 평균적인 지식을 가진 자에게 일 양상을 보다 완전하게 설명하기 위해서 제공되는 것이다.

[0141] **실험방법**

[0142] **1. 올리고뉴클레오티드 라이브러리 및 플라스미드 라이브러리의 제작**

[0143] Twist Bioscience Co.에 의뢰하여 총 17,840개 올리고뉴클레오티드 풀(pool)의 라이브러리를 제작하였다. 올리고뉴클레오티드 풀에 대한 표적 서열로서, 임의의 합성 서열은 sgRNA 또는 편집가능한 윈도우 내의 A 또는 C 포함 여부 등에 대한 어떠한 정보도 없이 생성되었다. 올리고뉴클레오티드 풀은 GeCKOv1 라이브러리로부터 임의로 선택된 9,824개의 표적 서열, 세포 표면 마커를 코딩(encoding)하는 유전자로부터 선택된 1,804개의 표적 서열, 베무라페닙, 셀루메티닙 및 6-티오구아닌에 대한 내성과 관련된 유전자로부터 선택된 2,484개의 표적 서열, GC 함량이 극히 낮거나 높은(≤20% 또는 ≥80%) 가이드 서열을 함유하는 998개 인풋 서열 및 관심 유전자와 관련된 인간 코딩 및 비코딩 유전자로부터 546개의 표적 서열이 포함되었다. 상기 546개의 표적 서열의 경우 표적 서열당 다섯개의 바코드를 사용하여 각 표적 부위에 대한 5-fold 커버리지(coverage)를 생성하였다. 종합하면, 17,840개의 올리고뉴클레오티드 세트는 9,824+1,804+2,484+998+(546X5) 개의 올리고뉴클레오티드로 구성되고, 9,824+1,804+2,484+998+546=15,656 쌍의 sgRNA-코딩 및 표적 서열을 포함한다. 이를 이용하여, HT_ABE_Train 및 HT_CBE_Train 데이터 세트를 생성하고, 이 중 546쌍은 HT_ABE_Test 및 HT_CBE_Test 데이터 세트 생성에 사용되었다.

[0144] 가이드 RNA 및 상응하는 표적 서열 쌍을 함유하는 플라스미드 라이브러리는 깁슨 어셈블리(Gibson assembly)와 후속하는 제한효소에 의한 절단(cutting) 및 결찰(ligation)의 2 단계 클로닝 과정을 수반하여 제작되었다.

[0145] **2. 플라스미드 벡터의 제조**

[0146] 개별 split-ABE 플라스미드의 제조를 위해, ABE7.10-코딩 서열(Addgene, no. 102919) 및 인테인-매개 split-Cas9-코딩 서열의 단편을 PCR 증폭시키고 lentiCas9-Blast(Addgene, no. 52962) 또는 pX601(Addgene, no. 61591) 플라스미드로 클로닝하였다. 생성된 플라스미드를 Lenti_Split-ABE-N-Blast, Lenti_SplitABE-C-Hygro-eGFP, AAV_Split-ABE-N 및 AAV_Split-ABE-C로 명명하였다.

[0147] 개별 split-BE4 플라스미드의 제조를 위해, BE4-코딩 서열(Addgene, no. 100802) 및 인테인-매개 split-Cas9-코딩 서열의 단편을 PCR 증폭시키고 lentiCas9-Blast(Addgene, no. 52962) 또는 pX601(Addgene, no. 61591) 플라스미드로 클로닝하였다. 생성된 플라스미드를 Lenti_Split-BE4-N-Blast, Lenti_Split-BE4-C-Hygro, AAV_Split-BE4-N 및 AAV_Split-BE4-C로 명명하였다.

[0148] 전장 ABE7.10을 코딩하는 렌티바이러스 벡터를 제조하기 위해, pLenti6/V5-GW/LacZ 플라스미드를 구매하고(ThermoFisher) 후속 복제를 위해 변형시켰다: EcoRV 제한 효소(NEB)로 LacZ 단편을 분해하고, 우드척 간염바이러스 전사후 조절요소(posttranscriptional regulatory element of woodchuck hepatitis)의 서열을 KpnI 효소-인식 부위에 삽입하였다. pCMV-ABE7.10 (Addgene, no. 107723)으로부터의 전장 ABE-코딩 서열을 변형된

pLenti6/V5-GW/LacZ 플라스미드에 클로닝하고 생성된 플라스미드를 Lenti-ABE-Blast로 명명하였다. pcDNA-BSD 플라스미드는 BSD 유전자를 PCR 증폭시키고 KpnI 및 EcoRI로 분해한 후 pcDNA 3.1(+) 벡터(Invitrogen)에 클로닝하여 준비하였다.

[0149] **3. 렌티바이러스의 생산**

[0150] 렌티바이러스 라이브러리 생산을 위해, 인간 배아 신장 세포인 HEK293T 세포(ATCC)를 준비하였다. 관심 유전자, psPAX2 및 pMD2.G를 함유하는 3개의 전달 플라스미드를 4:3:1의 중량비로 혼합하여 총 20㎍의 플라스미드 혼합물을 생성하고, 리포펙타민 2000 (Invitrogen)을 사용하여 이를 HEK293T 세포에 형질감염시켰다. 형질감염 후 12 시간에 신선한 배지를 세포에 가하고, 형질 감염 후 36 시간에 바이러스를 함유한 상층액을 수득하였다. 수득된 상층액은 Millex-HV 0.45 ㎛ 저-단백질 결합 멤브레인(Millipore)으로 여과하고, 분액은 사용시까지 -80℃에 보관하였다. 바이러스 수율은 Lenti-X p24 Rapid Titer Kit(Clontech)로 측정하여 검증하였다. 바이러스 역가 산출을 위해, 순차 희석된 바이러스 분액을 8 $\mu g m\ell^{-1}$의 폴리브렌(polybrene)의 존재에서 HEK293T 세포에 형질도입하고, 2 $\mu g m\ell^{-1}$ 퓨로마이신 또는 20 $\mu g m\ell^{-1}$ 블라스티시딘(blasticidin) S (InvivoGen)의 존재에서 배양하여 산출하였다.

[0151] 준비된 렌티바이러스 라이브러리의 형질도입을 위해, HEK293T 세포(9.0×10^6)를 배양 접시에 밤새 배양하였다. 감염다중도(multiplicity of infection: MOI) 0.3의 렌티바이러스 라이브러리를 8 $\mu g m\ell^{-1}$의 폴리브렌의 존재에서 HEK293T 세포에 형질도입하고, 세포를 15~18시간 동안 배양하였다. 세포를 2 $\mu g m\ell^{-1}$ 퓨로마이신의 존재에서 배양하여 형질도입되지 않은 세포를 제거하고, 9.0×10^6세포의 양으로 세포 라이브러리를 유지하였다.

[0152] **4. 세포 라이브러리에 ABE, CBE의 전달**

[0153] ABE의 경우, Lenti_Split-ABE-N-Blast 및 Lenti_Split-ABE-C-Hygro-Egfp를 1:1의 중량비로 혼합하여 총 240㎍의 플라스미드 혼합물을 생성하고, 리포펙타민 2000을 사용하여 이를 5.2×10^7 개 양의 세포 라이브러리로 전달하였다. CBE의 경우, Lenti_Split-BE4-N-Blast, Lenti_Split-BE4-C-Hygro 및 pcDNA-BSD를 9:9:2의 중량비로 혼합하여 총 20㎍의 플라스미드 혼합물을 생성하고, 네온 형질감염 시스템 (ThermoFisher Scientific)을 사용하여 2×10^6 라이브러리 세포에 전기 천공시켰다. 다음날 배양 배지를 10% FBS가 보충된 DMEM, 40㎍ ml^{-1} 블라스티시딘 S (InvivoGen) 및 80㎍ ml^{-1} 하이그로마이신(hygromycin) B 골드 (InvivoGen)로 교체하고, 형질감염 5일 후 배양물을 수집하여 사용하였다.

[0154] **5. 내인성 부위(endogenous sites)에서 염기 편집 빈도의 측정**

[0155] 내인성 부위에서의 ABE 및 CBE 활성 평가를 위해, 546개 내인성 표적 중 총 153개 표적 부위가 선택되었다(DHS 부위 70개, 비-DHS 부위 83개). HEK293T 세포에 sgRNA를 코딩하는 플라스미드 100ng (pRG2; Addgene no. 104174) 및 전장-길이 ABE7.10(Lenti-ABE-Blast) 또는 split-BE4(lenti-BE4-N-Blast, lenti-BE4-C-Hygro; 1:1 비율)를 코딩하는 플라스미드 100ng의 혼합물로 형질감염시켜 각각의 활성을 측정하였다. HCT116 세포는 sgRNA(pRG2)를 코딩하는 플라스미드 100ng, split-ABE (AAV-Split-ABE7.10-N, AAV-Split-ABE7.10-C; 1:1 비율) 또는 split-BE4 (AAV-Split-BE4-N, AAV-Split-BE4-C; 1:1 비율)를 코딩하는 플라스미드 200ng 및 eGFP 및 퓨로마이신 N-아세틸트랜스퍼라제를 코딩하는 플라스미드 50ng으로 형질감염시켰다. HEK293T 또는 HCT116 세포는 각 웰당 1.0×10^5 또는 4.0×10^4의 밀도로 형질감염시켰다. U2OS 세포는 sgRNA를 코딩하는 플라스미드 1㎍(pRG2, Addgene no. 104174), eGFP를 코딩하는 플라스미드 500ng 및 퓨로마이신 N-아세틸트랜스퍼라제(pEGFP-Puro, Addgene no. 45561) 및 전장 ABE7.10(Lenti-ABE-Blast) 또는 split-BE4 (Lenti-BE4-N-Blast, Lenti-BE4-C-Hygro; 1:1 비율)를 코딩하는 플라스미드 1㎍의 혼합물로 형질감염시켰다. 플라스미드 혼합물을 네온 형질 감염 시스템(ThermoFisher Scientific)을 사용하여 $1x10^6$ U2OS 세포로 전기 천공시켰다. 밤새 배양한 후, 배양 배지를 10% FBS 및 2㎍ ml^{-1}의 퓨로마이신 (InvivoGen)이 보충된 DMEM으로 교체하였다. ABE7.10 또는 BE4로 형질감염된 세포를 수집하고 5일 후(HEK293T 세포 및 U2OS 세포) 또는 3.5 일(HCT116 세포) 후 딥 시퀀싱(Deep sequencing) 하였다.

[0156] **6. 질환 모델링 및 질환-연관 돌연변이의 교정**

[0157] 인간 유도만능줄기세포에서 염기교정 유전자가위로 만들 수 있는 질환 모델링 및 질환-관련 돌연변이의 교정을

확인하였다. 먼저, ClinVar 데이터베이스의 질환 관련 점돌연변이들 중 다수 개(적어도 3개 이상)의 표적 A 또는 C들을 포함하고 좁은 편집가능한 윈도우(위치 4 내지 8)에 위치한 총 95개 돌연변이를 선택하였다. 정상 인간 iPSC를 Essential 8 배지(ThermoFisher Scientific)에서 배양하였다. 질환-관련 점돌연변이를 유도하기 위해, 인간 iPSC를 split-ABE (AAV-Split-ABE7.10-N, AAV-Split-ABE7.10-C; 1:1 비율) 또는 split-BE4 (AAV-Split-BE4-N, AAV-Split-BE4-C; 1:1 비율), sgRNA (pRG2, Addgene no. 104174)를 코딩하는 플라스미드 150 ng 및 eGFP와 퓨로마이신 N-아세틸트랜스퍼라제 (pEGFP-Puro, Addgene no. 45561)를 코딩하는 플라스미드 100ng의 혼합물 500 ng으로 형질감염시켰다. ABE 및 CBE-매개된 질환-관련 점돌연변이 교정 효율을 측정하기 위해, 병원성 점돌연변이를 포함하는 합성 표적 서열을 렌티바이러스로 정상 인간 iPSC에 전달하였다. 배양 배지에 4 μg ml^{-1} 블라스티시딘을 첨가하여 형질도입되지 않은 세포를 제거하였다. 다음으로, 리포펙타민 시약(ThermoFisher Scientific)을 사용하여 표적 서열이 도입된 iPSC를 상기 3종류의 플라스미드 혼합물로 형질감염시켰다. 형질감염 후 밤새 배양한 뒤, 배양 배지를 10 μM Y-27632(Sigma-Aldrich) 및 1 μg ml^{-1} 퓨로마이신(Gibco)으로 보충된 Essential 8 배지로 교체하였다. 퓨로마이신으로 선별한 후 24시간에 배지를 제거하고 웰 당 10 μM Y-27632로 보충된 Essential 8 배지를 세포에 가해주었다. 형질감염 3일 후, 세포를 수집하고 유전체 DNA를 딥 시퀀싱하여 유전자가위의 효율 및 염기 편집의 결과를 측정하였다.

[0158] **7. 딥 시퀀싱(Deep sequencing)**

[0159] Wizard Genomic DNA 정제 키트(Promega)를 사용하여 세포로부터 유전체 DNA를 분리하였다. PCR은 2X pfu PCR Smart mix (Solgent)를 사용하여 수행하였다. 고-처리량 실험을 위해, 첫번째 PCR에서 각 세포 라이브러리에 대해 총 264 μg의 유전체 DNA를 이용하여 라이브러리에 대해 1,400x 이상의 커버리지가 되도록 하였다. 생성된 PCR 산물을 단일 풀(pool)로 합한 후 MEGAquick-spin total fragment DNA 정제 키트 (iNtRON Biotechnology)로 정제하였다. 정제된 산물 중 20 ng의 시료를 Illumina 어댑터 및 바코드 서열을 함유한 프라이머를 사용하여 2차 PCR 증폭하였다.

[0160] 실험에 사용된 프라이머는 다음과 같다(5'-3').

[0161] 올리고뉴클레오티드 풀(pool) 증폭용 프라이머

[0162] - 정방향: TTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC (서열번호 1)

[0163] - 역방향: GAGTAAGCTGACCGCTGAAGTACAAGTGGTAGAGTAGAGATCTAGTTACGCCAAGCT (서열번호 2)

[0164] 1차 PCR 반응용 프라이머

[0165] - 정방향: ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGAAAAAGTGGCACCGAGTCG (서열번호 3)

[0166] ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTTGAAAAAGTGGCACCGAGTCG (서열번호 4)

[0167] ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGCTTGAAAAAGTGGCACCGAGTCG (서열번호 5)

[0168] - 역방향: GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTAAGTCGAGTAAGCTGACCGCTGAAG (서열번호 6)

[0169] GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTATTAAGTCGAGTAAGCTGACCGCTGAAG (서열번호 7)

[0170] GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTATTAAGTCGAGTAAGCTGACCGCTGAAG (서열번호 8)

[0171] 2차 PCR 반응용 프라이머(염기서열 중앙의 소문자는 8bp 바코드 서열을 의미함)

[0172] - 정방향: AATGATACGGCGACCACCGAGATCTACACtatagcctACACTCTTTCCCTACACGAC (서열번호 9)

[0173] AATGATACGGCGACCACCGAGATCTACACatagaggcACACTCTTTCCCTACACGAC (서열번호 10)

[0174] AATGATACGGCGACCACCGAGATCTACACcctatcctACACTCTTTCCCTACACGAC (서열번호 11)

[0175] AATGATACGGCGACCACCGAGATCTACACggctctgaACACTCTTTCCCTACACGAC (서열번호 12)

[0176] AATGATACGGCGACCACCGAGATCTACACaggcgaagACACTCTTTCCCTACACGAC (서열번호 13)

[0177] AATGATACGGCGACCACCGAGATCTACACtaatcttaACACTCTTTCCCTACACGAC (서열번호 14)

[0178] AATGATACGGCGACCACCGAGATCTACACcaggacgtACACTCTTTCCCTACACGAC (서열번호 15)

[0179] AATGATACGGCGACCACCGAGATCTACACgtactgacACACTCTTTCCCTACACGAC (서열번호 16)

[0180] - 역방향: CAAGCAGAAGACGGCATACGAGATcgagtaatGTGACTGGAGTTCAGACGTGT (서열번호 17)

[0181] CAAGCAGAAGACGGCATACGAGATctctccggaGTGACTGGAGTTCAGACGTGT (서열번호 18)

[0182] CAAGCAGAAGACGGCATACGAGATaatgagcgGTGACTGGAGTTCAGACGTGT (서열번호 19)

[0183] CAAGCAGAAGACGGCATACGAGATggaatctcGTGACTGGAGTTCAGACGTGT (서열번호 20)

[0184] CAAGCAGAAGACGGCATACGAGATttctgaatGTGACTGGAGTTCAGACGTGT (서열번호 21)

[0185] CAAGCAGAAGACGGCATACGAGATacgaattcGTGACTGGAGTTCAGACGTGT (서열번호 22)

[0186] CAAGCAGAAGACGGCATACGAGATagcttcagGTGACTGGAGTTCAGACGTGT (서열번호 23)

[0187] CAAGCAGAAGACGGCATACGAGATgcgcattaGTGACTGGAGTTCAGACGTGT (서열번호 24)

[0188] CAAGCAGAAGACGGCATACGAGATcatagccgGTGACTGGAGTTCAGACGTGT (서열번호 25)

[0189] CAAGCAGAAGACGGCATACGAGATttcgcggaGTGACTGGAGTTCAGACGTGT (서열번호 26)

[0190] CAAGCAGAAGACGGCATACGAGATgcgcgagaGTGACTGGAGTTCAGACGTGT (서열번호 27)

[0191] CAAGCAGAAGACGGCATACGAGATctatcgctGTGACTGGAGTTCAGACGTGT (서열번호 28)

[0192] **8. 염기 편집 효율 및 교정결과 빈도의 분석**

[0193] 염기 편집 효율 및 교정결과의 분석을 위해 Python scripts 프로그램을 변형하여 서열 데이터를 분석하였다. 표적 서열의 상류에 위치한 독특한 15-nt 바코드 서열을 이용하여 각각의 가이드 RNA 및 표적 서열 쌍을 식별하였다. 인델이 예상된 절단 부위(절단 부위의 가운데에 위치한 8-nt 영역) 주변의 삽입 또는 결실은 뉴클레아제로 유도된 인델로 간주하였다. 염기 편집의 효율 및 결과를 분석하기 위해, 독특한 바코드 서열로 소팅(sorting)한 리드(read)를 Python script로 정렬하고, Needleman-Wunsch 알고리즘을 사용하여 참조 서열과의 비교를 통해 인델을 포함하는 리드를 분류하였다. 편집가능한 윈도우 내의 ABE 및 CBE의 표적 뉴클레오티드가 각각 T 또는 G로 전환되었을 때, 이를 염기 편집된 것으로 카운트하였다. 그 다음, 하기 수학식에 따라 각 위치에서의 염기 편집 효율을 산출하였다.

수학식 1

[0194]
$$\text{염기 편집 효율 (\%)} = \frac{\text{넓은염기교정 범위(표적 서열의 위치3 내지10)에서 의도된표적 뉴클레오타드 전환을 포함하는 모든 서열의 총 리드(read)}}{\text{총 리드(read)}} \times 100$$

[0195] 총 리드 수는 인델을 포함하는 모든 염기 콜(call)의 합계이다. 염기 편집 빈도에 대한 분석의 정확성을 증가시키기 위해, 딥 시퀀싱 데이터 중 100개 미만의 총 딥 시퀀싱 리드 카운트를 갖는 표적 서열은 제외시켰다.

[0196] 위치 3 내지 10의 편집가능한 윈도우 내에 단 하나의 표적 뉴클레오티드만이 존재하는 경우, 염기 편집 효율은 표적 뉴클레오티드에서의 전환 효율과 동일하고, 염기 편집의 결과는 편집되거나, 되지 않거나의 두가지 경우로 수렴한다. 다만, 표적 서열 조성이 랜덤인 경우, 편집가능한 윈도우에서 단 하나의 표적 뉴클레오티드를 가질 확률은 13%이며, 나머지 87%의 경우 하나 이상의 표적 뉴클레오티드가 존재하여 염기 편집 후 복잡한 결과를 초래한다. 이에, 교정결과의 빈도를 분석하기 위해 Python script를 사용하여 염기 편집 윈도우의 교정결과에 따라 정렬된 리드를 다시 계산하였다. 표적 뉴클레오티드 전환 효율은 위치 3 내지 10의 편집가능한 윈도우 내에서의 뉴클레오티드의 위치에 대한 영향을 받지 않으므로, 표적 뉴클레오티드가 의도하지 않은 뉴클레오티드로 전환된 경우는 제외하였다. 각 염기 편집 결과의 빈도는 하기 수학식에 따라 산출하였다.

수학식 2

[0197]
$$\text{염기 편집결과 빈도} = \frac{\text{특정 염기}-\text{편집된 결과 서열의 리드(read)}}{\text{넓은염기교정 범위(표적 서열의 위치3 내지10)에서의도된 표적 뉴클레오타드 전환을 포함하는 모든 서열의 총 리드(read)}}$$

[0198] 염기 편집 결과에 대한 분석의 정확성을 높이기 위해, 변형 카운트가 100 미만인 표적 서열(또는 iPSC를 사용한 실험의 경우 200 미만)을 필터링하고, 의도되지 않은 뉴클레오티드 전환을 포함하는 리드를 제외하였다. 염기 편집 결과의 절대적 빈도는 염기 편집 결과 빈도와 염기 편집 효율을 곱하여 산출할 수 있으며, 하기 수학식에 따라 계산할 수 있다:

수학식 3

[0199]
$$\text{염기 편집 결과의 절대적 빈도}(\%) = \frac{\text{특정 염기}-\text{편집된 결과 서열의 리드(read)}}{\text{총 리드(read)}} \times 100$$

[0200]
$$= \frac{\text{넓은염기교정 범위(표적 서열의 위치3 내지10)에서 의도된표적 뉴클레오타드 전환을 포함하는 모든 서열의 총 리드(read)}}{\text{총 리드(read)}} \times \frac{\text{특정 염기}-\text{편집된 결과 서열의 리드(read)}}{\text{넓은염기교정 범위(표적 서열의 위치3 내지10)에서의도된 표적 뉴클레오타드 전환을 포함하는 모든 서열의 총 리드(read)}} \times 100$$

[0201]
$$= \text{염기 편집 효율} \times \text{교정결과 빈도}\,(\%)$$

[0202] **9. 염색질 접근성의 고려**

[0203] ENCODE42로부터 얻은 DNase-seq 좁은 피크 데이터를 염색질 접근성(chromatin accessibility) 고려에 사용하였다. 각 표적 위치에 대하여, bowtie43을 사용하여 23개 염기의 PAM + 프로토스페이서 서열을 hg19 인간 참조 유전체에 정렬하였다. DNase-seq 좁은 피크와 오버랩된 표적 부위를 DHS 부위로 간주하였다.

[0204] **10. 컨볼루션 신경망을 사용한 딥러닝**

[0205] 풀링층(pooling layer)이 없는 컨볼루션 신경망(Convolutional neural network: CNN)을 사용하여 생성한 데이터 세트에 대해 딥러닝을 수행하고 염기교정 유전자가위의 효율과 교정결과를 예측하는 모델을 개발하였다. 학습 데이터 세트에 모델이 과적합(overfitting)되는 것을 방지하기 위해 검증 스코어를 기반으로 조기 중지(early stopping)를 사용하고, 각 층(layer)에서 드롭아웃률 0.3을 사용하였다. 모든 모델은 필터 모양(1,3); 채널 차수, 4; 스트라이드(stride) (1,1) 및 컨볼루션 층(layer)은 패딩이 적용되지 않도록 하였고, ABE_proportion, ABE_efficiency, CBE_proportion 및 CBE_efficiency 모델에 대해 각각 150, 60, 60 및 150 필터(filter)를 사용하였다. 컨볼루션 층 이후에는 하나(ABE_proportion, ABE_efficiency) 또는 두 개(CBE_proportion, CBE_efficiency)의 완전히 연결된 층이 사용되었으며, 히든 레이어의 노드(node) 수는 ABE_proportion, ABE_efficiency, CBE_proportion 및 CBE_efficiency가 각각 256, 500, 256/256 및 500/50이었다. 합성곱 연산을 거친 후, ReLU 활성화 함수를 적용하여 연산하였다.

[0206] **11. 대칭적 KL 발산(Symmetrized KL divergence)**

[0207] 염기 편집 결과 분포의 유사성을 계산하기 위하여 대칭적 쿨백-라이블러 발산을 사용하였다. 0으로 나눠지는 것을 피하기 위해, 학습 데이터 세트에 0.001, 테스트 데이터 세트에 0.5의 거짓 카운트(pseudo count)를 추가한 뒤 하기 수학식에 의해 KL 발산을 계산하였다. P_i와 Q_i는 각각 예측 및 평가된 빈도를 의미한다.

$$Symmetrized\ KL = \sum_i P_i log \frac{P_i}{Q_i} + \sum_i Q_i log \frac{Q_i}{P_i}$$

[0208]

[0209] **12. ClinVar 데이터 분석**

[0210] 인간 질환-관련 돌연변이의 모델링 및 치료적 교정을 위한 염기 편집 효율 및 결과를 검증하기 위하여, 공개적으로 사용가능한 ClinVar(v. clinvar_20190219_hg38) 데이터 세트를 필터링하여 사용하였다. 필터링은 먼저, PAM(NGG) 서열에서 적절한 거리에 있으면서, 위치 3 내지 10의 편집가능한 윈도우 내에 A 또는 C를 가진 점돌연변이를 선택한 뒤, ClinVar 데이터베이스 상에서 병원성 또는 유사병원성으로 표시된 표적 점돌연변이를 선택하는 2단계로 수행되었다.

[0211] **13. 통계적 유의성**

[0212] DHS 부위와 비-DHS 부위 간 염기 편집 효율을 비교하기 위해 양측 스튜던트 t-test(two-tailed Student's t-test)를 사용하였다. DeepABE/CBE와 DeepABE-CA/CBE-CA(염색질 접근성을 고려한 모델) 간 성능을 비교하기 위해 Steiger's test를 사용하였다. 표적 염기 주위의 서열 컨텍스트가 염기 편집 빈도에 미치는 영향을 확인하기 위해 one-way ANOVA 및 Tukey의 사후 검정을 사용하였다. 통계적 유의성은 PASW Statistics (v.18.0, IBM) 및 Microsoft Excel (v.16.0, Microsoft Corporation)을 사용하여 결정하였다.

[0213] **14. Python script 및 딥 시퀀싱 데이터**

[0214] 일 구체예에서, DeepABE, DeepCBE, DeepCBE-CA의 Python script는 github (https://github.com/MyungjaeSong/Paired-Library, https://github.com/CRISPRJWCHOI/BaseEditing_tool)에서 제공되고, 일 구체예에서 생성한 딥 시퀀싱 데이터는 NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/)에서 접근번호 SRP150719 (PRJNA476544)로 접근가능하다.

[0215] **실시예**

[0216] **실시예 1. 유전자가위의 효율 및 교정결과 데이터 세트 생산**

[0217] **1-1. 데이터 세트의 생산**

[0218] 유전자가위 대량검증을 위해, 이전 연구(Kim et al, Nat Methods, 2017)에서 사용하였던 15,656개의 가이드 RNA 코딩 및 표적 서열 쌍의 렌티바이러스 라이브러리를 사용하여 유전자가위의 효율 및 교정결과 데이터 세트를 생산하였다. 총 15,656개의 서열 쌍 중에서, 프로토스페이서 영역의 5' 말단으로부터 3 내지 10 bp 위치(위치 20은 PAM(NGG)의 바로 상류에 자리함)에서 13,504개는 적어도 하나의 표적 아데닌을 포함하였고, 14,157개는 적어도 하나의 표적 시토신을 포함하였다. 세포 라이브러리를 ABE7.10 또는 BE4를 코딩하는 플라스미드로 형질감염시키고, 형질감염 5일 후 표적 뉴클레오티드에서의 염기 전환 효율 및 염기 편집 결과를 평가하였다. 이 때 ABE 또는 CBE를 코딩하는 플라스미드 벡터는 비교적 크기가 크기 때문에, 인테인(intein)-매개된 split-Cas9-기반의 ABE 및 CBE를 함께 사용하였다. 그 결과, 4개의 독립적인 데이터 세트가 생성되었고 이를 각각 HT_ABE_Train, HT_ABE_Test, HT_CBE_Train, 및 HT_CBE_Test로 명명하였다. 또한, 96개 및 102개의 내인성(endogenous) 표적 부위에서의 ABE 및 CBE활성에 대한 데이터 세트를 생성하고, 이를 각각 Endo_ABE 및 Endo_CBE로 명명하였다. 그 결과, DNase I 과민성 영역(DHS)에 해당하는 표적 부위의 위치 3 내지 10에서의 ABE 및 CBE 활성은 비-DHS 부위에 비해 각각 1.1배(P=0.55) 또는 1.8배(P=0.0077) 높게 나타남을 확인하였다. 상기 결과를 통해, CBE 활성은 염색질 접근성이 낮은 부위보다 높은 부위에서 증가하고, ABE 활성은 염색질 접근성에 거의 영향을 받지 않음을 확인하였다.

[0219] **1-2. 염기교정 유전자가위 활성에 영향을 미치는 요인 탐색**

[0220] 고-처리량(high-throughput) 평가를 통해 ABE- 및 CBE-지정된 염기 편집은 위치 3 내지 10에 해당하는 넓은 윈도우에서 발생하나, 상대적으로 높은 수준의 염기 편집은 ABE 및 CBE 모두 위치 4 내지 8의 좁은 윈도우에서 달성할 수 있음을 확인하였다(도 1의 a, b).

[0221] 동일한 표적 서열에서 Cas9의 활성과 ABE 또는 CBE의 활성을 비교할 때, ABE 또는 CBE 활성이 높을 때 대부분의 경우 Cas9 활성 역시 높아 크지는 않은 양의 상관관계가 확인되었다(도 1의 c, d). 그러나, Cas9 활성이 높은 경우 ABE 또는 CBE 활성은 때로는 높고 때로는 낮아, 결과적으로 ABE, CBE와 SpCas9 뉴클레아제 활성은 비대칭적 상관관계를 나타내었다. 나아가, 표적 서열의 각 위치에서 ABE 또는 CBE의 높은 활성 또는 Cas9의 높은 활성

과 관련된 뉴클레오티드 선호도를 결정하고, 상기 비대칭적 상관관계는 표적 서열의 특정 위치에서 SpCas9 및 ABE/CBE의 뉴클레오티드 선호도가 상이하기 때문임을 확인하였다. 구체적으로, ABE/CBE의 경우 표적 뉴클레오티드에 바로 인접한 뉴클레오티드(표적 뉴클레오티드 ±1 bp)에서 강한 뉴클레오티드 선호도가 관찰되었고, 이와 같은 선호는 표적 뉴클레오티드의 위치에 관계없이 보존되었다.

[0222] 나아가, ABE 및 CBE 활성에 대한 염기 편집에 대한 서열 컨텍스트 주변의 영향을 분석하기 위해 16개의 모든 가능한 NAN 서열에서 ABE의 염기 편집 효율을 분석한 결과, 염기 전환 효율은 TAB(특히 TAY, Y=C 또는 T)에서 가장 높고, AAC, GAA, CAA, AAT, AAG 및 AAA와 같이 A가 많은 컨텍스트(A-rich contexts)에서 가장 낮음을 확인하였다(도 1의 e). TAA의 경우, T에 따른 영향과 3' 위치의 A로 인한 영향이 상쇄되어 염기 편집 효율이 높지 않았다. 마찬가지로, 모든 가능한 NCN 서열에서 CBE의 염기 편집 효율을 분석한 결과, ABE와 유사하게 염기 전환 효율은 5' 위치가 T일 때(즉, TCN) 가장 높고, 5' 위치가 G일 때(예를 들어, GCC, GCT, GCG 및 GCA) 가장 낮음을 확인하였다(도 1의 f). ABE와 대조적으로, TCC, CCC, CCT, ACC, CCA, CCG 및 GCC와 같이 하나 또는 이웃하는 두개 위치에서 표적 뉴클레오티드가 반복되는 경우에는, C 염기의 전환 효율이 약간 더 높게 나타남을 확인하였다.

[0223] 이와 같은 컨텍스트 선호는 좁은 윈도우의 모든 위치(위치 4 내지 8)에서 관찰되어, 염기 편집 효율에 있어서 뉴클레오티드 컨텍스트가 강한 영향을 미침을 확인하였다(도 2). 구체적으로, 서열 컨텍스트와 관련된 표적 뉴클레오티드의 전환 효율에 대한 최대-배수 차이(maximum-fold difference)는 ABE와 CBE가 각각 45배, 13배로 나타남을 확인하였다.

[0224] 반면, SpCas9에 의한 인델 생성의 경우 이와 같은 컨텍스트 선호가 거의 관찰되지 않았다(도 3).

[0225] **실시예 2. 염기교정 유전자가위 활성 및 결과 예측 모델의 구축**

[0226] **2-1. 효율 예측 모델의 개발**

[0227] 실시예 1에서 생산한 유전자가위의 효율 및 교정결과 데이터와 딥 러닝(Deep learning) 기술을 이용하여 예측 모델을 구축하였다. 먼저, 딥러닝 프레임워크와 HT_ABE_Train 및 HT_CBE_Train 학습(training) 데이터 세트를 사용하여 ABE_efficiency 및 CBE_efficiency로 명명되는 유전자가위의 염기 편집 효율 예측 모델을 개발하였다(도 4). 이 때, 염기 편집 효율(base-editing efficiency)은 분석된 전체 DNA 카피 중에서 염기 편집된 뉴클레오티드의 수와 관계없이, 넓은 편집가능한 윈도우(ABE 및 CBE의 경우 위치 3 내지 10) 내 염기 편집된 서열을 가진 DNA 카피의 백분율을 지칭하며, 딥러닝 기법으로 신경망(newral network architecture)을 사용하였다. 10번 교차검증(10-fold cross validation)을 사용하여 히든 레이어(hidden layer)의 수와 인풋 서열의 길이를 결정하였다. ABE_efficiency의 경우 가장 높은 성능을 보인 2개 히든 레이어 모델과 3개 히든 레이어 모델이 비슷한 성능을 나타내어(각각 0.776 vs. 0.778) 2개의 히든 레이어 모델을 선택하였다(도 5).

[0228] 그 결과, ABE_efficiency 모델은 테스트 데이터 세트로 HT_ABE_Test 및 Endo_ABE_HEK293T를 사용할 때 각각 Spearman R=0.72, Pearson r=0.70(HT_ABE_Test) 및 Spearman R=0.76, Pearson r=0.70(Endo_ABE_HEK293T)의 상관 계수에 도달하여 우수한 성능을 나타내었고(도 6의 a), 유사하게 CBE_efficiency 모델 역시 HT_CBE_Test 및 Endo_CBE_HEK293T 사용 시 Spearman R=0.79, Pearson r=0.78(HT_CBE_Test) 및 Spearman R=0.69, Pearson r=0.60(Endo_CBE_HEK293T)의 상관 계수에 도달하여 우수한 성능을 나타냄을 확인하였다(도 6의 b).

[0229] **2-2. 교정결과 예측 모델의 개발**

[0230] 편집가능한 윈도우에 하나 이상의 표적 뉴클레오티드가 있는 경우 염기 편집의 결과로 다양한 서열이 생성되므로, 이러한 교정결과의 상대적 빈도를 예측하기 위해 또 다른 딥러닝 프레임워크와 HT_ABE_Train 및 HT_CBE_Train 학습 데이터 세트를 사용하여 ABE_proportion 및 CBE_proportion로 명명되는 교정결과 예측 모델을 개발하였다. Spearman 상관 계수 외에, 대칭적 쿨백-라이블러 발산(symmetric Kullback-Leibler (KL) divergence)을 함께 사용하여 교정결과 빈도의 유사성을 반영하였다.

[0231] 그 결과, ABE_proportion 모델은 각각 Pearson r=0.95(HT_ABE_Test) 및 Pearson r=0.93(Endo_ABE_HEK293T)의 높은 성능을 나타내었고(도 6의 a), 유사하게 CBE_proportion 모델 역시 Pearson r=0.95(HT_CBE_Test) 및 Pearson r=0.91(Endo_CBE_HEK293T)의 높은 성능을 나타내었다(도 6의 b). 나아가, Endo_ABE_HEK293T의 교정결과 빈도와 ABE_proportion으로부터의 예측값 간의 대칭적 KL 발산은 동일한 표적 서열에서의 Endo_ABE_HEK293T와 HT_ABE_Test 간 KL 발산 값(중앙값 KL = 0.10)과 유사하게 낮았으며(중앙값 KL = 0.11), Endo_CBE_HEK293T의 교정결과 빈도와 CBE_proportion으로부터의 예측값 간의 대칭적 KL 발산 역시 동일한 표적 서열에서의 Endo_CBE_HEK293T와 HT_CBE_Test 간 KL 발산 값(중앙값 KL = 0.18)과 유사하게 낮음을 확인하였다(중앙값 KL =

0.36)(도 6의 c).

[0232] 구체적으로, ABE 및 CBE 염기교정 유전자가위의 염기 편집 효율 및 교정결과를 예측한 결과를 하기 표 1 및 3에, 상기 염기교정 유전자가위로 만들거나 교정할 수 있는 인간 점돌연변이 질환의 예를 표 2 및 4에 나타내었다.

[0233] [표 1]

[0234] ABE

	참조 서열 (정상서열; 소문자는 점돌연변이의 위치를 나타냄)	표적 서열 (점돌연변이를 포함하는 서열; 소문자는 점돌연변이를 나타냄, 총 30 bps = 4 bp 이웃 서열 + 20 bp 프로토스페이서 + 3bp NGG PAM + 3 bp 이웃 서열)	염기 편집의 결과
1	CAGCTCAATgTAAGTAGACCGTTCAGGAAC	CAGCTCAATaTAAGTAGACCGTTCAGGAAC	CAGCTCAATgTAAGTAGACCGTTCAGGAAC
2	CTGTACCAAGTgCCACAAAGGTAGGGGCAA	CTGTACCAAGTaCCACAAAGGTAGGGGCAA	CTGTACCAAGTgCCACAAAGGTAGGGGCAA
3	CCACGGCAAGTgCAGAAACACCATTGGCAG	CCACGGCAAGTaCAGAAACACCATTGGCAG	CCACGGCAAGTgCAGAAACACCATTGGCAG
4	GCGTGTGAAgCCCGGACTCACTGGAGGCCT	GCGTGTGAAaCCCGGACTCACTGGAGGCCT	GCGTGTGAAgCCCGGACTCACTGGAGGCCT
5	ATTACAAATgTAAGGCCAAAAATCTGGCTG	ATTACAAATaTAAGGCCAAAAATCTGGCTG	ATTACAAATgTAAGGCCAAAAATCTGGCTG
6	TGGCTACAAGTgCCAGTGTGAGGAAGGCTT	TGGCTACAAGTaCCAGTGTGAGGAAGGCTT	TGGCTACAAGTgCCAGTGTGAGGAAGGCTT
7	GCGGATGATgTACTGAATCTGCGCTGGCGC	GCGGATGATaTACTGAATCTGCGCTGGCGC	GCGGATGATgTGCTGAATCTGCGCTGGCGC
8	GCTCTCCAACTgCAGCGTCTCCTTCGGCTG	GCTCTCCAACTaCAGCGTCTCCTTCGGCTG	GCTCTCCAACTgCAGCGTCTCCTTCGGCTG

[0235]

[0236] [표 2]

[0237] ABE

	예측된 결과 빈도 (%)	관련 질환
1	**47.57892529**	Usher_syndrome,_type_1F
2	**44.57045926**	TNF_receptor-associated_periodic_fever_syndrome_(TRAPS)
3	**43.53880158**	Marfan_syndrome
4	**42.98826047**	Maturity-onset_diabetes_of_the_young,_type_3
5	**41.97446974**	Congenital_stationary_night_blindness,_type_1F
6	**41.45525213**	Familial_hypercholesterolemia
7	**40.43030665**	Congenital_myasthenic_syndrome_13
8	**40.08395373**	Lynch_syndrome

[0238]

[0239] [표 3]

[0240] CBE

	참조 서열 (정상서열; 소문자는 점돌연변이의 위치를 나타냄)	표적 서열 (점돌연변이를 포함하는 서열; 소문자는 점돌연변이를 나타냄, 총 30 bps = 4 bp 이웃 서열 + 20 bp 프로토스페이서 + 3bp NGG PAM + 3 bp 이웃 서열)	염기 편집의 결과
1	CTCCAGTTCCtGACGGCTGAGGAGCGGAAG	CTCCAGTTCCcGACGGCTGAGGAGCGGAAG	CTCCAGTTTTtGACGGCTGAGGAGCGGAAG
2	TACAGATTCCtGGAGGAGATGCGGCGGCGG	TACAGATTCCcGGAGGAGATGCGGCGGCGG	TACAGATTTTtGGAGGAGATGCGGCGGCGG
3	CAGGGGTCTCtGCAGGACATGCTGTGGCAG	CAGGGGTCTCcGCAGGACATGCTGTGGCAG	CAGGGGTTTTtGCAGGACATGCTGTGGCAG
4	ACCCCGTCtATCAGAAGACCACAGAGGATG	ACCCCGTCcATCAGAAGACCACAGAGGATG	ACCCCGTTtATTAGAAGACCACAGAGGATG
5	AGGCAGGTTCCtAAGACACAGGGCAGGCAC	AGGCAGGTTCCcAAGACACAGGGCAGGCAC	AGGCAGGTTTTtAAGACACAGGGCAGGCAC
6	TCTGGTTCCtGGAAAGCATTAAGAAGGCAG	TCTGGTTCCcGGAAAGCATTAAGAAGGCAG	TCTGGTTTTtGGAAAGCATTAAGAAGGCAG
7	GCCGAGAACCtGGATACACAGCCGAGGAGA	GCCGAGAACCcGGATACACAGCCGAGGAGA	GCCGAGAATTtGGATACACAGCCGAGGAGA
8	GTCAGAGTCCtGTGAAAGAAACACAGGCAC	GTCAGAGTCCcGTGAAAGAAACACAGGCAC	GTCAGAGTTTtGTGAAAGAAACACAGGCAC

[0241]

[0242] [표 4]

[0243] CBE

	예측된 결과 빈도 (%)	관련 질환
1	**27.76197788**	Loeys-Dietz_syndrome_2
2	**27.26662656**	RETINITIS_PIGMENTOSA_80
3	**25.88935684**	Leptin_deficiency_or_dysfunction
4	**25.35618417**	Familial_hypercholesterolemia
5	**24.7623159**	Deafness,_autosomal_recessive_3
6	**23.55097809**	Retinitis_pigmentosa_51
7	**23.40125579**	Cholesterol_monooxygenase_(side-chain_cleaving)_deficiency
8	**22.88536345**	Progressive_myoclonic_epilepsy

[0244]

[0245] 표 2 및 4에 나타낸 바와 같이, 염기교정 유전자가위로 만들거나 교정할 수 있는 인간 점돌연변이 질환의 예로서 ABE의 경우 어셔 증후군(Usher syndrome), 종양괴사인자 수용체 관련 주기적 증후군(TNF receptor-associated periodic syndrome: TRAPS), 마판 증후군(marfan syndrome), 제3형 청년기 발병 당뇨병(Type 3 form of Maturity-Onset Diabetes of the Young: MODY3), 선천성 비진행성 야맹증(Congenital stationary night blindness type 1F), 가족성 고콜레스테롤혈증(Familial hypercholesterolemia), 선천근육무력증후군(congenital myasthenic syndrome: CMS), 린치증후군(Lynch syndrome) 등이 확인되었고, CBE의 경우 로이-디에츠 증후군(Loeys-Dietz syndrome: LDS), 망막색소변성증(retinitis pigmentosa), 렙틴 결핍 또는 장애(Leptin deficiency 또는 dysfunction), 가족성 고콜레스테롤혈증(Familial hypercholesterolemia), 상염색체 열성 청각장애(autosomal recessive deafness), 콜레스테롤 모노옥시다제 결핍(cholesterol monooxygenase (side-chain-cleaving) deficiency), 진행성 근간대성간질(progressive myoclonus epilepsy) 등이 확인되었다.

[0246] 상기 결과로부터, 개발된 ABE_proportion 및 CBE_proportion 모델이 염기 편집의 결과 빈도를 높은 정확성으로 예측할 수 있음을 확인하였다. 한편, 이와 같은 효과는 전통적인 머신러닝 또는 얕은 신경망(예를 들어, AdaBoost, Boosted RT, SVM, Ridge, Lasso, ElasticNet, Random Forest 등)을 사용하는 경우 확인되지 않아, 일 양상에 따른 예측 시스템은 딥러닝을 이용함에 따라 매우 높은 정확도로 염기교정 유전자가위의 염기교정 효율과 교정결과를 예측할 수 있는 것임을 알 수 있었다.

[0247] 한편, 본 발명자들은 이전 연구를 통해 내인성 부위에서의 Cas12a 효율 예측은 염색질 접근성(chromatin accessibility)을 고려함으로써 개선될 수 있음을 확인하였다. 이에 일 구현예에서, CBE는 염색질 접근성에 영향을 받았으나 ABE는 그렇지 않음을 확인하여, 염색질 접근성의 고려가 염기교정 유전자가위 효율 예측을 향상시킬 수 있는지에 대한 추가 테스트를 수행하였다. 먼저, DNase I 과민성 영역(DNase I hypersensitive: DHS) 및 비-DHS(non-DHS) 부위의 계층에서 계층화된 무작위 샘플링을 통해 Endo_CBE 데이터 세트를 짝지어진 하위 세트(paired subset)로 나누어 유사한 비율의 DHS/비-DHS 부위가 각 하위 세트에 할당되도록 하고, 각각을 Endo_CBE_1A 및 Endo_CBE_1B로 명명하였다. 상기 무작위 샘플링을 반복하여 Endo_CBE_2A, Endo_CBE_2B 등으로 명명된 4개의 데이터 세트를 추가로 생성하고, Endo_CBE_1A 데이터 세트 및 ENCODE(Encyclopedia of DNA elements)로부터 얻은 이진법의 염색질 접근성 정보를 사용하여 DeepCBE를 미세조정함으로써, 표적 서열 정보 및 염색질 접근성 둘 모두에 기초한 CBE-기반 유전자 편집 효율 및 결과 예측 모델인 DeepCBE-CA(Chromatin Accessibility)-1A를 개발하였다.

[0248] 다음으로, 테스트 데이터 세트로 다른 데이터 세트(Endo_CBE_1B)를 사용하여 DeepCBE-CA-1A를 평가하였다. 테스트 데이터 세트와 학습(training) 데이터 세트를 서로 바꾸어 미세조정 및 후속 테스트를 반복하고(예를 들어, 미세조정을 위한 학습 데이터 세트로 Endo_CBE_1B, 테스트 데이터 세트로 Endo_CBE_1A를 사용), 다른 4쌍의 데이터 세트를 사용하여 이를 반복하였다. 총 10회의 미세조정 및 후속 테스트의 결과, 이들 미세조정된 모델의 Spearman 상관 관계는 DeepCBE의 Spearman 상관 관계와 유사하여(도 6의 d), 염색질 접근성 정보를 고려한 미세조정은 DeepCBE의 정확도를 향상시키지 않음을 확인하였다. ABE 역시, DeepABE와 Endo_ABE를 사용하여 동일하게 총 10회의 테스트를 수행한 결과, 염색질 접근성 정보의 고려는 ABE에 대한 예측 정확도를 향상시키지 않았다.

[0249] **2-3. DeepABE 및 DeepCBE의 개발**

[0250] 다음으로, 염기 편집된 결과의 절대적 빈도를 예측하기 위해 실시예 2-1 및 2-2에서 개발한 ABE_efficiency와 ABE_proportion을, CBE_efficiency와 CBE_proportion을 결합하여 염기교정 효율 및 염기교정 유전자가위가 만

들 수 있는 모든 염기 편집 결과물들에 대한 예측 수행이 가능한 DeepABE 및 DeepCBE 모델을 생성하였다.

[0251] 그 결과, HT_ABE_Test 및 Endo_ABE_HEK293T로 테스트했을 때 DeepABE는 Spearman R=0.90, Pearson r=0.92(HT_ABE_Test) 및 Spearman R=0.86, Pearson r=0.80(Endo_ABE_HEK293T)의 높은 상관 계수에 도달하여 염기 편집의 결과 빈도 예측에 탁월한 성능을 나타내었다(도 6의 a). 유사하게, DeepCBE 역시 Spearman R=0.86, Pearson r=0.87(HT_CBE_Test) 및 Spearman R=0.83, Pearson r=0.71(Endo_CBE_HEK293T)의 높은 상관 계수에 도달하여 매우 우수한 성능을 나타냄을 확인하였다(도 6의 b).

[0252] 한편, 상기 ABE_efficiency, CBE_efficiency, ABE_proportion, 및 CBE_proportion 모델은 강력한 컨볼루션 신경망(convolutional neural networks: CNNs)을 사용하는 딥러닝 프레임워크를 기반으로 하며, 구체적으로 다음과 같이 개발되었다: (1) 표적 서열 및 이웃하는 서열을 함유하는 인풋(input) 서열을 4차원 이진(binary) 행렬로 변환; (2) 위치 가중치 매트릭스(position weight matrices)를 결정하기 위해, 3-nt의 긴 필터(filter)가 4차원 이진 매트릭스를 통해 이동; (3) 완전히 연결된 층(fully connected layers)에서, 추출된 특징이 가중치 합산(weighted sum), 수정된 선형 유닛(rectified linear unit: ReLU) 활성화 함수에 따라 합쳐짐; (4) 아웃풋 층(output layer)에서, 선형 회귀를 수행하고 각 표적 서열에 대한 활성(activity) 스코어 또는 결과 빈도를 예측. DeepABE 및 DeepCBE 스코어는 ABE_proportion 및 ABE_efficiency 또는 CBE_proportion 및 CBE_efficiency 각각의 스코어를 곱하여 간단히 얻었다(도 7).

[0253] **2-4. DeepABE 및 DeepCBE의 정확성 검증**

[0254] 생물학적 복제시료(biological replicate)를 사용하여 실시예 2에서 구축한 예측 모델 ABE_efficiency, ABE_proportion, DeepABE, CBE_efficiency, CBE_proportion 및 DeepCBE의 정확성을 검증하였다. 구체적으로, HCT116 세포, U2OS 세포 및 HEK293T 세포의 내인성 부위(endogenous sites)에서 ABE_efficiency, ABE_proportion, DeepABE, CBE_efficiency, CBE_proportion 및 DeepCBE의 성능을 평가하였다. 그 결과, ABE_proportion, CBE_proportion 뿐만 아니라 DeepABE 및 DeepCBE 모델 모두가 모든 세포 유형에 걸쳐 우수한 성능을 나타냄을 확인하였다(도 8 내지 11).

[0255] **실시예 3. 인간 점돌연변이 질환에 대한 예측 모델의 적용**

[0256] **3-1. 인간 점돌연변이 질환 데이터에 대한 적용**

[0257] 개발된 예측 모델과 ClinVar4에서 보고된 인간 점돌연변이 질환 데이터를 사용하여, 질환 모델링 및 염기교정 유전자가위의 인간 점돌연변이의 치료적 교정에 대한 염기 편집 효율 및 결과를 예측하였다. 질환 모델링을 위해, 적절한 거리에 PAM(NGG)이 있는 위치 3 내지 10의 편집가능한 윈도우를 사용하여 생성될 수 있는 병원성 및 유사병원성 점돌연변이를 탐색한 결과, ABE 및 CBE를 사용하여 이론적으로 생성될 수 있는 점돌연변이는 각각 2,917개 및 8,759개임을 확인하였다. 이들 점돌연변이 중 점돌연변이가 단 하나의 A 또는 C를 가지는 편집가능한 윈도우에 생기는 경우는 ABE의 경우 이중 24%(2,917개 중 691개), CBE의 경우 13%(8,759개 중 1,113개)로 나타났으며, 나머지 76% 및 87%의 점돌연변이는 각각 하나 이상의 A 또는 C를 포함하였다(도 12의 a).

[0258] 나아가, 염기교정 유전자가위를 사용하여 교정 가능한 병원성 및 유사병원성 점돌연변이를 탐색한 결과, 원칙적으로 8,930개 및 2,873개의 돌연변이가 ABE 및 CBE를 사용하여 야생형 서열로 각각 전환될 수 있음을 확인하였다. 이들 돌연변이 중, ABE의 경우 21%(8,930개 중 1,834개), CBE의 경우 12%(2,873개 중 336개)가 편집가능한 윈도우에서 단 하나의 A 또는 C를 가진 것으로 나타났으며, 이를 통해 대부분의 염기 편집 가능한 돌연변이는 다수의 표적 뉴클레오티드를 가진 윈도우에서 발생함을 확인하였다(도 12의 b).

[0259] 상기 결과를 통해, ABE로 만들 수 있는 인간 질환의 수는 총 2,917개, CBE로 만들 수 있는 인간 질환의 수는 총 8,759개, ABE를 사용하여 교정할 수 있는 인간 질환의 수는 총 8,930개, CBE를 사용하여 교정할 수 있는 인간 질환의 수는 총 2,873개임을 확인하였다.

[0260] **3-2. 인간 유도만능줄기세포에 대한 적용**

[0261] 인간 유도만능줄기세포(human induced pluripotent stem cells: iPSCs)와 개발된 예측 모델을 사용하여, 질환 모델링 및 염기교정 유전자가위의 인간 점돌연변이의 치료적 교정에 대한 염기 편집 효율 및 결과를 예측하였다. 그 결과, 질환 모델링과 치료 모두에서 예측된 교정효율과 측정된 교정결과 빈도 간에 유의미한 상관 관계가 있음을 확인하였다(도 12의 c, d).

[0262] **3-3. 염기교정 유전자가위의 효율 예측**

[0263] 상기 실시예 3-1 및 3-2에서 수행한 모델링을 바탕으로, 인간 점돌연변이 질환에 대한 염기교정 유전자가위의 염기교정 효율을 예측하여 효율에 따른 분포를 확인하였다(도 13). 관심있는 돌연변이를 포함하는 염기 편집 결과의 빈도가 5%보다 높을 때, 해당 모델링을 "효율적(efficient)"인 것으로 정의한다면, 편집가능한 윈도우 내 단일 A의 돌연변이 691개 중 639개(92%), 다수 A의 돌연변이 2,226개 중 1,225개(55%)가 효율적으로 모델링 되는 것으로 나타났다. 교정 가능한 점돌연변이의 경우, 염기 편집 결과 야생형 서열의 빈도가 5%보다 높을 때, 해당 모델링을 "효율적(efficient)"인 것으로 정의한다면, 편집가능한 윈도우 내 단일 A의 돌연변이 1,834개 중 1,728개(94%), 다수 A의 돌연변이 7,096개 중 4,038개(57%)가 야생형으로 전환되어 효율적으로 모델링됨을 확인하였다.

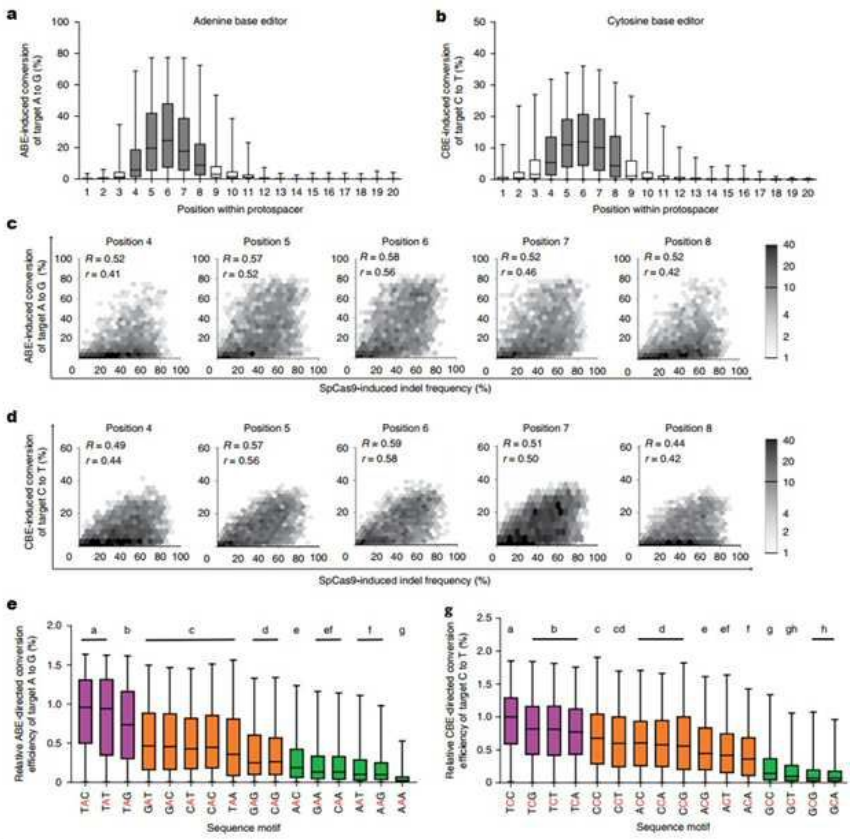
[0264] 상기 결과를 통해, 일 양상에 따른 예측 시스템을 사용하면 인간 점돌연변이 질환에 대한 염기교정 유전자가위의 염기교정 효율과 가능한 교정결과들의 빈도를 모두 예측할 수 있어, 최적의 염기교정 유전자가위를 선별 및 실제 질환에서 상기 선별된 염기교정 유전자가위가 치료효과를 나타낼지에 관한 1차적 결과를 제공할 수 있음을 확인하였다.

[0265] 특히, 염기교정 유전자가위의 경우 기존의 크리스퍼 유전자가위와 달리 효율 예측만으로는 유전자가위 선별에 부족함이 있고 효율뿐만 아니라 염기교정 결과도 예측할 필요가 있으므로, 일 양상에 따른 예측 시스템은 딥러닝을 통해 염기교정 유전자가위의 효율 및 다양한 염기교정 결과물의 빈도를 예측함으로써 안전한 교정이 가능한 유전자가위를 선별하고, 그에 따라 적합한 가이드 RNA를 설계하여, 유전질환의 치료에 유용하게 활용할 수 있을 것으로 기대된다.

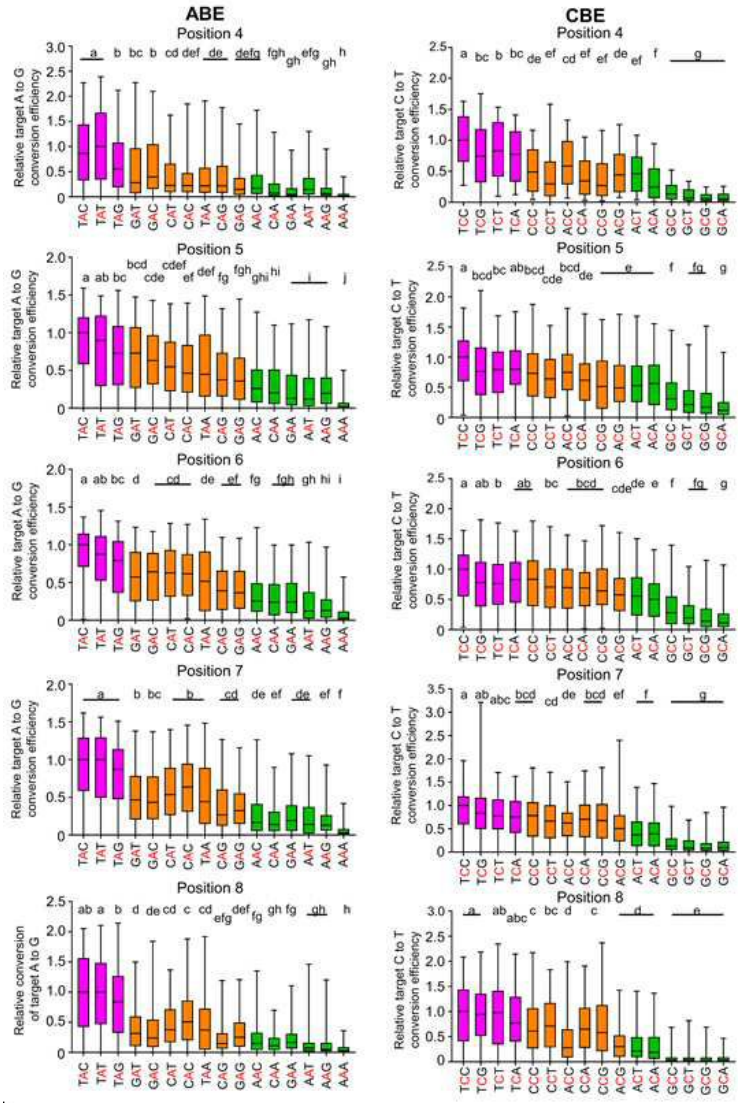
[0266] 전술한 설명은 예시를 위한 것이며, 본 발명이 속하는 기술분야의 통상의 지식을 가진 자는 본 발명의 기술적 사상이나 필수적인 특징을 변경하지 않고서 다른 구체적인 형태로 쉽게 변형이 가능하다는 것을 이해할 수 있을 것이다. 그러므로 이상에서 기술한 실시예들은 모든 면에서 예시적인 것이며 한정적이 아닌 것으로 이해해야만 한다.

도면

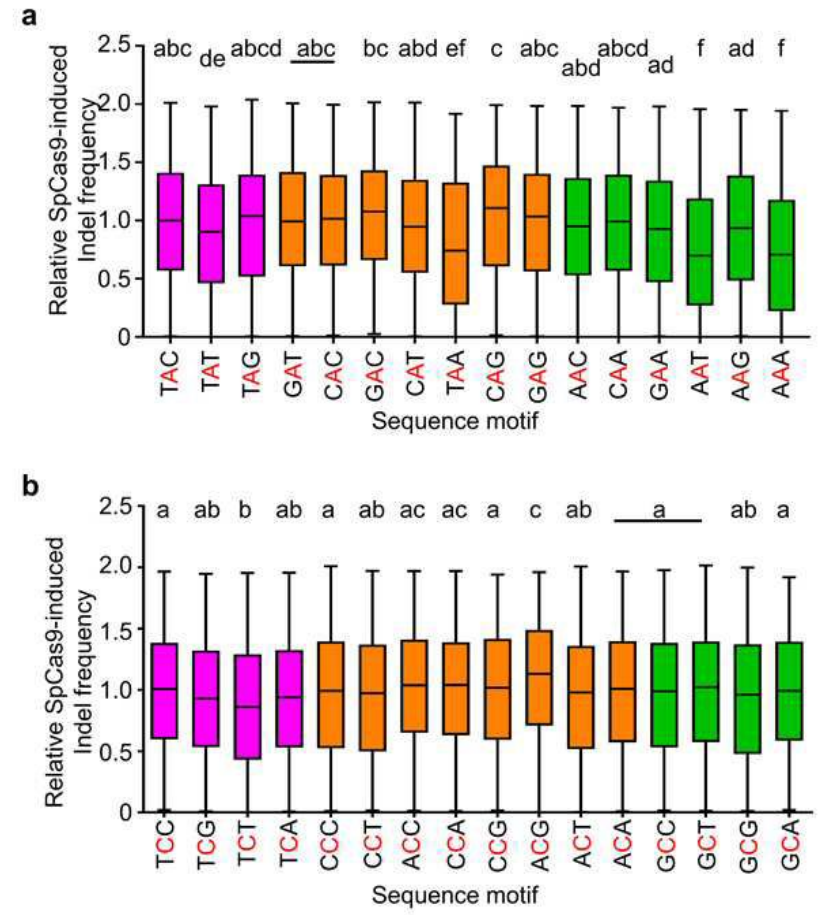
도면1



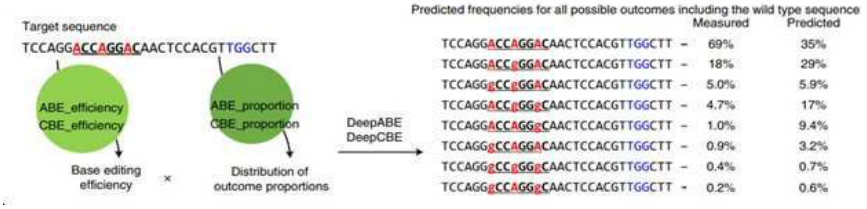
도면2


ABE
CBE
Position 4
Position 5
Position 6
Position 7
Position 8
Relative target A to G conversion efficiency
Relative target C to T conversion efficiency
Relative conversion of target A to G
Relative target C to T conversion efficiency
TAC TAT TAG GAT GAC CAT CAC TAA CAG GAG AAC CAA GAA AAT AAG AAA
TCC TCG TCT TCA CCC CCT ACC CCA CCG ACG ACT ACA GCC GCT GCG GCA

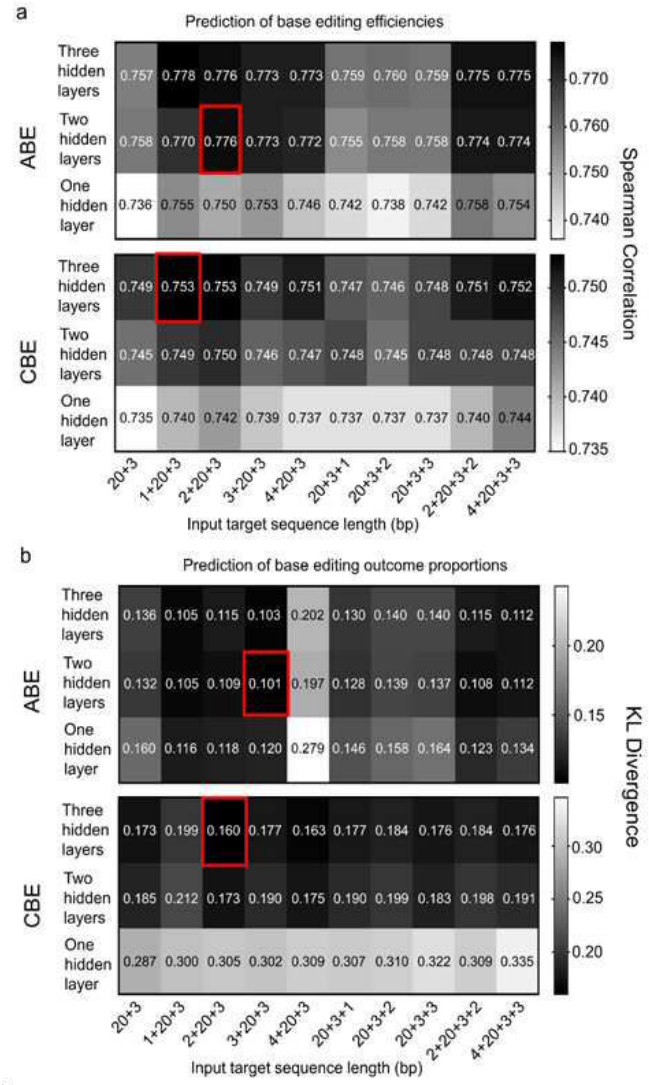
도면3


a
Relative SpCas9-induced Indel frequency
abc de abcd abc bc abd ef c abc abd abcd ad f ad f
TAC TAT TAG GAT CAC GAC CAT TAA CAG GAG AAC CAA GAA AAT AAG AAA
Sequence motif
b
Relative SpCas9-induced Indel frequency
a ab b ab a ab ac ac a c ab a ab a
TCC TCG TCT TCA CCC CCT ACC CCA CCG ACG ACT ACA GCC GCT GCG GCA
Sequence motif

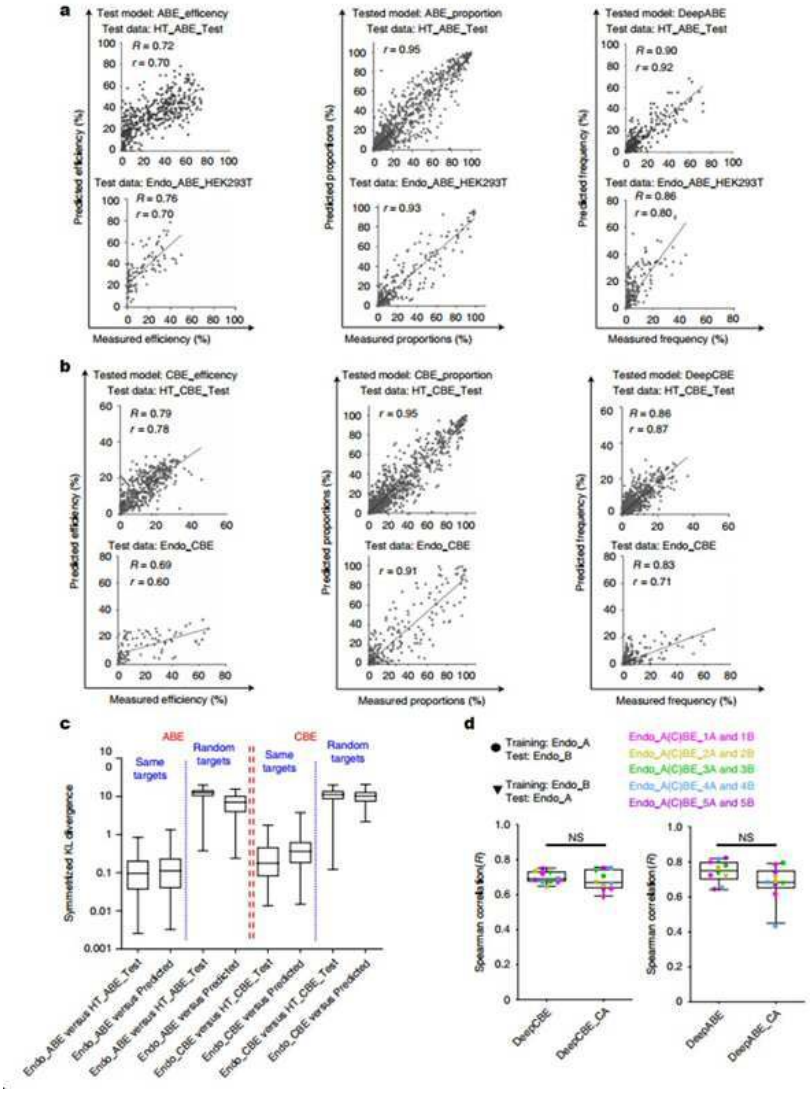
도면4


Target sequence
TCCAGGACCAGGACAACTCCACGTTGGCTT
ABE_efficiency
CBE_efficiency
Base editing efficiency
ABE_proportion
CBE_proportion
Distribution of outcome proportions
DeepABE
DeepCBE
Predicted frequencies for all possible outcomes including the wild type sequence
Measured
Predicted
TCCAGGACCAGGACAACTCCACGTTGGCTT - 69% 35%
TCCAGGACCgGGACAACTCCACGTTGGCTT - 18% 29%
TCCAGGgCCgGGACAACTCCACGTTGGCTT - 5.0% 5.9%
TCCAGGACCgGGgCAACTCCACGTTGGCTT - 4.7% 17%
TCCAGGACCAGGgCAACTCCACGTTGGCTT - 1.0% 9.4%
TCCAGGgCCAGGACAACTCCACGTTGGCTT - 0.9% 3.2%
TCCAGGgCCgGGgCAACTCCACGTTGGCTT - 0.4% 0.7%
TCCAGGgCCAGGgCAACTCCACGTTGGCTT - 0.2% 0.6%

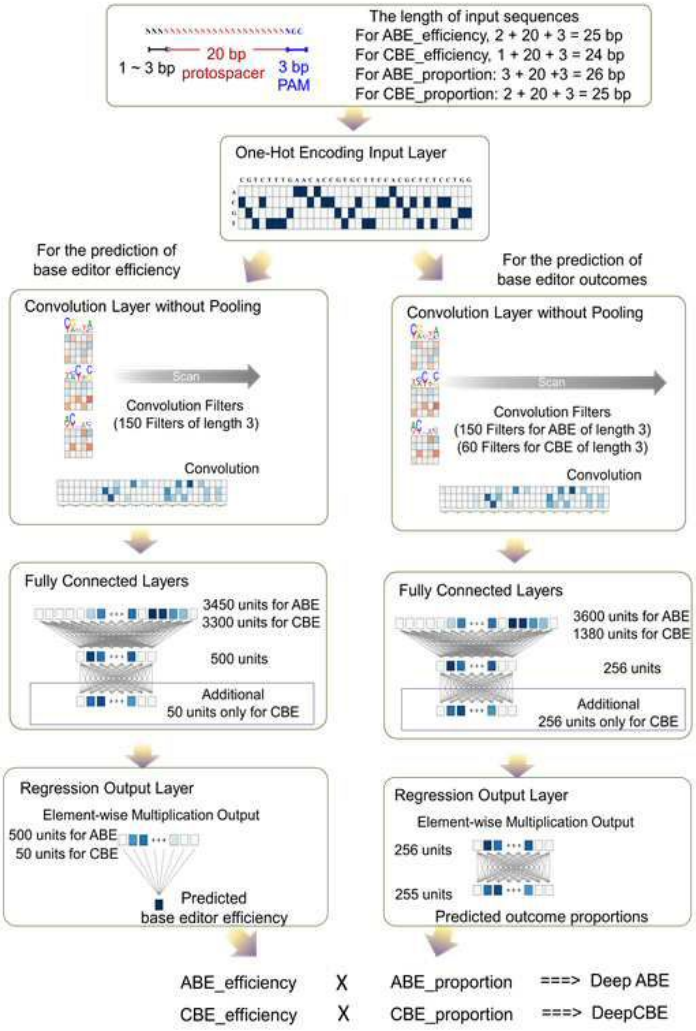
도면5


a
Prediction of base editing efficiencies
ABE
Three hidden layers
Two hidden layers
One hidden layer
0.757 0.778 0.776 0.773 0.773 0.759 0.760 0.759 0.775 0.775
0.758 0.770 0.776 0.773 0.772 0.755 0.758 0.758 0.774 0.774
0.736 0.755 0.750 0.753 0.746 0.742 0.738 0.742 0.758 0.754
CBE
0.749 0.753 0.753 0.749 0.751 0.747 0.746 0.748 0.751 0.752
0.745 0.749 0.750 0.746 0.747 0.748 0.745 0.748 0.748 0.748
0.735 0.740 0.742 0.739 0.737 0.737 0.737 0.737 0.740 0.744
0.770
0.760
0.750
0.740
0.750
0.745
0.740
0.735
Spearman Correlation
20+3
1+20+3
2+20+3
3+20+3
4+20+3
20+3+1
20+3+2
20+3+3
2+20+3+2
4+20+3+3
Input target sequence length (bp)
b
Prediction of base editing outcome proportions
ABE
Three hidden layers
Two hidden layers
One hidden layer
0.136 0.105 0.115 0.103 0.202 0.130 0.140 0.140 0.115 0.112
0.132 0.105 0.109 0.101 0.197 0.128 0.139 0.137 0.108 0.112
0.160 0.116 0.118 0.120 0.279 0.146 0.158 0.164 0.123 0.134
CBE
0.173 0.199 0.160 0.177 0.163 0.177 0.184 0.176 0.184 0.176
0.185 0.212 0.173 0.190 0.175 0.190 0.199 0.183 0.198 0.191
0.287 0.300 0.305 0.302 0.309 0.307 0.310 0.322 0.309 0.335
0.20
0.15
0.30
0.25
0.20
KL Divergence
20+3
1+20+3
2+20+3
3+20+3
4+20+3
20+3+1
20+3+2
20+3+3
2+20+3+2
4+20+3+3
Input target sequence length (bp)

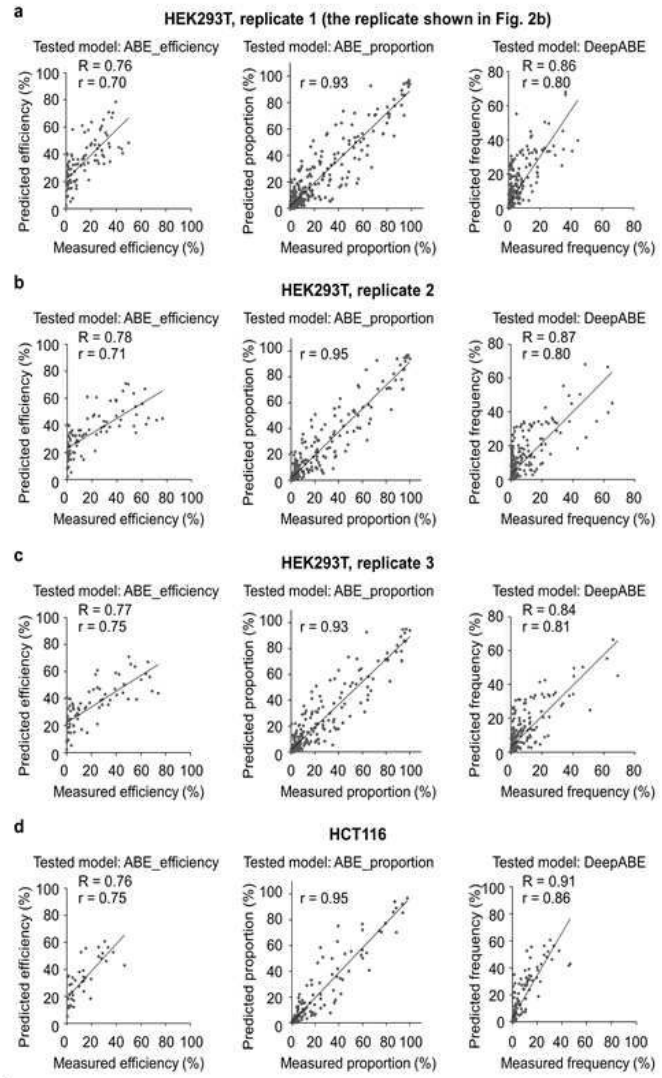
도면6


a
Test model: ABE_efficency
Test data: HT_ABE_Test
R = 0.72
r = 0.70
Test data: Endo_ABE_HEK293T
R = 0.76
r = 0.70
Predicted efficiency (%)
Measured efficiency (%)
Tested model: ABE_proportion
Test data: HT_ABE_Test
r = 0.95
Test data: Endo_ABE_HEK293T
r = 0.93
Predicted proportions (%)
Measured proportions (%)
Tested model: DeepABE
Test data: HT_ABE_Test
R = 0.90
r = 0.92
Test data: Endo_ABE_HEK293T
R = 0.86
r = 0.80
Predicted frequency (%)
Measured frequency (%)
b
Tested model: CBE_efficency
Test data: HT_CBE_Test
R = 0.79
r = 0.78
Test data: Endo_CBE
R = 0.69
r = 0.60
Predicted efficiency (%)
Measured efficiency (%)
Tested model: CBE_proportion
Test data: HT_CBE_Test
r = 0.95
Test data: Endo_CBE
r = 0.91
Predicted proportions (%)
Measured proportions (%)
Tested model: DeepCBE
Test data: HT_CBE_Test
R = 0.86
r = 0.87
Test data: Endo_CBE
R = 0.83
r = 0.71
Predicted frequency (%)
Measured frequency (%)
c
ABE
CBE
Same targets
Random targets
Same targets
Random targets
Symmetrized KL divergence
Endo_ABE versus HT_ABE_Test
Endo_ABE versus Predicted
Endo_ABE versus HT_ABE_Test
Endo_ABE versus Predicted
Endo_CBE versus HT_CBE_Test
Endo_CBE versus Predicted
Endo_CBE versus HT_CBE_Test
Endo_CBE versus Predicted
d
Training: Endo_A
Test: Endo_B
Training: Endo_B
Test: Endo_A
Endo_A(C)BE_1A and 1B
Endo_A(C)BE_2A and 2B
Endo_A(C)BE_3A and 3B
Endo_A(C)BE_4A and 4B
Endo_A(C)BE_5A and 5B
NS
NS
Spearman correlation (R)
DeepCBE
DeepCBE_CA
DeepABE
DeepABE_CA

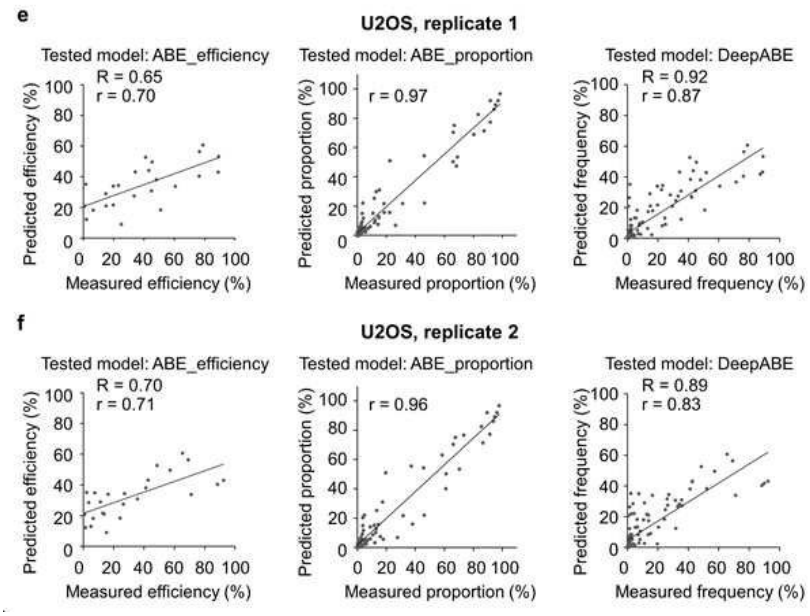
도면7


The length of input sequences
For ABE_efficiency, 2 + 20 + 3 = 25 bp
For CBE_efficiency, 1 + 20 + 3 = 24 bp
For ABE_proportion: 3 + 20 +3 = 26 bp
For CBE_proportion: 2 + 20 + 3 = 25 bp
1 ~ 3 bp
20 bp protospacer
3 bp PAM
One-Hot Encoding Input Layer
For the prediction of base editor efficiency
For the prediction of base editor outcomes
Convolution Layer without Pooling
Scan
Convolution Filters (150 Filters of length 3)
Convolution
Convolution Layer without Pooling
Scan
Convolution Filters (150 Filters for ABE of length 3) (60 Filters for CBE of length 3)
Convolution
Fully Connected Layers
3450 units for ABE
3300 units for CBE
500 units
Additional 50 units only for CBE
Fully Connected Layers
3600 units for ABE
1380 units for CBE
256 units
Additional 256 units only for CBE
Regression Output Layer
Element-wise Multiplication Output
500 units for ABE
50 units for CBE
Predicted base editor efficiency
Regression Output Layer
Element-wise Multiplication Output
256 units
255 units
Predicted outcome proportions
ABE_efficiency X ABE_proportion ===> Deep ABE
CBE_efficiency X CBE_proportion ===> DeepCBE

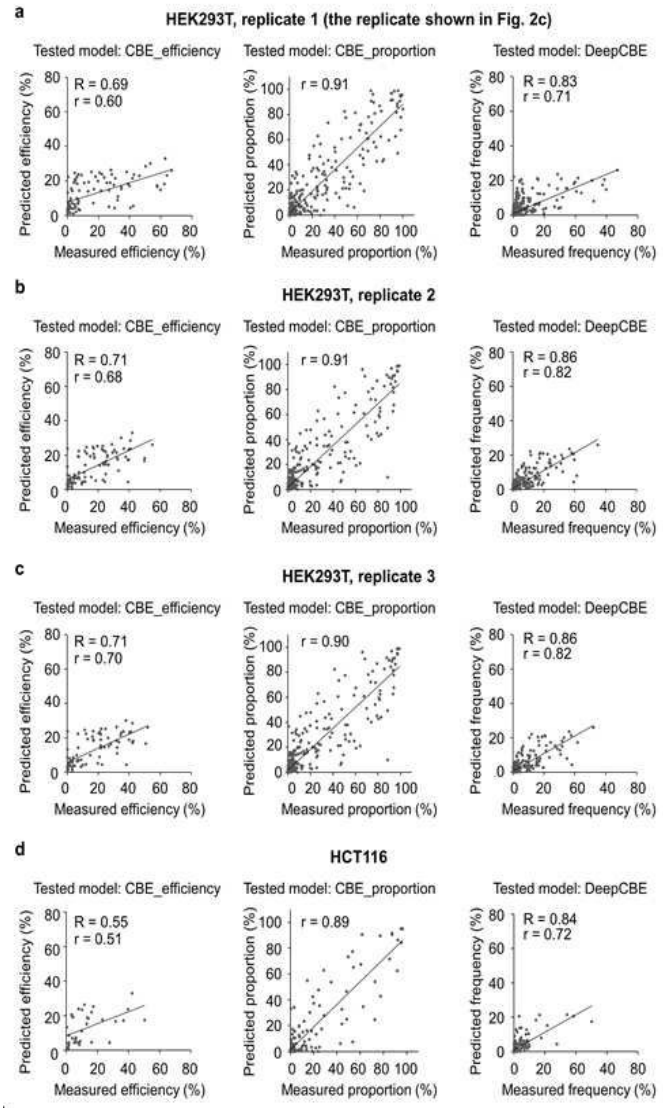
도면8


a
HEK293T, replicate 1 (the replicate shown in Fig. 2b)
Tested model: ABE_efficiency
R = 0.76
r = 0.70
Predicted efficiency (%)
Measured efficiency (%)
Tested model: ABE_proportion
r = 0.93
Predicted proportion (%)
Measured proportion (%)
Tested model: DeepABE
R = 0.86
r = 0.80
Predicted frequency (%)
Measured frequency (%)
b
HEK293T, replicate 2
Tested model: ABE_efficiency
R = 0.78
r = 0.71
Tested model: ABE_proportion
r = 0.95
Tested model: DeepABE
R = 0.87
r = 0.80
c
HEK293T, replicate 3
Tested model: ABE_efficiency
R = 0.77
r = 0.75
Tested model: ABE_proportion
r = 0.93
Tested model: DeepABE
R = 0.84
r = 0.81
d
HCT116
Tested model: ABE_efficiency
R = 0.76
r = 0.75
Tested model: ABE_proportion
r = 0.95
Tested model: DeepABE
R = 0.91
r = 0.86

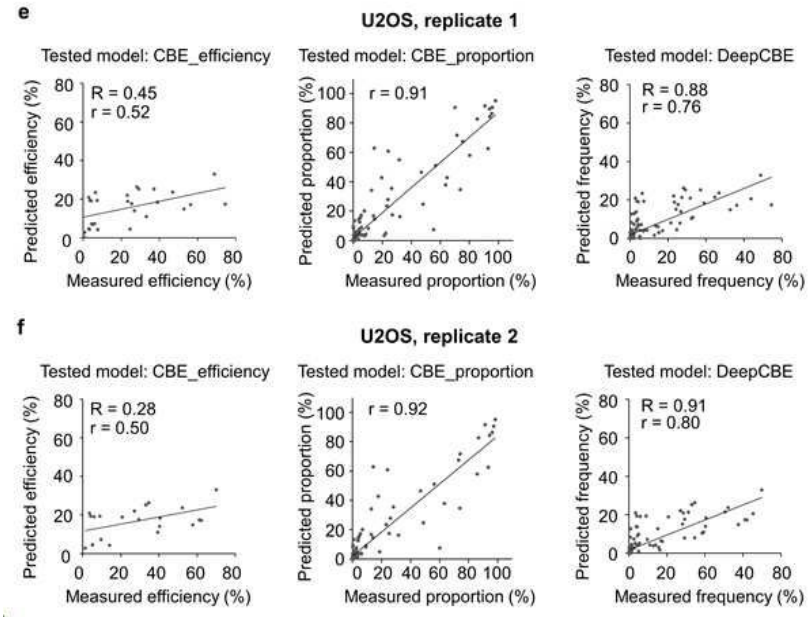
도면9


e
U2OS, replicate 1
Tested model: ABE_efficiency
R = 0.65
r = 0.70
Predicted efficiency (%)
Measured efficiency (%)
Tested model: ABE_proportion
r = 0.97
Predicted proportion (%)
Measured proportion (%)
Tested model: DeepABE
R = 0.92
r = 0.87
Predicted frequency (%)
Measured frequency (%)
f
U2OS, replicate 2
Tested model: ABE_efficiency
R = 0.70
r = 0.71
Tested model: ABE_proportion
r = 0.96
Tested model: DeepABE
R = 0.89
r = 0.83

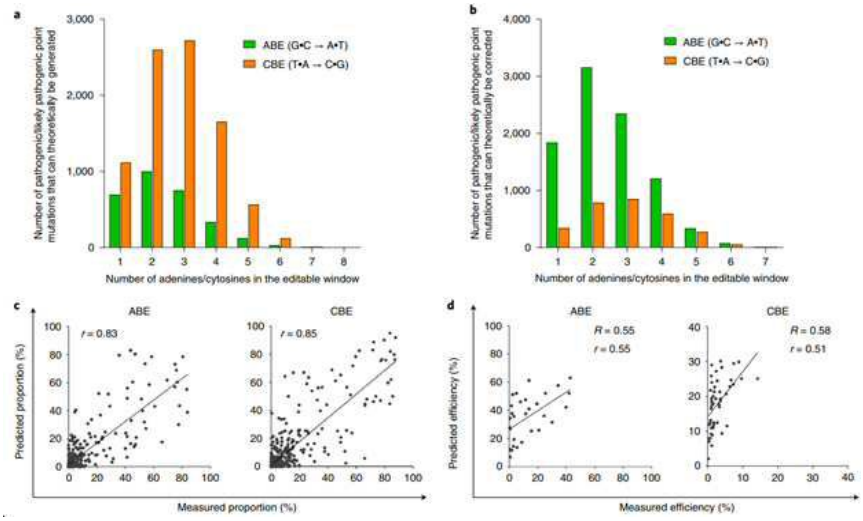
도면10


a
HEK293T, replicate 1 (the replicate shown in Fig. 2c)
Tested model: CBE_efficiency
Predicted efficiency (%)
R = 0.69
r = 0.60
Measured efficiency (%)
Tested model: CBE_proportion
Predicted proportion (%)
r = 0.91
Measured proportion (%)
Tested model: DeepCBE
Predicted frequency (%)
R = 0.83
r = 0.71
Measured frequency (%)
b
HEK293T, replicate 2
Tested model: CBE_efficiency
R = 0.71
r = 0.68
Tested model: CBE_proportion
r = 0.91
Tested model: DeepCBE
R = 0.86
r = 0.82
c
HEK293T, replicate 3
Tested model: CBE_efficiency
R = 0.71
r = 0.70
Tested model: CBE_proportion
r = 0.90
Tested model: DeepCBE
R = 0.86
r = 0.82
d
HCT116
Tested model: CBE_efficiency
R = 0.55
r = 0.51
Tested model: CBE_proportion
r = 0.89
Tested model: DeepCBE
R = 0.84
r = 0.72

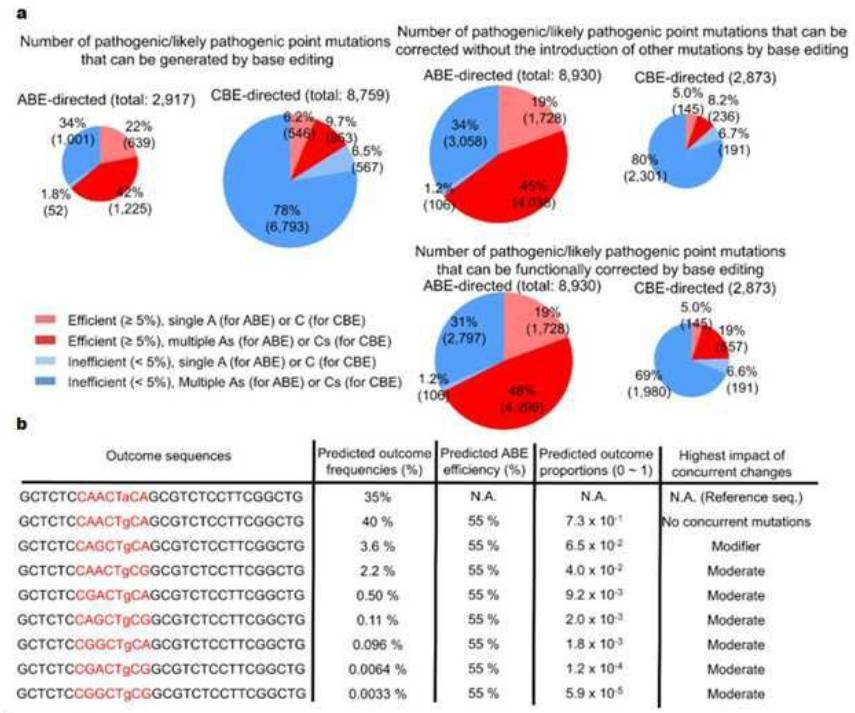
도면11


e
U2OS, replicate 1
Tested model: CBE_efficiency
Tested model: CBE_proportion
Tested model: DeepCBE
Predicted efficiency (%)
Predicted proportion (%)
Predicted frequency (%)
R = 0.45
r = 0.52
r = 0.91
R = 0.88
r = 0.76
Measured efficiency (%)
Measured proportion (%)
Measured frequency (%)
f
U2OS, replicate 2
Tested model: CBE_efficiency
Tested model: CBE_proportion
Tested model: DeepCBE
Predicted efficiency (%)
Predicted proportion (%)
Predicted frequency (%)
R = 0.28
r = 0.50
r = 0.92
R = 0.91
r = 0.80
Measured efficiency (%)
Measured proportion (%)
Measured frequency (%)

도면12


a
Number of pathogenic/likely pathogenic point mutations that can theoretically be generated
ABE (G•C → A•T)
CBE (T•A → C•G)
Number of adenines/cytosines in the editable window
b
Number of pathogenic/likely pathogenic point mutations that can theoretically be corrected
ABE (G•C → A•T)
CBE (T•A → C•G)
Number of adenines/cytosines in the editable window
c
ABE
CBE
r = 0.83
r = 0.85
Predicted proportion (%)
Measured proportion (%)
d
ABE
CBE
R = 0.55
r = 0.55
R = 0.58
r = 0.51
Predicted efficiency (%)
Measured efficiency (%)

도면13



b

Outcome sequences	Predicted outcome frequencies (%)	Predicted ABE efficiency (%)	Predicted outcome proportions (0 ~ 1)	Highest impact of concurrent changes
GCTCTCCAACTaCAGCGTCTCCTTCGGCTG	35%	N.A.	N.A.	N.A. (Reference seq.)
GCTCTCCAACTgCAGCGTCTCCTTCGGCTG	40 %	55 %	7.3×10^{-1}	No concurrent mutations
GCTCTCCAGCTgCAGCGTCTCCTTCGGCTG	3.6 %	55 %	6.5×10^{-2}	Modifier
GCTCTCCAACTgCGGCGTCTCCTTCGGCTG	2.2 %	55 %	4.0×10^{-2}	Moderate
GCTCTCCGACTgCAGCGTCTCCTTCGGCTG	0.50 %	55 %	9.2×10^{-3}	Moderate
GCTCTCCAGCTgCGGCGTCTCCTTCGGCTG	0.11 %	55 %	2.0×10^{-3}	Moderate
GCTCTCCGGCTgCAGCGTCTCCTTCGGCTG	0.096 %	55 %	1.8×10^{-3}	Moderate
GCTCTCCGACTgCGGCGTCTCCTTCGGCTG	0.0064 %	55 %	1.2×10^{-4}	Moderate
GCTCTCCGGCTgCGGCGTCTCCTTCGGCTG	0.0033 %	55 %	5.9×10^{-5}	Moderate

서 열 목 록

<110> Industry-Academic Cooperation Foundation, Yonsei University

<120> A system for predicting base-editing efficiency and outcome product frequencies of base editors

<130> PN134603KR

<160> 28

<170> KoPatentIn 3.0

<210> 1

<211> 55

<212> DNA

<213> Artificial Sequence

<220><223> primer_F for oligonucleotide pool amplification

<400> 1

ttgaaagtat ttcgatttct tggctttata tatcttgtgg aaaggacgaa acacc 55

<210> 2

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_R for oligonucleotide pool amplification

<400> 2

gagtaagctg accgctgaag tacaagtggt agagtagaga tctagttacg ccaagct 57

<210> 3

<211> 56

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_1 for 1st PCR

<400> 3

acactctttc cctacacgac gctcttccga tctcttgaaa aagtggcacc gagtcg 56

<210> 4

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_2 for 1st PCR

<400> 4

acactctttc cctacacgac gctcttccga tcttcttgaa aaagtggcac cgagtcg 57

<210> 5

<211> 58

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_3 for 1st PCR

<400> 5

acactctttc cctacacgac gctcttccga tctcgcttga aaaagtggca ccgagtcg 58

<210> 6

<211> 61

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_1 for 1st PCR

<400> 6

gtgactggag ttcagacgtg tgctcttccg atctttaagt cgagtaagct gaccgctgaa 60

g 61

<210> 7

<211> 62

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_2 for 1st PCR

<400> 7

gtgactggag ttcagacgtg tgctcttccg atctattaag tcgagtaagc tgaccgctga 60

ag 62

<210> 8

<211> 63

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_3 for 1st PCR

<400> 8

gtgactggag ttcagacgtg tgctcttccg atcttattaa gtcgagtaag ctgaccgctg 60

aag 63

<210> 9

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_1 for 2nd PCR

<400> 9

aatgatacgg cgaccaccga gatctacact atagcctaca ctctttccct acacgac 57

<210> 10

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_2 for 2nd PCR

<400> 10

aatgatacgg cgaccaccga gatctacaca tagaggcaca ctctttccct acacgac 57

<210> 11

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_3 for 2nd PCR

<400> 11

aatgatacgg cgaccaccga gatctacacc ctatcctaca ctctttccct acacgac 57

<210> 12

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_4 for 2nd PCR

<400> 12

aatgatacgg cgaccaccga gatctacacg gctctgaaca ctctttccct acacgac 57

<210> 13

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_5 for 2nd PCR

<400> 13

aatgatacgg cgaccaccga gatctacaca ggcgaagaca ctctttccct acacgac 57

<210> 14

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_6 for 2nd PCR

<400> 14

aatgatacgg cgaccaccga gatctacact aatcttaaca ctctttccct acacgac 57

<210> 15

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_7 for 2nd PCR

<400> 15

aatgatacgg cgaccaccga gatctacacc aggacgtaca ctctttccct acacgac 57

<210> 16

<211> 57

<212> DNA

<213> Artificial Sequence

<220><223> primer_F_8 for 2nd PCR

<400

> 16

aatgatacgg cgaccaccga gatctacacg tactgacaca ctctttccct acacgac 57

<210> 17

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_1 for 2nd PCR

<400> 17

caagcagaag acggcatacg agatcgagta atgtgactgg agttcagacg tgt 53

<210> 18

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_2 for 2nd PCR

<400> 18

caagcagaag acggcatacg agattctccg gagtgactgg agttcagacg tgt 53

<210> 19

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_3 for 2nd PCR

<400> 19

caagcagaag acggcatacg agataatgag cggtgactgg agttcagacg tgt 53

<210> 20

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_4 for 2nd PCR

<400> 20

caagcagaag acggcatacg agatggaatc tcgtgactgg agttcagacg tgt 53

<210> 21

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_5 for 2nd PCR

<400

> 21

caagcagaag acggcatacg agatttctga atgtgactgg agttcagacg tgt 53

<210> 22

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_6 for 2nd PCR

<400> 22

caagcagaag acggcatacg agatacgaat tcgtgactgg agttcagacg tgt 53

<210> 23

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_7 for 2nd PCR

<400> 23

caagcagaag acggcatacg agatagcttc aggtgactgg agttcagacg tgt 53

<210> 24

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_8 for 2nd PCR

<400> 24

caagcagaag acggcatacg agatgcgcat tagtgactgg agttcagacg tgt 53

<210> 25

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_9 for 2nd PCR

<400> 25

caagcagaag acggcatacg agatcatagc cggtgactgg agttcagacg tgt 53

<210> 26

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_10 for 2nd PCR

<

400> 26

caagcagaag acggcatacg agatttcgcg gagtgactgg agttcagacg tgt 53

<210> 27

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_11 for 2nd PCR

<400> 27

caagcagaag acggcatacg agatgcgcga gagtgactgg agttcagacg tgt 53

<210> 28

<211> 53

<212> DNA

<213> Artificial Sequence

<220><223> primer_R_12 for 2nd PCR

<400> 28

caagcagaag acggcatacg agatctatcg ctgtgactgg agttcagacg tgt 53