(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0014711
(43) 공개일자 2022년02월07일

(51) 국제특허분류(Int. Cl.)
G16B 5/00 (2019.01) *G06N 20/00* (2019.01)
G06N 3/08 (2006.01) *G16B 30/10* (2019.01)
G16B 35/10 (2019.01) *G16B 50/00* (2019.01)
(52) CPC특허분류
G16B 5/00 (2019.02)
G06N 20/00 (2021.08)
(21) 출원번호 10-2020-0094684
(22) 출원일자 2020년07월29일
심사청구일자 2020년07월29일

(71) 출원인
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
김형범
서울특별시 마포구 토정로18길 11, 107동 1702호 (현석동, 래미안웰스트림)
김희권
서울특별시 종로구 송월길 130, 104동 1901호(행촌동, 경희궁자이 1단지)
유구상
인천광역시 남동구 구월로 192, 1105동 202호(구월동, 힐스테이트롯데캐슬골드1단지아파트)
(74) 대리인
리앤목특허법인

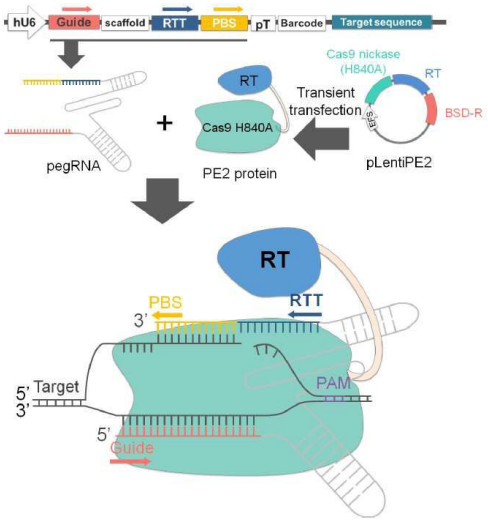
전체 청구항 수 : 총 20 항

(54) 발명의 명칭 **딥러닝을 이용한 프라임에디팅 효율 예측 시스템 및 방법**

(57) 요 약

딥러닝을 이용한 프라임에디팅 효율 예측 시스템, 상기 시스템을 구축하는 방법, 상기 시스템을 이용한 프라임에디팅 효율 예측 방법 및 상기 방법을 컴퓨터로 실행하기 위한 프로그램이 기록된 컴퓨터 판독가능 기록매체를 제공한다.

대 표 도 - 도1



(52) CPC특허분류
G06N 3/08 (2013.01)
G16B 30/10 (2019.02)
G16B 35/10 (2019.02)
G16B 50/00 (2019.02)

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

프라임에디터(Prime editor)의 프라임에디팅(Prime editing) 효율에 대한 데이터를 입력받는 정보 입력부;

상기 정보 입력부에서 입력 받은 데이터를 이용하여 프라임에디팅 효율에 영향을 미치는 특징과 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하여 프라임에디팅 효율 예측 모델을 생성하는 예측 모델 생성부;

프라임에디팅의 후보 표적 서열을 입력받는 후보 서열 입력부; 및

상기 후보 서열 입력부에 입력된 후보 표적 서열을 상기 예측 모델 생성부에서 생성된 효율 예측 모델에 적용하여 프라임에디팅 효율을 예측하는 효율 예측부를 포함하는,

딥러닝을 이용한 프라임에디팅(Prime editing) 효율 예측 시스템.

청구항 2

청구항 1에 있어서, 상기 프라임에디터는 프라임에디터2인 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 3

청구항 1에 있어서, 상기 프라임에디팅 효율에 대한 데이터는 표적 서열 내에서 의도하지 않은 돌연변이 없이 프라임에디터 및 pegRNA에 의해 유도된 편집이 발생한 비율로 나타내지는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 4

청구항 1에 있어서, 상기 프라임에디팅 효율에 대한 데이터는,

pegRNA를 암호화하는 뉴클레오티드 서열 및 상기 pegRNA가 목적하는 표적 뉴클레오티드 서열을 포함하는 올리고뉴클레오티드를 포함하는 세포 라이브러리에 프라임에디터(Prime editor)를 도입하는 단계;

상기 프라임에디터가 도입된 세포 라이브러리로부터 수득한 DNA를 이용하여 딥시퀀싱을 수행하는 단계; 및

상기 딥시퀀싱으로 수득한 데이터로부터 프라임에디팅 효율을 분석하는 단계를 포함하는 방법을 수행하여 수득된 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 5

청구항 4에 있어서, 상기 올리고뉴클레오티드는 바코드 서열을 더 포함하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 6

청구항 1에 있어서, 상기 프라임에디팅 효율에 영향을 미치는 특징은 pegRNA 및 표적 서열 정보로부터 추출된 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 7

청구항 6에 있어서, 상기 pegRNA 및 표적 서열 정보는 역전사효소(reverse transcriptase, RT) 주형 서열 정보, PBS(primer binding site) 서열 정보, 및 표적 서열 정보 중 어느 하나 이상을 포함하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 8

청구항 1에 있어서, 상기 예측 모델 생성부는 pegRNA 및 표적 서열 정보로부터 프라임에디팅 효율에 영향을 미

치는 특징을 추출하는 특징 추출 모듈을 포함하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 9

청구항 1에 있어서, 상기 예측 모델 생성부는 컨볼루션 신경망(convolutional neural network. CNN) 또는 다층 퍼셉트론(multilayer percentron, MLP)을 기반으로 하여 딥러닝을 수행하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 10

청구항 1에 있어서, 상기 후보 표적 서열은 PAM (protospacer adjacent motif), 및 프로토스페이서 서열을 포함하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 11

청구항 1에 있어서, 상기 효율 예측부는 프라임에디터 및 pegRNA에 의한 후보 표적 서열의 프라임에디팅 효율을 예측하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 12

청구항 1에 있어서, 효율 예측부에서 예측된 프라임에디팅 효율을 출력하는 출력부를 더 포함하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템.

청구항 13

프라임에디터의 프라임에디팅 효율 데이터 세트를 수득하는 단계; 및

상기 효율 데이터 세트를 이용하여 프라임에디팅 효율에 영향을 미치는 특징과 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하여 프라임에디팅 효율 예측 모델을 생성하는 단계를 포함하는,

딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법.

청구항 14

청구항 13에 있어서, 상기 효율 데이터 세트를 수득하는 단계는,

pegRNA를 암호화하는 뉴클레오티드 서열 및 상기 pegRNA가 목적하는 표적 뉴클레오티드 서열을 포함하는 올리고 뉴클레오티드를 포함하는 세포 라이브러리에 프라임에디터를 도입하는 단계;

상기 프라임에디터가 도입된 세포 라이브러리로부터 수득한 DNA를 이용하여 딥시퀀싱을 수행하는 단계; 및

상기 딥시퀀싱으로 수득한 데이터로부터 프라임에디팅 효율을 분석하는 단계를 포함하는 것인,

딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법.

청구항 15

청구항 13에 있어서, 상기 프라임에디팅 효율은 표적 서열 내에서 의도하지 않은 돌연변이 없이 프라임에디터 및 pegRNA에 의해 유도된 편집이 발생한 비율로 계산되는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법.

청구항 16

청구항 13에 있어서, 상기 프라임에디팅 효율에 영향을 미치는 특징은 pegRNA 및 표적 서열 정보로부터 추출된 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법.

청구항 17

청구항 16에 있어서, 상기 pegRNA 및 표적 서열 정보는 RT 주형 서열 정보, PBS 서열 정보, 및 표적 서열 정보 중 어느 하나 이상을 포함하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법.

청구항 18

청구항 13에 있어서, 상기 예측 모델을 생성하는 단계에서, 컨볼루션 신경망(convolutional neural network. CNN) 또는 다층 퍼셉트론(multilayer percentron, MLP)을 기반으로 하여 딥러닝을 수행하는 것인, 딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법.

청구항 19

프라임에디팅의 후보 표적 서열을 설계하는 단계; 및

상기 설계된 후보 표적 서열을 청구항 1 내지 12 중 어느 한 항의 효율 예측 시스템에 적용하여 프라임에디팅 효율을 예측하는 단계를 포함하는,

프라임에디팅 효율 예측 방법.

청구항 20

청구항 19에 따른 방법을 컴퓨터로 실행하기 위한 프로그램이 기록된 컴퓨터 판독가능 기록매체.

발명의 설명

기 술 분 야

[0001] 딥러닝을 이용한 프라임에디팅 효율 예측 시스템, 상기 시스템을 구축하는 방법, 상기 시스템을 이용한 프라임에디팅 효율 예측 방법 및 상기 방법을 컴퓨터로 실행하기 위한 프로그램이 기록된 컴퓨터 판독가능 기록매체에 관한 것이다.

배 경 기 술

[0002] 프라임에디팅(Prime Editing)은 donor DNA 또는 이중가닥 나누기(double-strand breaks, DSBs) 없이, 거의 모든 크기의 유전자 변화를 도입할 수 있는 혁신적인 신규 게놈 편집 방법이다(Anzalone, A.V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149-157 (2019)). 이러한 변화에는 삽입, 결실, 및 모든 가능한 12가지 점 돌연변이뿐만 아니라 이러한 변화들의 조합을 포함한다.

[0003] 프라임에디터(Prime editor, PE)는 기본적으로 Cas9 nickase-reverse transcriptase (RT) 융합 단백질 및 프라임에디팅 가이드 RNA(prime editing guide RNA, pegRNA)로 구성되며; pegRNA는 표적 서열을 인식하는 가이드 서열, tracrRNA 스캐폴드 서열, 역전사 개시에 필요한 프라이머 결합 부위(primer binding site, PBS), 및 원하는 유전적 변화를 포함하며 표적 서열에 상동성인 RT 주형(RT template)을 포함한다. 4가지 유형의 프라임에디터가 개발되었다: PE1, PE2, PE3, 및 PE3b.

[0004] 프라임에디팅은 다양한 조건에 따라 편집 효율이 크게 달라질 수 있다. 프라임에디팅 효율에 영향을 미치는 인자에 대해 일부 연구가 이루어지고 있으나, 아직 초기 단계에 있다.

[0005] 따라서, 프라임에디팅 효율에 영향을 미치는 인자를 식별하고, 주어진 표적 서열에서의 프라임에디팅 활성을 예측하는 계산 모델의 개발은 프라임에디팅을 크게 촉진할 것이다.

발명의 내용

해결하려는 과제

[0006] 딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 제공한다.

[0007] 딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법을 제공한다.

[0008] 상기 효율 예측 시스템을 이용한 프라임에디팅 효율 예측 방법을 제공한다.

[0009] 상기 방법을 컴퓨터로 실행하기 위한 프로그램이 기록된 컴퓨터 판독가능 기록매체를 제공한다.

과제의 해결 수단

[0010] 일 양상은 딥러닝을 이용한 프라임에디팅(Prime editing) 효율 예측 시스템을 제공한다.

[0011] 상기 딥러닝을 이용한 프라임에디팅 효율 예측 시스템은

[0012] 프라임에디터의 프라임에디팅 효율에 대한 데이터를 입력받는 정보 입력부;

[0013] 상기 정보 입력부에서 입력 받은 데이터를 이용하여 프라임에디팅 효율에 영향을 미치는 특징과 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하여 프라임에디팅 효율 예측 모델을 생성하는 예측 모델 생성부;

[0014] 프라임에디팅의 후보 표적 서열을 입력받는 후보 서열 입력부; 및

[0015] 상기 후보 서열 입력부에 입력된 후보 표적 서열을 상기 예측 모델 생성부에서 생성된 효율 예측 모델에 적용하여 프라임에디팅 효율을 예측하는 효율 예측부를 포함한다.

[0017] 본 발명자들은 고처리량(high-throughput) 실험을 통해, 54,836 쌍의 pegRNA 암호화 서열 및 상응하는 표적 서열을 사용하여 프라임에디팅 효율 데이터 세트를 구성하였고, 이를 이용하여 프라임에디팅 효율과 관련된 특징을 추출하였으며, 주어진 표적 서열에서 프라임에디팅 효율을 예측하는 시스템을 구축하였다.

[0019] 상기 프라임에디팅 효율 예측 시스템은 프라임에디터(Prime editor)의 프라임에디팅(Prime editing) 효율에 대한 데이터를 입력받는 정보 입력부를 포함한다.

[0020] “프라임에디팅(Prime editing)”은 4세대 유전자 가위에 의한, DNA 이중가닥 절단 없이 한 가닥의 DNA만 절단하여 유전자 변화를 도입할 수 있는 게놈 편집 방법이다.

[0021] 프라임에디팅은 “프라임에디터(Prime editor, PE)”에 의해 수행된다. 프라임에디터의 종류로는 PE1, PE2, PE3, 및 PE3b 등이 있으나, 이에 제한되지 않는다. 일 구체예에서, 상기 프라임에디터는 프라임에디터2(PE2)일 수 있다. 프라임에디터는 Cas9 nickase-reverse transcriptase (RT) 융합 단백질 및 프라임에디팅 가이드 RNA (pegRNA)를 포함한다. 본 명세서에서, 프라임에디터는 Cas9 nickase-RT 융합 단백질만을 포함하는 것을 의미할 수도 있고, Cas9 nickase-RT 융합 단백질과 pegRNA를 함께 포함하는 것을 의미할 수도 있다. 예를 들어, 세포 내에 pegRNA를 별도로 도입한 경우, 여기에 프라임에디터를 도입하였다는 것은 Cas9 nickase-RT 융합 단백질만을 도입한 것을 의미할 수 있다. 즉, pegRNA가 이미 도입되어 있는 경우 프라임에디터의 도입은 Cas9 nickase-RT 융합 단백질만을 도입한 것을 의미할 수 있다. 일 구체예에서, 프라임에디터는 Cas9 nickase-RT 융합단백질을 의미할 수 있다. 상기 Cas9 nickase는 Cas9 H850A일 수 있다.

[0022] 프라임에디터에 사용되는 “Cas9 nickase”는 한 가닥의 DNA를 절단 (nick)하도록 변형된 것일 수 있다.

[0023] “프라임에디팅 효율”은 프라임에디터에 의한 유전자 편집 효율을 의미한다. 프라임에디팅 효율은 프라임에디팅을 수행하였을 때, 표적 서열 내에서 의도하지 않은 돌연변이 없이 프라임에디터 및 pegRNA에 의해 유도된 편집이 발생하는 비율로 계산될 수 있다. 상기 프라임에디팅 효율은 백분율로 표시될 수 있다.

[0024] “프라임에디팅 효율에 대한 데이터”는 기존의 공지된 데이터일 수도 있고, 당업자가 적절히 채택할 수 있는 임의의 방법으로 직접 수득한 데이터일 수 있으며, 프라임에디팅 효율을 예측할 수 있는 예측 모델을 생성할 수 있는 데이터라면, 데이터가 수득되는 방법은 제한되지 않는다. 일 구체예에서, 고처리량(high-throughput) 실험을 통해 pegRNA 및 그에 상응하는 표적 서열을 사용하여 분석한 프라임에디팅 효율 데이터일 수 있다.

[0025] 구체적으로, 상기 프라임에디팅 효율에 대한 데이터는, pegRNA를 암호화하는 뉴클레오티드 서열 및 상기 pegRNA가 목적하는 표적 뉴클레오티드 서열을 포함하는 올리고뉴클레오티드를 포함하는 세포 라이브러리에 프라임에디터를 도입하는 단계; 상기 프라임에디터가 도입된 세포 라이브러리로부터 수득한 DNA를 이용하여 딥시퀀싱을 수행하는 단계; 및 상기 딥시퀀싱으로 수득한 데이터로부터 프라임에디팅 효율을 분석하는 단계를 포함하는 방법을 수행하여 수득된 것일 수 있다.

[0026] “역전사 효소(reverse transcriptase, RT)”는 RNA를 주형으로 하고, 이에 상보적인 새로운 DNA를 합성하는 효소이다.

[0027] “pegRNA(prime editing guide RNA)”는 표적 서열을 인식하는 가이드 서열(guide sequence), tracrRNA 스캐폴드 서열, 역전사 개시에 필요한 프라이머 결합 부위(primer binding site, PBS), 및 원하는 유전적 변화를 포함하는 RT 주형(RT template)을 포함한다.

[0028] 상기 pegRNA에서 가이드 서열은 표적 서열과 전부 또는 일부 상보적인 서열을 포함한다.

[0029] “표적 서열(target sequence)”은 pegRNA가 목적하는 표적 뉴클레오티드 서열을 의미한다. 상기 표적 서열은 pegRNA가 표적으로 할 것으로 예상되는 서열일 수 있다. 상기 표적 서열은 공지된 게놈 서열 중 일부 서열일 수 있고, 본 발명의 시스템을 이용하는 당업자가 분석하고자 하는 서열을 임의로 설계한 서열일 수도 있다.

[0030] “올리고뉴클레오티드(oligonucleotide)”는 수 개 내지 수백 개의 뉴클레오티드가 포스포다이에스터 결합으로 연결된 물질을 의미한다. 상기 올리고뉴클레오티드의 길이는 100 nts 내지 300 nts, 100 nts 내지 250 nts, 또는 100 nts 내지 200 nts일 수 있으나, 이에 제한되는 것은 아니며, 당업자가 적절히 조절할 수 있다.

[0031] 상기 올리고뉴클레오티드에 포함되는 pegRNA를 암호화하는 뉴클레오티드 서열은 가이드 서열, RT 주형 서열, PBS 서열 등을 포함할 수 있다.

[0032] 상기 올리고뉴클레오티드에 포함되는 표적 뉴클레오티드 서열은 PAM (protospacer adjacent motif) 및 RT 주형 결합 영역을 포함할 수 있다. 상기 RT 주형 결합 영역은 RT 주형에 전부 또는 일부 상보적인 서열을 포함할 수 있다.

[0033] 상기 올리고뉴클레오티드는 바코드 서열(barcode sequence)을 더 포함할 수 있다. 따라서, 상기 올리고뉴클레오티드는 pegRNA를 암호화하는 서열, 바코드 서열 및 상기 pegRNA가 목적하는 표적 서열을 포함할 수 있다. 상기 바코드 서열의 개수는 1개, 2개, 또는 그 이상일 수 있다. 상기 바코드 서열은 당업자가 목적에 따라 적절히 설계할 수 있다. 예를 들어, 상기 바코드 서열은 딥시퀀싱 수행 후 각각의 pegRNA 및 그에 상응하는 표적 서열 쌍이 식별될 수 있게 하는 것일 수 있다.

[0034] 상기 올리고뉴클레오티드는 PCR 증폭될 수 있도록 프라이머가 결합될 수 있는 추가의 서열을 더 포함할 수 있다.

[0035] “라이브러리”는 특성이 다른 동종의 물질이 2종 이상 포함된 집단 (pool 또는 population)을 의미한다. 따라서, 올리고뉴클레오티드 라이브러리는 뉴클레오티드 서열이 다른 2종 이상의 올리고뉴클레오티드, 예컨대 pegRNA, 및/또는 표적 서열이 다른 2종 이상의 올리고뉴클레오티드를 포함하는 집단일 수 있다. 또한, 세포 라이브러리는 특정이 다른 2종 이상의 세포, 예컨대 세포에 포함되는 올리고뉴클레오티드가 다른 세포들의 집단일 수 있다.

[0036] “벡터”는 상기 올리고뉴클레오티드를 세포 내에 전달할 수 있도록 하는 매개체를 의미할 수 있다. 구체적으로, 벡터는 각각의 pegRNA 암호화 서열 및 표적 서열을 포함하는 올리고뉴클레오티드를 포함할 수 있다. 상기 벡터는 바이러스 벡터 또는 플라스미드 벡터일 수 있으나, 이에 제한되지 않는다. 상기 바이러스 벡터는 렌티바이러스 벡터 또는 레트러바이러스 벡터 등이 사용될 수 있으나, 이에 제한되지 않는다. 상기 벡터는 개체의 세포 내에 존재하는 경우 삽입물, 즉 올리고뉴클레오티드가 발현될 수 있도록 삽입물에 작동가능하게 연결된 필수적인 조절 요소를 포함할 수 있다. 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.

[0037] 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. 상기 도입된 벡터는 세포 내에서 벡터 자체로 존재하거나, 염색체 내에 통합될 수 있으나, 이에 제한되는 것은 아니다.

[0038] 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.

[0039] 본원에서 제조된 세포 라이브러리는 pegRNA 암호화 서열 및 표적 서열을 포함하는 올리고뉴클레오티드가 도입된 세포 집단을 말한다. 이때 각각의 세포들은 pegRNA 암호화 서열 및/또는 표적 서열이 다른 올리고뉴클레오티드가 도입된 것일 수 있다.

[0040] 상기 세포 라이브러리에 프라임에디팅을 유도하기 위하여 프라임에디터를 도입할 수 있다. 상기 프라임에디터는 Cas9 nickase-RT 융합 단백질을 의미할 수 있다. 상기 프라임에디터는 벡터에 의해 세포 내로 도입될 수도 있고, 프라임에디터 그 자체로 세포 내에 도입될 수도 있으며, 세포 내에서 프라임에디터가 활성을 나타낼 수 있는 한 그 도입 방법은 제한되지 않는다. 여기에서, 벡터에 관한 설명은 상술한 바와 같다.

[0041] 상기 세포 라이브러리에서는 도입된 pegRNA 및 표적 서열을 포함하는 올리고뉴클레오티드, 및 프라임에디터에 의해 프라임에디팅이 일어날 수 있다. 즉, 도입된 표적 서열에 대하여 유전자 편집이 일어날 수 있다.

[0042] 상기 프라임에디터가 도입된 세포 라이브러리로부터 DNA를 수득하는 방법은 당업계에 공지된 다양한 DNA 분리 방법을 이용하여 수행될 수 있다.

[0043] 세포 라이브러리를 구성하는 각각의 세포들은 도입된 표적 서열에서 유전자 편집이 발생한 것으로 예상되므로, 표적 서열을 서열 분석하여 유전자 편집 효율을 검출할 수 있다. 상기 서열 분석 방법은 프라임에디팅 효율 데이터를 얻을 수 있다면, 특정 방법에 제한되는 것은 아니나, 예를 들어 딥시퀀싱을 이용할 수 있다.

[0044] 상기 딥시퀀싱으로 수득한 데이터로부터 프라임에디팅 효율을 분석하는 단계는 프라임에디팅 효율을 계산하는 단계를 포함할 수 있다.

[0045] 프라임에디팅 효율은 pegRNA 서열 및 표적 서열의 종류 및/또는 길이에 따라 다르게 나타날 수 있다.

[0046] 상기 프라임에디팅 효율에 대한 데이터는 데이터 세트로 제공될 수 있다.

[0047] 상기 “정보 입력부”는 상술한 프라임에디팅 효율 데이터를 입력 받는 구성 요소이다. 상기 정보 입력부는 시스템의 사용자로부터 직접 프라임에디팅 효율 데이터를 입력 받거나, 또는 미리 저장된 효율 데이터를 입력 받는 것일 수 있으나, 이에 제한되지 않는다.

[0048] 상기 시스템에 있어서, 미리 수득한 프라임에디팅 효율 데이터 또는 공지된 프라임에디팅 효율 데이터가 저장된 저장부를 더 포함할 수 있으나, 이에 제한되지 않는다. 상기 저장부를 포함하는 경우, 상기 정보 입력부는 상기 저장부로부터 설정된 크기 또는 범위의 데이터를 입력 받아, 프라임에디팅 효율을 예측하는데 이용할 수 있다.

[0049] 일 구체예에서, 상기 시스템은 프라임에디팅 효율 데이터가 저장된 데이터베이스를 더 포함할 수 있다. 상기 정보 입력부는 상기 데이터베이스로부터 프라임에디팅 효율 데이터를 입력받는 것일 수 있으나, 이에 제한되지 않는다.

[0051] 상기 프라임에디팅 효율 예측 시스템은 상기 정보 입력부에서 입력 받은 데이터를 이용하여 프라임에디팅 효율에 영향을 미치는 특징과 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하여 프라임에디팅 효율 예측 모델을 생성하는 예측 모델 생성부를 포함한다.

[0052] “예측 모델 생성부”는 상기 정보 입력부를 통해 입력된 프라임에디팅 효율 데이터를 이용하여, 프라임에디팅 효율에 영향을 미치는 특징과 프라임에디팅 효율 간의 관계를 학습할 수 있는 구성을 의미한다. 상기 예측 모델 생성부는 학습된 정보를 기반으로 예측 모델을 생성한다. 따라서, 사용자는 상기 예측 모델을 통해 프라임에디팅 효율을 예측할 수 있다.

[0053] 상기 프라임에디팅 효율에 영향을 미치는 특징은 프라임에디팅에 관여하는 요소에 대한 정보로부터 추출된 것일 수 있다. 상기 프라임에디팅에 관여하는 요소는 프라임에디터를 구성하는 구성요소 및 표적 서열을 포함할 수 있다. 상기 프라임에디터를 구성하는 구성요소는 Cas9-nickase, 역전사 효소, pegRNA를 포함할 수 있다.

[0054] 일 구체예에서, 상기 프라임에디팅 효율에 영향을 미치는 특징은 pegRNA 및 표적 서열 정보로부터 추출된 것일 수 있다.

[0055] 상기 pegRNA 및 표적 서열 정보는 RT 주형 서열 정보, PBS 서열 정보, 및 표적 서열 정보 중 어느 하나 이상을 포함할 수 있다. 구체적으로, 상기 pegRNA 및 표적 서열 정보는 RT 주형의 길이; RT 주형의 구체적인 서열; 편집 유형; 편집 위치; 편집 길이; PBS의 길이; PBS의 구체적인 서열; 표적 서열의 구체적인 뉴클레오티드 서열; 융해 온도; GC 수; 표적 서열, PBS 및 RT 주형 서열의 최소 자가폴딩 자유 에너지; 및 표적 서열에서 Cas9-

sgRNA 활성과 관련된 indel 빈도 중 어느 하나 이상의 정보를 포함할 수 있으며, 프라임에디팅 효율에 영향을 미칠 수 있는 특징이라면 그 종류를 제한하지 않고 모두 포함될 수 있다.

[0056] 상기 편집 유형은 치환 (substitution), 삽입 (insertion), 삭제 (deletion) 등을 포함할 수 있으나, 이에 제한되지 않는다. 상기 편집 유형은 표적 서열에서 치환, 삽입, 또는 삭제되는 뉴클레오티드의 종류 (예: A, G, C, T) 또는 수 (예: 1 nt, 2nts, 3nts)를 포함할 수 있다.

[0057] 상기 편집 위치는 닉킹 부위를 기준으로 계산되는 것일 수 있다. 예를 들어, 상기 편집 위치는, 닉킹 부위로부터 +1, +2, +3 등으로 표현될 수 있다.

[0058] “닉킹 부위 (nicking site)”란 표적 서열에서 Cas9-nickase에 의해 절단되는 부위를 의미한다.

[0059] “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 딥-러닝 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. 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.

[0060] 상기 프라임에디팅 효율에 영향을 미치는 특징은 프라임에디팅 효율에 영향을 미치는 공지된 특징일 수 있고, 상기 프라임에디팅 효율 데이터를 분석하여 추출한 특징일 수도 있다. 상기 프라임에디팅 효율에 영향을 미치는 특징은 상기 예측 모델 생성부에 의해 추출될 수도 있고, 별도의 방법을 수행하여 추출된 특징을 이용할 수도 있다. 상기 별도의 방법은 상기 프라임에디팅 효율 데이터를 이용하여 특징 중요도 (feature importance) 평가를 수행하는 것일 수 있으나, 이에 제한되지 않는다. 예를 들어, 상기 특징 중요도의 평가는 Tree SHAP 방법을 이용할 수 있으나, 이에 제한되지 않는다.

[0061] 상기 예측 모델 생성부는 컨볼루션 신경망(convolutional neural network. CNN) 또는 다층 퍼셉트론(multilayer percentron, MLP)을 기반으로 하여 딥러닝을 수행할 수 있다.

[0062] 일 구체예에서, 상기 프라임에디팅 효율에 영향을 미치는 특징은 PBS 길이 및 RT 주형 길이일 수 있다. 따라서, 상기 예측 모델 생성부는 상기 정보 입력부에서 입력 받은 데이터를 이용하여 컨볼루션 신경망을 기반으로 하여 PBS 길이 및 RT 주형 길이와 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하여 프라임에디팅 효율 예측 모델을 생성할 수 있다.

[0063] 일 구체예에서, 상기 프라임에디팅 효율에 영향을 미치는 특징은 융해 온도, GC 수, GC 함량, 최소 자가-폴딩 자유 에너지 등을 더 포함할 수 있다.

[0064] 상기 예측 모델 생성부는 상기 정보 입력부에서 입력 받은 데이터 중 뉴클레오티드 서열에 관한 데이터는 4차원 이진 매트릭스로 전환시킬 수 있다. 4차원 이진 매트릭스로의 전환은 one-hot 인코딩에 의해 수행될 수 있다.

[0065] 상기 예측 모델은 컨볼루션 레이어 및 완전히 연결된 레이어를 포함하는 것일 수 있다.

[0066] 상기 예측 모델은 컨볼루션 레이어, 완전히 연결된 레이어, 및 회귀 출력 레이어를 포함하는 것일 수 있다.

[0067] 상기 컨볼루션 신경망을 기반으로 하여 딥러닝을 수행하는 단계는,

[0068] 컨볼루션 레이어를 통해 표적 서열, 및 RT 주형 및 PBS 서열에서 2개의 임베딩 벡터를 얻고, 임베딩 벡터를 프라임에디팅 효율에 영향을 미치는 특징과 연결시키는 단계;

[0069] 완전히 연결된 레이어를 통해 상기 벡터에 ReLU(Rectified-linear-unit) 활성 기능을 곱하는 단계; 및

[0070] 회귀 출력 레이어를 통해 출력의 선형 변환을 수행하여 프라임에디팅 효율에 대한 예측 점수를 계산하는 단계를 포함할 수 있다.

[0071] 상기 예측 모델은 풀링 레이어를 포함하지 않는 것일 수 있다.

[0072] 일 실시예에서, 48,000 쌍의 pegRNA와 표적 서열을 갖는 세포 라이브러리를 사용하여 얻은 프라임에디팅 효율 데이터를 이용하여 컨볼루션 신경망을 기반으로 하여 PBS 길이 및 RT 주형 길이와 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하였다. 그 결과, 주어진 표적 서열에 대해 프라임에디팅 효율을 예측할 수 있는 모델 DeepPE를 생성하였다. 상기 DeepPE를 사용하여, 주어진 표적 서열에서 특정 유형의 편집을 의도하는 경우,

PBS 및 RT 주형의 길이에 따른 프라임에디팅 효율을 예측할 수 있었다.

[0073] 다른 구체예에서, 상기 프라임에디팅 효율에 영향을 미치는 특징은 편집 유형, 편집 위치, 또는 이들의 조합일 수 있다. 따라서, 상기 예측 모델 생성부는 상기 정보 입력부에서 입력 받은 데이터를 이용하여 다층 퍼셉트론을 기반으로 하여 편집 유형, 편집 위치, 또는 이들의 조합과 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하여 프라임에디팅 효율 예측 모델을 생성할 수 있다.

[0074] 일 실시예에서, 6,800 쌍의 pegRNA와 표적 서열을 갖는 세포 라이브러리를 사용하여 얻은 프라임에디팅 효율 데이터를 이용하여 다층 퍼셉트론을 기반으로 하여 편집 유형 또는 편집 위치와 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하였다. 그 결과, 주어진 표적 서열에 대해 프라임에디팅 효율을 예측할 수 있는 모델 PE_type 및 PE_position을 생성하였다. 상기 PE_type 및 PE_position을 사용하여, 주어진 표적 서열에서 편집 유형 및/또는 편집 위치에 따른 프라임에디팅 효율을 예측할 수 있었다.

[0075] 동일한 원리를 이용하여, 임의의 표적 서열에서 특정 유형의 편집을 의도하는 경우, 프라임에디팅 효율에 영향을 미치는 각 특징의 특정값에 따른 프라임에디팅 효율을 예측할 수 있는 모델을 생성할 수 있다.

[0076] 상기 예측 모델 생성부는 pegRNA 및 표적 서열 정보로부터 프라임에디팅 효율에 영향을 미치는 특징을 추출하는 특징 추출 모듈을 포함할 수 있으나, 이에 제한되지 않는다. 또한, 상기 예측 모델 생성부는 상기 특징 추출 모듈에서 추출된 특징을 조합하는 조합 모듈을 더 포함할 수 있으나, 이에 제한되지 않는다.

[0078] 상기 프라임에디팅 효율 예측 시스템은 프라임에디팅의 후보 표적 서열을 입력받는 후보 서열 입력부를 포함한다.

[0079] 상기 “후보 서열 입력부”는 상기 후보 표적 서열을 입력 받기 위한 프라임에디팅 효율 예측 시스템의 구성이다.

[0080] 상기 후보 표적 서열은 프라임에디팅 효율을 분석 또는 예측하고자 하는 pegRNA의 표적 뉴클레오티드 서열을 의미한다. 상기 후보 표적 서열은 프라임에디팅 효율을 확인하고자 하는 개체의 유전체 서열에서 유래한 것일 수 있고, 또는 당업계에 공지된 방법으로 설계 및 합성된 임의의 서열일 수도 있으나, 프라임에디팅 효율 예측을 위해 본 발명의 시스템에 적용될 수 있는 서열이라면, 그 종류를 제한하지 않는다.

[0081] 일 구체예에서, 상기 후보 표적 서열은 10개 내지 100개, 20개 내지 100개, 30개 내지 100개, 10개 내지 90개, 20개 내지 90개, 30개 내지 90개, 10개 내지 80개, 20개 내지 80개, 30개 내지 80개, 10개 내지 70개, 20개 내지 70개, 30개 내지 70개, 10개 내지 60개, 20개 내지 60개, 30개 내지 60개, 10개 내지 50개, 20개 내지 50개, 또는 30개 내지 50개의 뉴클레오티드로 구성된 것일 수 있으나, 이에 제한되지 않는다.

[0082] 상기 후보 표적 서열은 PAM (protospacer adjacent motif), 및 프로토스페이서 서열을 포함할 수 있으나, 이에 제한되지 않는다. 상기 PAM 및 프로토스페이서 서열은 프라임에디터가 표적 서열을 인식하는 과정에 관여하는 서열이다.

[0084] 상기 프라임에디팅 효율 예측 시스템은 상기 후보 서열 입력부에 입력된 후보 표적 서열을 상기 예측 모델 생성부에서 생성된 효율 예측 모델에 적용하여 프라임에디팅 효율을 예측하는 효율 예측부를 포함한다.

[0085] “효율 예측부”는 기 설정된 방법으로 구축된 효율 예측 모델에 후보 서열 입력부를 통해 입력된 후보 표적 서열을 적용하여, 프라임에디팅 효율을 예측하는 구성이다.

[0086] 상기 시스템에 있어서, 상기 효율 예측부는 프라임에디터에 의한 후보 표적 서열의 프라임에디팅 효율을 예측하는 것일 수 있다.

[0087] 일 실시예에서, DeepPE에 입력된 특정 표적 서열에 대해, 특정 유형의 편집을 의도하는 경우, RT 주형 및 PBS 길이에 따른 프라임에디팅 효율을 예측하였다.

[0088] 다른 실시예에서, PE_type 및 PE_position에 입력된 특정 표적 서열에 대해, 편집 종류(예: 편집 유형, 편집 위치, 편집된 뉴클레오티드의 수 등)에 따른 프라임에디팅 효율을 예측하였다.

[0089] 따라서, 본 시스템의 사용자는 상기 예측 모델에 의해 예측된 프라임에디팅 효율을 참고하여 주어진 표적 서열

에 유전자 편집을 유도하기 위한 pegRNA 서열, 구체적으로 RT 주형 및/또는 PBS 서열을 설계할 수 있다.

[0091] 상기 프라임에디팅 효율 예측 시스템은 효율 예측부에서 예측된 프라임에디팅 효율을 출력하는 출력부를 더 포함할 수 있다.

[0092] 상기 출력부가 출력하는 프라임에디팅 효율에 대한 정보는, 프라임에디팅 효율에 대해 산출된 수치, 또는 미리 설정된 기준값에 대한 상대적인 수치로 나타낼 수 있으나, 출력되는 정보의 형태나 종류는 제한되지 않는다.

[0094] 다른 양상은 딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법을 제공한다.

[0095] 상기 딥러닝을 이용한 프라임에디팅 효율 예측 시스템을 구축하는 방법은,

[0096] 프라임에디터의 프라임에디팅 효율 데이터 세트를 수득하는 단계; 및

[0097] 상기 효율 데이터 세트를 이용하여 프라임에디팅 효율에 영향을 미치는 특징과 프라임에디팅 효율 간의 관계를 학습하는 딥러닝을 수행하여 프라임에디팅 효율 예측 모델을 생성하는 단계를 포함한다.

[0099] 상기 효율 데이터 세트를 수득하는 단계는, pegRNA를 암호화하는 뉴클레오티드 서열 및 상기 pegRNA가 목적하는 표적 뉴클레오티드 서열을 포함하는 올리고뉴클레오티드를 포함하는 세포 라이브러리에 프라임에디터를 도입하는 단계; 상기 프라임에디터가 도입된 세포 라이브러리로부터 수득한 DNA를 이용하여 딥시퀀싱을 수행하는 단계; 및 상기 딥시퀀싱으로 수득한 데이터로부터 프라임에디팅 효율을 분석하는 단계를 포함할 수 있다.

[0100] 상기 올리고뉴클레오티드는 바코드 서열을 더 포함할 수 있다. 상기 바코드 서열에 대한 설명은 상술한 바와 같다.

[0101] 상기 프라임에디팅 효율은 표적 서열 내에서 의도하지 않은 돌연변이 없이 프라임에디터 및 pegRNA에 의해 유도된 편집이 발생한 비율로 계산되는 것일 수 있다.

[0102] 상기 프라임에디팅 효율에 영향을 미치는 특징은 pegRNA 및 표적 서열 정보로부터 추출된 것일 수 있다. “프라임에디팅 효율에 영향을 미치는 특징”, “pegRNA 및 표적 서열 정보”에 관한 설명은 상술한 바와 같다.

[0103] 상기 pegRNA 및 표적 서열 정보는 RT 주형 서열 정보, PBS 서열 정보, 및 표적 서열 정보 중 어느 하나 이상을 포함하는 것일 수 있으나, 이에 제한되지 않는다.

[0104] 상기 예측 모델을 생성하는 단계에서, 컨볼루션 신경망(convolutional neural network. CNN) 또는 다층 퍼셉트론(multilayer percentron, MLP)을 기반으로 하여 딥러닝을 수행할 수 있다.

[0105] 상기 예측 모델을 생성하는 단계 이후에, 생성된 예측 모델을 검증하는 단계를 더 포함할 수 있다. 상기 검증은 당업계에 알려진 방법을 통해 검증할 수 있다.

[0107] 다른 양상은 프라임에디팅 효율 예측 방법을 제공한다.

[0108] 상기 프라임에디팅 효율 예측 방법은,

[0109] 프라임에디팅의 후보 표적 서열을 설계하는 단계; 및

[0110] 상기 설계된 후보 표적 서열을 일 양상에 따른 프라임에디팅 효율 예측 시스템에 적용하여 프라임에디팅 효율을 예측하는 단계를 포함한다.

[0111] 상기 후보 표적 서열, 및 프라임에디팅 효율 예측 시스템에 대한 설명은 상술한 바와 같다.

[0113] 다른 양상은 상기 프라임에디팅 효율 예측 방법을 컴퓨터로 실행하기 위한 프로그램이 기록된 컴퓨터 판독가능 기록매체를 제공한다.

[0114] 상기 프로그램은 상기 프라임에디팅 효율 예측 시스템 또는 상기 프라임에디팅 효율 예측 방법을 컴퓨터 프로그

래밍 언어로 구현한 것일 수 있다.

[0115] 상기 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.

[0116] 상기 프로그램은 온라인 또는 오프라인으로 제공될 수 있다.

발명의 효과

[0117] 일 양상에 따른 딥러닝을 이용한 프라임에디팅 효율 예측 시스템은 기존의 기계 학습 기반 예측 방법에 비해 높은 정확도로 프라임에디팅 효율을 예측할 수 있다. 따라서, 상기 시스템은 유전자 편집에 의한 질병 치료 등 유전자 가위를 적용하는 모든 분야에서 유용하게 사용될 수 있다.

도면의 간단한 설명

[0118] 도 1은 프라임에디팅 구성요소를 나타낸 개략도이다. PE2 단백질은 일시적 트랜스펙션(transient transfection)에 의해 발현되었다. 인간 U6 프로모터 (hU6)는 PE2를 표적 서열로 안내하는 pegRNA의 발현을 위해 사용되었다. Guide, 가이드 서열; RTT, RT 주형; PBS, 프라이머 결합 부위; RT, 역전사효소; BSD-R, 블라스티시딘 내성 유전자.

도 2는 라이브러리 1 및 2의 구성을 나타낸 것이다. 라이브러리 1에서, 2,000개의 가이드 서열에 대해, 각각 상이한 PBS 및 RT 주형 길이의 24개 조합을 생성하여 48,000개 pegRNA를 구성하였다. 라이브러리 2에서, 2,000개의 가이드 서열을 34개의 서로 다른 조합의 PBS 및 RT 주형과 연결하여, 상이한 위치에서 다양한 유형의 편집을 생성하도록 하여, 6,800개의 pegRNA를 구성하였다.

도 3은 pegRNA, cDNA 및 넓은 표적 서열 내에서 위치가 어떻게 지정되는지를 나타낸 개략도이다. pegRNA 및 pegRNA로부터 생성된 cDNA 내의 위치는 Cas9 nickase의 닉킹 부위에서 시작하여 넘버링하였다. 넓은 표적 서열 내의 위치는 PAM으로부터 상류의 20번째 뉴클레오티드가 위치 1이고, NGG PAM의 뉴클레오티드가 위치 21-23이 되도록 지정하였다.

도 4는 프라임에디팅 효율의 고처리량 평가 절차의 개략도이다.

도 5는 두 개의 다른 실험에 의해 독립적으로 PE2 암호화 플라스미드로 형질감염된 반복실험에서 PE 효율의 상관관계를 나타낸 것이다. 라이브러리 1 및 2의 결과를 결합하였다. 분석의 정확도를 증가시키기 위해, 딥시퀀싱 판독 수가 200 미만이거나, 백그라운드 프라임에디팅 빈도가 5% 이상인 경우의 pegRNA 및 표적 서열 쌍을 제거하였다.

도 6은 내인성 부위에서 측정된 PE 효율과 상응하는 통합된 표적 서열에서의 PE 효율 간의 상관관계를 나타낸 것이다. 초기 연구(Anzalone, A.V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 576, 149-157 (2019))에서 공개된 PE3 효율의 데이터 세트를 사용하였다.

도 7은 내인성 부위에서 측정된 PE 효율과 상응하는 통합된 표적 서열에서의 PE 효율 간의 상관관계를 나타낸 것이다. 데이터 세트는 Endo-BR1-TR1, Endo-BR1-TR2, Endo-BR2-TR1, Endo-BR2-TR2, Endo-BR2-TR3, 또는 Endo-BR3를 사용하였다.

도 8은 SpCas9 유도 indel 빈도 및 동일한 표적 서열에서 결정된 PE2 효율 간의 상관관계를 나타낸 것이다. PBS 및 RT 주형 길이의 영향을 최소화하기 위해, 상이한 PBS 및 RT 주형 길이를 갖는 24개의 pegRNA 중에서 가장 높은 효율을 나타내는 pegRNA를 각 표적 서열 당 선택하였다. pegRNA 및 표적 서열 쌍의 수는 n = 1,956이었다.

도 9는 라이브러리 1을 사용하여 SpCas9 유도 indel 빈도 및 동일한 표적 서열에서 결정된 PE2 효율 간의 상관관계를 나타낸 것이다. 모든 24개 조합의 PBS 및 RT 주형 길이를 고려하여 상관관계를 평가하였다. pegRNA 및 표적 서열 쌍의 수는 n = 21,288이었다.

도 10은 PBS 및 RT 주형 길이의 PE2 효율에 대한 영향을 나타낸 것이다. 히트맵은 주어진 길이의 PBS 및 RT 주

형에서 평균 편집 효율을 나타낸다.

도 11은 PBS 및 RT 주형 길이의 프라임에디팅 효율에 대한 영향을 나타낸 것이다. (A) RT 주형의 길이는 12 nt로 고정된 경우, 다양한 길이의 PBS에서의 PE 효율; (B) PBS의 길이는 13 nt로 고정된 경우, 다양한 길이의 RT 주형에서의 PE 효율. PE 효율에서 통계적으로 유의한 차이가 없는 실험군의 서브세트 ($P < 0.05$)는 a, b, c, 및 d와 같은 문자로 나타내었다. 박스에서, 탑, 중간, 및 바텀 라인은 각각 25, 50, 및 75 백분위수를 나타내며, 위스커(whiskers)는 10 및 90 백분위수를 나타내며, 특이치는 개별점으로 표시된다. X 축에 지정된 실험군 당 pegRNA 및 표적 서열 쌍의 수는 n = 1,772 - 1,826이다.

도 12는 주어진 PBS 길이 및 RT 주형 길이에 대해 5% 이상의 PE2 효율을 갖는 pegRNA의 빈도이다.

도 13은 (A) 주어진 PBS 길이 및 RT 주형 길이에 대해 5% 미만의 편집 효율을 갖는 pegRNA의 빈도; (B) 주어진 PBS 길이 및 RT 주형 길이에 대해 5% 이상의 편집 효율을 갖는 pegRNA의 빈도이다.

도 14는 주어진 표적 서열 당 가장 높은 편집 효율을 유도하는 PBS 및 RT 주형 길이 조합의 빈도를 나타낸 것이다.

도 15는 각 표적에서 가장 높은 편집 효율을 나타낸 PBS 및 RT 주형 길이의 조합을 선택할 때 평균 편집 효율을 나타낸 것이다.

도 16은 Tree SHAP (XGBoost classifier)에 의해 결정된 PE2 효율과 연관된 가장 중요한 10개의 특징을 나타낸 것이다. 오른쪽 그래프에서, 각 표적 서열은 점으로 표시되고; X축에서 점의 위치는 SHAP 값을 나타낸다. 높고 낮은 SHAP 값은 각각 높고 낮은 프라임에디팅 효율과 연결된다. 점의 색은 특정 표적 서열에 대한 관련 특징 값을 나타내며; 빨간색 및 파란색은 관련 특징의 높고 낮은 값을 나타낸다. 겹쳐진 지점은 Y축 방향에서 약간 분리되어 밀도를 분명하게 하였다.

도 17은 Tree SHAP에 의해 결정된 PE2 효율과 연관된 가장 중요한 1 내지 51번째 특징을 나타낸 것이다.

도 18은 Tree SHAP에 의해 결정된 PE2 효율과 연관된 가장 중요한 52 내지 100번째 특징을 나타낸 것이다.

도 19는 PBS 및 RT 주형에서 GC 함량 및 GC 수의 프라임에디팅 효율에 대한 영향을 나타낸 것이다.

도 20은 PBS의 융해 온도 및 RT 주형에 상응하는 표적 DNA 영역의 프라임에디팅 효율에 대한 영향을 나타낸 것이다. PBS 및 RT 주형의 길이는 각각 13 nt 및 12 nt였다. X축에 지정된 실험군 당 pegRNA 및 표적 서열 쌍의 수는 n = 13-736이었다.

도 21은 1-bp 삽입, 결실, 및 치환의 경우 PE2 효율을 나타낸 것이다. pegRNA 및 표적 서열 쌍의 수는 삽입의 경우 739개, 결실의 경우 178개, 치환의 경우 566개였다.

도 22는 삽입된 뉴클레오티드 유형 및 수의 PE2 효율에 대한 영향을 나타낸 것이다. pegRNA 및 표적 서열 쌍의 수는 A, C, G, T, AG, AGGAA(5 bp), 및 AGGGAATCATG(10bp) 삽입 각각에 대해 183, 183, 188, 185, 184, 179, 및 163이었다.

도 23은 결실 길이의 PE2 효율에 대한 영향을 나타낸 것이다. pegRNA 및 표적 서열 쌍의 수는 1bp, 2bp, 5bp, 및 10 bp 결실 각각에 대해 178, 189, 185, 및 169이었다.

도 24는 치환 유형의 PE2 효율에 대한 영향을 나타낸 것이다. pegRNA 및 표적 서열 쌍의 수는 C에서 T로의 변환, C에서 G로의 변환, A에서 G로의 변환, A에서 C로의 변환, A에서 T로의 변환, G에서 T로의 변환, T에서 A로의 변환, T에서 C로의 변환, G에서 C로의 변환, G에서 A로의 변환, C에서 A로의 변환, T에서 G로의 변환 각각에 대해 88, 87, 36, 35, 34, 44, 21, 20, 45, 45, 90, 및 21이었다.

도 25는 치환의 유형의 프라임에디팅 효율에 대한 영향을 나타낸 것이다. pegRNA 및 표적 서열 쌍의 수는 A에서 T로의 변환, C에서 G로의 변환, G에서 C로의 변환, 및 T에서 A로의 변환에 대해 52, 40, 50, 및 35이고(왼쪽 그래프), A에서 T로의 변환, C에서 G로의 변환, G에서 C로의 변환, 및 T에서 A로의 변환에 대해 49, 44, 43, 및 42이고(가운데 그래프), 및 A에서 T로의 변환, C에서 G로의 변환, G에서 C로의 변환, 및 T에서 A로의 변환에 대해 29, 46, 51, 47이었다(오른쪽 그래프).

도 26은 1-bp 변환 치환의 경우 편집 위치의 PE2 효율에 대한 영향을 나타낸 것이다. X축에 나타낸 편집 위치는 닉킹 부위로부터 카운트되었다. pegRNA 및 표적 서열 쌍의 수는 위치 +1, +2, +3, +4, +5, +6, +7, +8, +9, +11, 및 +14에 대해 각각 179, 186, 184, 180, 173, 184, 182, 178, 177, 178, 및 173이었다.

도 27는 2개 위치에서 1-bp 변환 치환의 경우 편집 위치의 프라임에디팅 효율에 대한 영향을 나타낸 것이다. pegRNA 및 표적 서열 쌍의 수는 위치 +1 및 +2, 위치 +1 및 +5, 위치 +1 및 +10, 위치 +2 및 +3, 위치 +2 및 +5, 위치 +2 및 +10, 위치 +5 및 +6, 위치 +5 및 +10, 및 위치 +10 및 +11 각각에 대해 190, 181, 186, 190, 177, 180, 183, 170, 및 169이었다.

도 28은 도 27에 설명된 두 개의 편집 위치 간의 거리에 따른 일부 편집의 상대적 빈도를 나타낸 것이다.

도 29는 두 개의 뉴클레오티드가 치환의 목적일 때 프라임에디팅 분석 결과를 나타낸 것이다. 히트맵은 일부(1 nt) 및 전부(2 nt) 편집의 평균 빈도를 나타낸다. pegRNA 및 표적 서열 쌍의 수는 위치 +1 및 +2, 위치 +1 및 +5, 위치 +1 및 +10, 위치 +2 및 +3, 위치 +2 및 +5, 위치 +2 및 +10, 위치 +5 및 +6, 위치 +5 및 +10, 및 위치 +10 및 +11 각각에 대해 190, 181, 186, 190, 177, 180, 183, 170, 및 169이었다.

도 30은 사용된 기계 학습 프레임워크에 따른 예측 모델의 교차 검증 결과를 나타낸 것이다.

도 31은 데이터 세트 HT-Test (pegRNA 및 표적 서열 쌍의 수 n = 4,457) 및 Endo-BR1-TR1 (n = 26)를 사용한 DeepPE의 평가 결과를 나타낸 것이다.

도 32는 DeepPE를 데이터 세트 HT-Test를 사용한 다른 예측 모델과 성능 비교한 결과이다. 바 그래프는 측정된 PE2 효율과 예측된 활성 점수 간의 Spearman 상관계수를 나타낸다. pegRNA 및 표적 서열 쌍의 수 n = 4,457이었다.

도 33은 pegRNA 및 PE2를 암호화하는 플라스미드를 HEK293T 세포로 일시적 트랜스펙션 한 후, 내인성 부위에서 PE2 효율을 측정하여 얻은 6개의 데이터 세트를 사용한 DeepPE의 평가 결과를 나타낸 것이다. 데이터 세트 Endo-BR1-TR1, Endo-BR1-TR2, Endo-BR2-TR1, Endo-BR2-TR2, Endo-BR2-TR3, 및 Endo-BR3 각각에 대해 표적 서열의 수는 26, 25, 23, 23, 23, 및 16이었다.

도 34는 HCT116 및 MDA-MB-231 세포를 사용한 DeepPE의 평가 결과를 나타낸 것이다. DeppPE의 훈련에 사용된 적 없는 렌티바이러스 통합된 표적 서열에서 HCT116 (HCT로 약칭함) 및 MDA-MB-231 (MDA로 약칭함) 세포주를 사용하여 PE2 효율의 8개 데이터 세트를 생성하였다. pegRNA 및 표적 서열 쌍의 수는 HCT-BR1-TR1, HCT-BR1-TR2, HCT-BR2-TR1, HCT-BR2-TR2, MDA-BR1-TR1, MDA-BR1-TR2, MDA-BR2-TR1 및 MDA-BR2-TR2 각각에 대해 72, 75, 75, 75, 71, 73, 74, 및 75이었다. 세포주 당 두 개의 생물학적 반복실험 (BR1 및 BR2)을 평가하였고, 각 생물학적 반복실험은 두 개의 기술적 반복실험 (TR1 및 TR2)를 가졌다.

도 35은 주어진 표적 서열에서 PBS 및 RT 주형 길이의 가능한 24개 조합 중에서 가장 효율적인 조합을 선택하기 위한 DeepPE 및 방법의 성능 비교를 나타낸 것이다. 예를 들어, "13-nt PBS & 12 nt-PT template"란 표적 서열에 관계 없이 이러한 길이의 조합을 선택하는 것을 의미한다. 초기 연구 권장사항 A 및 B는 13-nt PBS 및 12-nt RT 주형(RTT)을 사용하고, 필요에 따라 RTT 길이를 변경하는 것에 의해 마지막 주형 뉴클레오티드로서 G를 사용하지 않는 것을 기반으로 한다. 권장사항 A에서, 마지막 주형 뉴클레오티드가 G이면, 12-nt 보다 10-nt RTT가 선택된다. 이러한 변경 후 마지막 주형 뉴클레오티드가 다시 G이면, 15-nt RTT가 선택된다. 권장사항 B에서, 마지막 주형 뉴클레오티드가 G이면, 12-nt 보다 15-nt RTT가 선택된다. 이러한 변경 후에 마지막 주형 뉴클레오티드가 다시 G이면, 10-nt RTT가 선택된다. 대조군으로서, pegRNA를 무작위로 선택하였다(Random 1 및 Random 2). 표적 서열의 수는 그룹 당 97개이다.

도 36은 사용된 기계 학습 프레임워크에 따른 PE_type의 교차 검증 결과를 나타낸 것이다.

도 37은 사용된 기계 학습 프레임워크에 따른 PE_position의 교차 검증 결과를 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0119] 이하 본 발명을 실시예를 통하여 보다 상세하게 설명한다. 그러나, 이들 실시예는 본 발명을 예시적으로 설명하기 위한 것으로 본 발명의 범위가 이들 실시예에 한정되는 것은 아니다.

[0121] **실시예 1: 재료의 준비**

[0122] **실시예 1-1: 프라임에디터2 (PE2) 발현 벡터 pLenti-PE2-BSD의 구축**

[0123] 유전자 가위 프라임에디터2 (Prime Editor 2, PE2) 발현 벡터는 다음과 같이 구축하였다. LentiCas9-Blast 플라스미드 (Addgene #52962)를 AgeI 및 BamHI 제한 효소 (NEB)로 37℃에서 4시간 동안 분해(digestion)하고, 1

μl Quick-CIP (NEB)로 37℃에서 10분 동안 처리하였다. 다음으로, 선형화된 플라스미드를 MEGAquick-spin 전체 프래그먼트 DNA 정제 키트 (iNtRON Biotechnology)를 사용하여 겔 정제하였다. pCMV-PE2 (Addgene #132775)로부터의 PE2 암호화 서열을 Solg™ 2× pfu PCR Smart mix (Solgent)를 사용하여 PCR에 의해 증폭하였다. 앰플리콘은 NEBuilder HiFi DNA assembly kit (NEB)를 사용하여 선형화된 LentiCas9-Blast 플라스미드로 어셈블시켰다. 어셈블된 플라스미드를 pLenti-PE2-BSD로 명명하였다.

[0125] **실시예 1-2: 올리고뉴클레오티드 라이브러리 디자인**

[0126] 54,836 쌍의 pegRNA 및 표적 서열을 포함하는 올리고뉴클레오티드 풀을 Twist Bioscience(San Francisco, CA)에서 합성하였다.

[0127] 각각의 올리고뉴클레오티드는 하기 구성요소를 함유하였다: 19-nt 가이드 서열, BsmBI 제한 부위 #1, 15-nt 바코드 서열 (바코드 1), BsmBI 제한 부위 #2, RT 주형 서열, PBS (primer binding site) 서열, 폴리 T 서열, 18-nt 바코드 서열 (바코드 2), 및 PAM (protospacer adjacent motif)과 RT 주형 결합 영역을 포함하는 상응하는 43~47-nt 넓은 표적 서열.

[0128] 바코드 1은 BsmBI로 절단하여 제거할 수 있는 스터퍼(stuffer)이다. 바코드 2 (표적 서열의 업스트림에 위치함)는 개별 pegRNA 및 표적 서열 쌍이 딥시퀀싱 후 식별될 수 있게 한다. 이들의 서열에서 의도하지 않은 BsmBI 제한 부위를 포함하는 올리고뉴클레오티드는 제외하였다.

[0129] PBS 및 RT 주형 길이의 PE2 효율에 대한 영향을 테스트하기 위해, 2,000 쌍의 가이드 및 표적 서열에 대하여, 24개의 PBS 및 RT 주형 길이 조합(6개의 PBS 길이 (7, 9, 11, 13, 15, 17 뉴클레오티드(nts)) x 4개의 RT 주형 길이(10, 12, 15, 20 nts) = 24개)을 갖는 pegRNA를 제조하여, 총 48,000개(=24 x 2,000) 쌍의 pegRNA 및 표적 서열이 되도록 하였다 (라이브러리 1). pegRNA는 닉킹 부위로부터 위치 +5에서 G에서 C로의 전환 돌연변이를 생성하도록 설계되었다. 2,000개의 표적 서열은 인간 단백질-암호화 유전자로부터 무작위로 선택하였다. 여기에서 SpCas9에 의해 유도된 indel 빈도를 이전 연구에서 측정한 바 있으며(Kim, H.K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci Adv* **5**, eaax9249 (2019)), 이는 동일한 표적 서열에서 SpCas9과 PE 효율 사이의 상관관계를 결정할 수 있게 한다.

[0130] 또한, 유전자 편집 위치, 유형, 및 길이의 PE2 효율에 대한 영향을 평가하기 위해 라이브러리 2로 명명한 다른 라이브러리를 준비하였다. 구체적으로, 라이브러리 1에 사용된 2,000개의 표적 서열에서 200개 표적 서열을 무작위로 선택하고, 하기와 같이 각각의 표적 서열에 대한 34개의 상이한 RT 주형을 설계하였다.

[0131] i) 편집 위치의 영향 (11개의 RT 주형): RT 주형은 닉킹 부위로부터 위치 +1, +2, …, +8, +9, +11, 및 +14에서 전환 돌연변이를 도입하도록 설계하였다. PBS 및 RT 주형의 길이는 각각 13 및 20 nts로 고정하였다.

[0132] ii) 편집 유형 및 길이의 영향 (14개 RT 주형): RT 주형은 닉킹 부위로부터 위치 +1에서 삽입 (삽입된 서열 = A, G, C, T, AG, AGGAA, 및 AGGAATCATG), 삭제 (1-, 2-, 5-, 및 10-nt), 및 단일 염기 치환 (모든 가능한 1-nt 치환)을 도입하도록 설계하였다. PBS 및 RT 주형의 우측상동부위(right homology arm)의 길이는 각각 13 및 14 nts로 고정하였다.

[0133] iii) PAM 편집의 영향 (9개 RT 주형): RT 주형은 위치 +1 및 +2, +1 및 +5, +1 및 +10, +2 및 +3, +2 및 +5, +2 및 +10, +5 및 +6, +5 및 +10, 및 +10 및 +11에서 2-bp 전환 돌연변이를 도입하도록 설계하였다. PBS 및 RT 주형의 길이는 각각 13 및 16 nts로 고정하였다.

[0134] 또한, 초기 프라임에디팅 연구(Anzalone, A.V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149-157 (2019))에서 사용된 표적 서열 당 5개의 고유한 바코드를 갖는 36 쌍의 pegRNA 및 표적 서열을 포함시켰다. 이 세트는 통합된 서열과 내인성 부위에서의 프라임에디팅 효율의 상관관계 결정하기 위해 사용되었다.

[0135] 이와 같이 모두 합쳐서, 총 54,836 쌍의 pegRNA 및 표적 서열 - 48,000쌍 (라이브러리 1에서, 2,000 x 24) + 6,800쌍 (라이브러리 2에서, 200 x 34) + 36쌍 (초기 프라임에디팅 연구에서)로 구성됨 - 을 사용하였다.

[0137] **실시예 1-3: 플라스미드 라이브러리 제작**

[0138] 상기 pegRNA 암호화 서열 및 상응하는 표적 서열의 쌍을 함유하는 플라스미드 라이브러리는 2단계 클로닝 공정을 사용하여 제조하였다:

[0139] (단계 I) 깁슨 어셈블리 및

[0140] (단계 II) 제한 효소-유도된 절단 및 결찰.

[0141] PCR을 통한 올리고뉴클레오티드 증폭 동안, 쌍을 이룬 가이드 RNA와 표적 서열의 분리는 이러한 2단계 공정에 의해 효과적으로 방지된다. 멀티단계 절차는 이전에 보고된 방법(Shen, J.P. et al. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. *Nat Methods* **14**, 573-576 (2017))을 조정 및 수정하였다.

[0142] **(1) 단계 I: pegRNA 암호화 서열 및 표적 서열의 쌍을 함유하는 초기 플라스미드 라이브러리의 구축**

[0143] 올리고뉴클레오티드 풀을 Phusion Polymerase (NEB)를 사용하여 15 사이클의 PCR을 통해 증폭시킨 후, 앰플리콘을 겔 정제하였다. Lenti_gRNA-Puro 벡터 (Addgene #84752)를 BsmBI 효소 (NEB)로 55℃에서 6시간 동안 분해(digestion)시켰다. 선형화된 벡터를 1μl의 Quick CIP로 37℃에서 10분 동안 처리하고, 겔 정제하였다. 깁슨 어셈블리를 사용하여 올리고뉴클레오티드의 증폭된 풀을 선형화된 Lenti_gRNA-Puro 벡터와 어셈블하였다. 컬럼 정제 후, 어셈블된 산물을 MicroPulser (Bio-Rad)를 사용하여 전기적격(electrocompetent) 세포 (Lucigen)로 형질전환시켰다. 이어서, SOC 배지 (2 ml)를 형질전환 혼합물에 첨가하고, 이를 37℃에서 1시간 동안 인큐베이션하였다. 이어서, 세포를 50 μg/ml 카르베니실린(carbenicillin)을 함유하는 Luria-Bertani (LB) 아가 플레이트에 시딩하고 인큐베이션하였다. 배양물의 작은 분획 (0.1, 0.01, 및 0.001 μl)을 별도로 시딩하여 라이브러리 범위(coverage)를 결정할 수 있게 하였다. 총 수확된 콜로니로부터 플라스미드를 추출하였다. 이 초기 플라스미드 라이브러리의 계산된 범위는 올리고뉴클레오티드 수의 113배였다.

[0144] **(2) 단계 II: sgRNA 스캐폴드 삽입**

[0145] 단계 I에서 제조한 초기 플라스미드 라이브러리를 BsmBI로 8시간 동안 분해(digestion)한 후, 1μl의 Quick CIP로 37℃에서 10분 동안 처리하였다. 분해(digestion)된 산물을 0.6% 아가로스 겔에서 크기선택(size-selection) 후 겔 정제하였다. pRG2 플라스미드 (Addgene #104174)에서 sgRNA 스캐폴드 서열을 Phusion polymerase 및 쌍의 각 멤버에서 BsmBI 제한 부위를 갖는 프라이머 쌍을 사용하여 30 사이클 PCR 증폭하였다. 생성된 앰플리콘을 BsmBI로 적어도 12시간 동안 분해(digestion)하고, 2% 아가로스 겔에서 겔 정제하였다. 정제된 삽입물 (10 ng)을 T4 리가아제 (Enzynomics)를 사용하여 16℃에서 16시간 동안 분해된 초기 플라스미드 라이브러리 벡터 (200 ng)로 라이게이션하였다. 라이게이션 산물을 컬럼 정제하고, Endura 전기적격 세포(Lucigen)로 전기천공시켰다. 콜로니를 수확하고, 최종 플라스미드 라이브러리를 추출하였다. 최종 플라스미드 라이브러리의 계산된 범위는 785x였다.

[0147] **실시예 1-4: 렌티바이러스의 생산**

[0148] HEK293T 세포 (4.0×10^6 또는 8.0×10^6)를 DMEM(Dulbecco's Modified Eagle Medium)를 함유하는 100-mm 또는 150-mm 세포 배양 접시에 시딩하였다. 15시간 후, DMEM을 25μM 클로로퀸 디포스페이트를 함유하는 새로운 배지로 교환한 후, 세포를 추가 5시간 동안 인큐베이션하였다. 플라스미드 라이브러리 및 psPAX2 (Addgene #12260)를 pMD2.G (Addgene #12259)와 1.3:0.72:1.64의 몰비로 혼합하고, 폴리에틸렌이민을 사용하여 HEK293T 세포로 공동-트랜스펙션시켰다. 트랜스펙션 후 15 시간에, 세포를 유지 배지로 리프레시(refresh)하였다. 트랜스펙션 후 48 시간에, 렌티바이러스 함유 상청액을 수집하고, Millex-HV 0.45-μm 저 단백질 결합 막 (Millipore)을 통해 여과하고, 분취하고, -80℃에서 저장하였다. 바이러스 역가를 결정하기 위해, 바이러스 분취의 연속 희석물을 폴리브렌 (8 μg/ml)의 존재 하에 HEK293T 세포로 형질도입시켰다. 형질도입되지 않은 세포와 연속 희석된 바이러스로 처리된 세포를 2 μg/ml 퓨로마이신 (Invitrogen)의 존재 하에 배양하였다. 거의 모든 형질도입되지 않은 세포가 죽었을 때, 바이러스 처리된 개체군에서 살아있는 세포의 수를 카운트하여 바이러스 역가를 추정하였다.

[0150] **실시예 1-5: 세포 라이브러리의 생성**

[0151] 렌티바이러스 형질도입을 준비하기 위해, HEK293T 세포를 9개의 150-mm 디쉬에 시딩하고 (디쉬 당 1.6 x 10^7 세포의 밀도), 밤새 인큐베이션 하였다. 렌티바이러스 라이브러리를 0.3의 MOI (multiplicity of infection)로 세포로 형질도입하여 올리고뉴클레오티드의 초기 수에 비해 500배 이상의 범위(coverage)를 달성하였다. 이어서, 세포를 밤새 인큐베이션한 후, 이후 5일 동안 2 μg/ml 퓨로마이신에서 유지하여 형질도입되지 않은 세포를 제거하였다. 이의 다양성을 보존하기 위해, 세포 라이브러리를 연구기간 동안 적어도 3.0 x 10^7 세포의 수로 유지하였다.

[0153] **실시예 1-6: 세포 라이브러리로의 PE2 전달**

[0154] 총 3.0 x 10^7 세포 (각각 1.0 x 10^7 세포를 함유하는 3개의 150-mm 배양 디쉬)를 80μl 리포펙타민 2000 (Thermo Fisher Scientific)을 사용하여 제조사의 지시에 따라 pLenti-PE2-BSD 플라스미드 (디쉬 당 80 μg)로 트랜스펙션 하였다. 배양 배지를 트랜스펙션 후 6시간에 10% 소태아혈청 및 20 μg/ml 블라스티시딘 S (InvivoGen)로 보충된 DMEM으로 교체하였다. 트랜스펙션 후 4.8일에, 세포를 수확하였다.

[0156] **실시예 2: 실험 방법 및 결과 측정**

[0157] **실시예 2-1: 내인성 부위에서 프라임에디터2(PE2) 효율의 측정**

[0158] 고처리량 실험의 결과를 검증하기 위해, 최종 플라스미드 라이브러리로부터 무작위로 33개의 개별 pegRNA 암호화 플라스미드를 선택하였다. 트랜스펙션을 준비하기 위해, HEK293T 세포를 16-18시간 전에 웰 당 5.0 x 10^4 또는 1.0 x 10^5 세포의 밀도로 48-웰 플레이트에 시딩하였다. 1,000 ng의 DNA 당 1 μl 의 리포펙타민 2000 또는 TransIT-2020 트랜스펙션 시약을 사용하여 제조사의 지시에 따라 세포를 PE2를 암호화하는 플라스미드 (pLenti-PE2-BSD, 1.0 x 10^4 세포 당 75 ng)와 pegRNA 암호화 플라스미드 (1.0 x 10^4 세포 당 25 ng)의 혼합물로 트랜스펙션 하였다. 밤새 인큐베이션 한 후, 배양 배지를 퓨로마이신 (2 μg/ml)을 함유하는 DMEM으로 교체하였다. 트랜스펙션 후 4.5일(Endo-BR1 및 Endo-BR2의 경우) 또는 7일(Endo-BR3)에, 세포를 수확하였다.

[0160] **실시예 2-2: HCT116 및 MDA-MB-231 세포주에서 PE2 효율의 측정**

[0161] HCT116 및 MDA-MB-231 세포를 각각 10%(v/v) FBS (fetal bovine serum)으로 보충된 DMEM 및 RPMI에서 5% CO_2의 존재 하에 37℃에서 각각 계대 배양하였다. PE2 발현 세포주를 생성하기 위해, PE2 암호화 렌티바이러스 벡터를 8 μg/ml 폴리브렌을 함유하는 배양 배지에서 MOI(multiplicity of infection) 0.3으로 HCT116 및 MDA-MB-231 세포로 형질도입하였다. 밤새 인큐베이션 한 후, 세포를 10 μg/ml 블라스티시딘 S의 존재 하에 7일 동안 배양하여 형질도입되지 않은 세포를 제거하였다.

[0162] pegRNA 암호화 서열 및 상응하는 표적 서열의 쌍을 함유하는 75개의 플라스미드를 플라스미드 라이브러리 1로부터 무작위로 선택하였다; 플라스미드 아이덴티티는 생어 염기서열 분석(Sanger sequencing)에 의해 결정하였다. 이어서, 플라스미드의 풀로부터 렌티바이러스 라이브러리를 생성하였다. PE2 발현 HCT116 및 MDA-MB-231 세포를 웰 당 2.0 x 10^5 세포의 밀도로 6-웰 플레이트에 시딩하고, 밤새 인큐베이션하고, 렌티바이러스 라이브러리로 형질도입하였다. 밤새 인큐베이션 한 후, 배양 배지를 HCT116 및 MDA-MB-231 세포주에 대해 각각 1 μg/ml 퓨로마이신 및 10 μg/ml 블라스티시딘 S를 함유하는 DMEM, 또는 2 μg/ml 퓨로마이신 및 10 μg/ml 블라스티시딘 S를 함유하는 RPMI로 교체하였다. 형질도입 4.5일 후에, 세포를 수확하고 분석하였다.

[0164] **실시예 2-3: 딥시퀀싱의 수행**

[0165] Wizard Genomic DNA purification kit(Promega)를 사용하여 수확된 세포로부터 게놈 DNA를 추출하였다.

[0166] 고처리량 실험을 위해, 통합된 바코드 및 표적 서열을 2X Taq PCR Smart mix(SolGent)를 사용하여 PCR 증폭하였다. 각각의 세포 라이브러리에 대해, 제1 PCR은 총 400 μg 의 게놈 DNA를 포함하였고; 10^6 세포 당 10 μg 게

놈 DNA를 가정하면, 적용 범위는 라이브러리 보다 700배 이상일 것이다. 반응 당 5 μg 의 초기 게놈 DNA 농도로 80개의 독립적인 50-μl PCR 반응을 수행한 후, 생성물을 풀링하고 MEGAquick-spin total fragment DNA purification kit (iNtRON Biotechnology)로 겔 정제하였다. 이어서, 100-ng 정제된 DNA를 Illumina 어댑터 및 바코드 서열을 모두 포함하는 프라이머를 사용하여 PCR에 의해 증폭시켰다.

[0167] 내인성 부위에서의 PE2 효율을 측정하기 위해, 독립적인 제1 PCR을 샘플 당 초기 게놈 DNA 주형 200 ng을 포함하는 40-μL 반응 부피에서 수행하였다. 이어서, Illumina 어댑터 및 바코드 서열을 부착시키기 위한 제2 PCR을 30 μl 반응 부피에서 제1 PCR로부터의 20 ng의 정제된 생성물을 사용하여 수행하였다. 겔 정제 후, 생성된 엠플리콘을 HiSeq 또는 MiniSeq (Illumina, San Diego, CA)를 사용하여 분석하였다.

[0169] **실시예 2-4: 프라임에디팅 효율의 분석**

[0170] 딥시퀀싱 데이터의 분석을 위해, 파이썬 스크립트 (Python scripts)를 사용하였다. 각각의 pegRNA 및 표적 서열 쌍은 22 nt 서열(18 nt 바코드 및 바코드의 상류에 위치한 4 nt 서열)을 통해 확인되었다. 넓은 표적 서열 내에 의도하지 않은 돌연변이가 없는 특정 편집을 포함하는 판독 (reads)은 PE2-유도된 돌연변이를 나타내는 것으로 간주되었다. 어레이 합성 및 PCR 증폭 절차에서 발생하는 백그라운드 프라임에디팅 빈도를 배제하기 위해, 아래에 나타낸 바와 같이 관찰된 프라임에디팅 빈도에서 PE2가 없을 때 측정된 백그라운드 프라임에디팅 빈도를 뺐다.

[0171] 프라임에디팅 효율 (%)

[0172] $$= \frac{\text{의도된 편집 및 특정 바코드를 갖는 판독 수} - (\text{특정 바코드를 갖는 전체 판독 수} \times \text{백그라운드 프라임에디팅 빈도}) + 100}{\text{특정 바코드를 갖는 전체 판독 수} - (\text{특정 바코드를 갖는 전체 판독 수} \times \text{백그라운드 프라임에디팅 빈도}) + 100} \times 100$$

[0173] 딥시퀀싱 데이터를 필터링하여 분석의 정확도를 개선하였다. 구체적으로, 딥시퀀싱 판독 카운트가 200 미만이고 백그라운드 프라임에디팅 빈도가 5%를 초과하는 pegRNA 및 표적 서열 쌍은 배제하였다.

[0175] **실시예 2-5: 특징 중요도 (feature importance)의 평가**

[0176] PE2 효율을 예측하기 위한 특징 중요도를 측정하기 위해, Tree SHAP method (XGBoost 알고리즘으로 통합된 SHapley Additive explanations)를 사용하였다(Lundberg, S.M. et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56-67 (2020)). 5배 교차 검증에서 결정된 최고의 하이퍼파라미터 구성으로 특징 및 훈련된 XGBoost 모델을 추출하였다. Tree SHAP 방법에서, 훈련된 XGBoost 모델의 각 특징에 샘플 당 중요도 점수가 할당되었다. 중요도 점수는, 모델 출력에서 기본값에 대한 특징의 효과를 나타내고, 최적의 신용 할당을 위한 게임 이론적 Shapley 값을 기반으로 계산되었다. 전체 데이터 세트에 대한 SHAP 값 분포를 보여주거나 평균 절대 값을 제공하여 모델에서 특징 중요도의 전반적인 개요를 제공하였다.

[0178] **실시예 2-6: 딥러닝-기반 계산 모델의 개발**

[0179] **(1) DeepPE의 개발**

[0180] DeepPE는 닉킹 사이트로부터 위치 +5에서 G에서 C로의 전환 돌연변이를 도입하는 PBS 및 RT 주형 길이의 최적 조합을 예측하는 딥러닝-기반 계산 모델이다.

[0181] 본 발명자들은 PE2 및 38,692개 pegRNA에 의해 유도된 프라임에디팅 효율로 구성된 훈련 데이터 세트를 사용하였다; 이러한 훈련 데이터는 47 nt 넓은 표적 서열, 17~37 nt RT 주형 + PBS 서열, 및 20개의 추가적인 특징 (예: 융해 온도, GC 수, GC 함량 및 최소 자가-폴딩 자유 에너지 등)을 포함한다. 뉴클레오티드 서열은 one-hot 인코딩에 의해 4차원 이진 매트릭스로 전환시켰다.

[0182] DeepPE는 컨볼루션 레이어 및 완전히 연결된 레이어를 사용하여 개발되었다.

[0183] 컨볼루션 레이어는 3 nt 길이의 10개 필터를 사용하여 넓은 표적 서열 및 RT 주형 + PBS 서열에서 2개의 임베딩 벡터를 얻었다. 이어서, 임베딩 벡터를 20개의 생물학적 특징과 연결시켰다. 딥 강화 학습 알고리즘이 로컬 정

보를 유지하기 위해 구현되었으므로, 풀링 레이어는 제외되었다.

[0184] 1,000 단위의 완전히 연결된 레이어는 벡터에 ReLU(Rectified-linear-unit) 활성 기능을 곱했다.

[0185] 회귀 출력 레이어는 출력의 선형 변환을 수행하고 PE2 효율에 대한 예측 점수를 계산하였다.

[0186] 9개의 서로 다른 모델(하이퍼파라미터; 컨볼루션 레이어 및 완전히 연결된 레이어 각각에 대해 필터(10, 20, 40) 및 유닛(200, 500, 1000)의 수)을 테스트한 후, 5배 교차 검증 동안 실험적으로 측정된 활성 수준과 예측된 활성 수준 사이의 가장 높은 Spearman 상관계수를 나타낸 모델을 선택하였다. 드롭아웃을 사용하여 0.3의 비율로 과적합을 피하였다. 목적 함수인 평균-제곱 오차, 및 학습 속도가 10^{-3}인 Adam optimizer를 사용하였다.

[0187] DeepPE는 TensorFlow를 사용하여 구현되었다.

[0189] **(2) PE_type 및 PE_position의 개발**

[0190] PE_type은 주어진 표적 서열에 대해 편집 유형에 따른 프라임에디팅 효율을 예측하는 딥러닝-기반 계산 모델이다.

[0191] PE_position은 주어진 표적 서열에 대해 편집 위치에 따른 프라임에디팅 효율을 예측하는 딥러닝-기반 계산 모델이다.

[0192] 다양한 편집 유형 및 위치에 대한 PE2 효율을 예측하기 위한 딥러닝-기반 알고리즘을 개발하기 위해, 컨볼루션 신경망 대신 다층 퍼셉트론(multilayer percentron, MLP)을 사용하였다. 교차 검증을 수행하여, DeepPE와 유사한 아키텍처와 파라미터 수를 갖지만 컨볼루션이 없는 18개의 MLP 모델 중에서 선택하였다. 고려된 하이퍼파라미터 구성은 다음과 같다: 레이어 수 ([2, 3]에서 선택됨), 각 히든 레이어에서 유닛 수 (제1 히든 레이어의 경우 [1000, 200, 50]에서 선택되고, 제2 히든 레이어의 경우 [50]에서 선택됨), 드롭아웃 정규화 파라미터, 학습 속도 ([0.01, 0.001, 0.0001]에서 선택됨), 및 ReLU 활성화 기능.

[0194] **실시예 2-7: 기존 기계 학습-기반 모델과의 비교**

[0195] **(1) 기계 학습을 위한 데이터 서브세트의 생성**

[0196] 라이브러리 1을 사용하여 얻은 PE2 효율 데이터를 계층화된 무작위 샘플링에 의해 HT-training 및 HT-test로 나누어, 동일한 표적 서열이 두 데이터 세트 간에 공유되지 않도록 하였다. 유사하게, 라이브러리 2를 사용하여 얻은 PE2 효율 데이터를 Type-training, Type-test, Position-training 및 Position-test로 나누어, 동일한 표적 서열이 훈련 데이터 세트 및 테스트 데이터 세트 간에 공유되지 않도록 하였다. 데이터 세트 Endo-BR1, Endo-BR2, Endo-BR3, HCT-BR1, HCT-BR2, MDA-BR1, 및 MDA-BR2의 생성에 사용된 표적 서열은 상응하는 테스트 데이터 세트에 포함시켜, 훈련 데이터 세트 및 테스트 데이터 세트 간에 표적 서열이 공유되지 않도록 하였다.

[0198] **(2) 기계 학습-기반 모델 훈련**

[0199] 기존의 기계 학습 알고리즘인 XGBoost, 그래디언트 부스티드 회귀 트리 (gradient-boosted regression tree), 랜덤 포레스트 (random forest), L1-정규화 선형 회귀 (L1-regularized linear regression), L2-정규화 선형 회귀 (L2-regularized linear regression), L1L2-정규화 선형 회귀 (L1L2-regularized linear regression), 및 SVM (support vector machine)을 기반으로 각각 학습하여 DeepPE의 성능과 비교하였다. 상기 모델들은 XGBoost 파이썬(Python) 패키지 (ver 0.90), scikit-learn (ver 0.19.1)로 구현하였다.

[0200] 넓은 표적 서열과 PBS 및 RT 주형 서열로부터 총 1,766개의 특징이 추출되었다. 그 특징은 위치-독립적 및 위치-의존적 뉴클레오티드 및 디뉴클레오티드, 융해 온도, GC 수, 및 넓은 표적 서열, PBS 및 RT 주형 서열의 최소 자가폴딩 자유 에너지, 및 DeepSpCas9 점수(Kim, H.K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci Adv* **5**, eaax9249 (2019))를 포함하였다. 융해 온도는 세포핵 환경을 고려하지 않고 기본 설정을 사용한 프로그램 (https://biopython.org/docs/1.74/api/Bio.SeqUtils.MeltingTemp.html)에 의해 계산되었다. 정규화 파라미터와 하이퍼파라미터 구성 중에서 모델 선택을 위해, 5배 교차 검증을 수행하였다.

[0201] XGBoost 및 그래디언트 부스티드 회귀 트리의 경우, 하기의 하이퍼파라미터 구성에서 선택된 144개가 넘는 모델을 검색하였다: 베이스 추정량의 수 ([5, 10, 50, 100]에서 선택됨), 개별 회귀 추정량의 최대 깊이 ([5, 10, 50, 100]에서 선택됨), 리프(leaf) 노드에 있는 최소 샘플 수 ([1,2,4]에서 선택됨), 학습 속도 ([0.05, 0.1, 0.2]에서 선택됨).

[0202] 랜덤 포레스트의 경우, XGBoost에 대해 학습 속도를 제외한 상기 나열된 동일한 하이퍼파라미터 구성에서 선택한 144개가 넘는 모델을 검색하였다; 최상의 분리(split)를 찾을 때 고려할 최대 특징 수를 검색하였다([모든 특징, 모든 특징의 제곱근, 모든 특징의 이진 로그(binary logarithm)]에서 선택됨).

[0203] L1-, L2 및 L1L2-정규화 선형 회귀의 경우, 정규화 파라미터를 최적화하기 위해, 로그 공간에서 10^{-6}과 10^{6} 사이에 균등한 간격으로 144개가 넘는 점을 검색하였다.

[0204] SVM의 경우, 하기 하이퍼파라미터로부터 144개가 넘는 모델을 검색하였다: 패널티 파라미터 C 및 커널 파라미터 γ, 10^{-3}과 10^{3} 사이에 균등한 간격으로 12개 점.

[0206] **실시예 2-8: 통계적 유의성**

[0207] 서로 다른 pegRNA를 사용한 실험 사이의 프라임에디팅 효율을 비교하기 위해, 일원분산분석 (one-way ANOVA) 후 Tukey의 사후검정을 사용하였다. 예측 모델의 예측 점수 간의 Spearman 상관관계를 비교하기 위해, 정확히 동일한 데이터 세트에서 두 개의 종속 상관계수를 테스트하는 방법인 Steiger의 테스트를 사용하였다. 카이-제곱 테스트를 수행하여 표적 서열 당 PBS 길이 및 RT 주형 길이의 가장 효율적인 조합이 선택될 때의 이들 두 파라미터 간의 관계를 결정하였다. 카이-제곱 분석의 정확도를 높이기 위해, 두 파라미터의 가장 효율적인 조합이 선택되었음에도 10% 미만의 프라임에디팅 효율을 나타내는 표적 서열은 분석에서 걸러내었다. DeepPE를 사용하거나 또는 주어진 표적 서열에서의 초기 연구의 권장사항을 사용하여 선택된 PBS 및 RT 주형 길이를 갖는 pegRNA의 PE2 효율을 비교하기 위하여, two-tailed paired t-테스트를 사용하였다. 통계적 유의성을 결정하기 위해, PASW Statistics (version 18.0, IBM) 및 Microsoft Excel (version 16.0, Microsoft Corporation)을 사용하였다.

[0209] **실시예 2-9: 데이터 가용성**

[0210] 이 연구의 딥시퀀싱 데이터는 NCBI Sequence Read Archive(SRA; https://www.ncbi.nlm.nih.gov/sra/)에 accession no. SRR11529289로 제출되었다.

[0212] **실험예 1: 프라임에디팅 효율 데이터의 수집**

[0213] PE2 효율의 고처리량 분석을 위해, 쌍 라이브러리 접근법을 사용하였다.

[0214] 도 1은 프라임에디팅 구성요소를 나타낸 개략도이다.

[0215] 도 2는 라이브러리 1 및 2의 구성을 나타낸 것이다.

[0216] 도 3은 pegRNA, cDNA 및 넓은 표적 서열 내에서 위치가 어떻게 지정되는지를 나타낸 개략도이다.

[0217] 본 발명자들은 48,000 쌍의 pegRNA-암호화 서열 및 상응하는 표적 서열(=2,000 표적 서열 × 24개 조합의 PBS 및 RT 주형/표적 서열)을 포함하는 올리고뉴클레오티드 풀로부터 라이브러리 1로 명명된 렌티바이러스 플라스미드 라이브러리를 제조하였다.

[0218] PE2 효율에 대한 PBS 및 RT 주형 길이의 영향을 테스트하기 위해, 라이브러리는 닉킹 부위(넓은 표적 서열 내에 위치 22)로부터 위치 +5에서 G에서 C로의 전환 돌연변이를 유도하는, 2,000 쌍의 가이드 및 표적 서열에 대한 24개의 상이한 PBS 및 RT 주형 길이의 조합(6개 PBS 길이(7, 9, 11, 13, 15, 17 nts) x 4개 RT 주형 길이(10, 12, 15, 20 nts) = 24개 조합)을 포함하였다. 즉, 48,000 (=24 x 2,000) 쌍의 pegRNA 및 표적 서열을 포함한다 (도 2).

[0219] 또한, PE2 효율에 대한 PBS 및 RT 주형 길이 이외의 인자의 영향을 평가하기 위해, 본 발명자들은 라이브러리 2

로 명명된 하나 이상의 라이브러리를 생성하였고, 이는 6,800 쌍의 pegRNA-암호화 서열 및 상응하는 표적 서열을 포함한다. 라이브러리 2를 사용하여 테스트한 인자는 편집 위치, 편집 유형(예: 삽입, 삭제 또는 치환), 및 2개-위치 편집의 위치를 포함한다(도 2).

[0220] 도 4는 프라임에디팅 효율의 고처리량 평가 절차의 개략도이다.

[0221] 도 4에 나타낸 바와 같이, HEK293T 세포를 플라스미드 라이브러리로부터 생성된 렌티바이러스로 형질도입하여 0.3 MOI에서 세포 라이브러리를 구축하고 형질도입되지 않은 세포는 퓨로마이신 선택에 의해 제거하였다. 이 라이브러리에서 각 세포는 pegRNA를 발현하고 상응하는 통합된 표적 서열을 포함한다. 이어서, 이 세포 라이브러리를 PE2를 암호화하는 플라스미드로 형질감염시키고 형질감염되지 않은 세포를 블라스티시딘 선별에 의해 제거하였다. PE2 플라스미드로 형질감염시키고 4일 반 후에, 게놈 DNA(genomic DNA)를 세포로부터 분리하고 PCR을 수행하여 표적 서열을 증폭시켰다. 엠플리콘을 딥시퀀싱하여 PE2에 의해 유도된 돌연변이 빈도를 측정하였다.

[0222] 생어 염기서열 분석에 따르면, 플라스미드 라이브러리에서 카피의 8.5% (=12/142)가 가이드 서열, 스캐폴드, PBS, RT 주형 또는 표적 서열 영역에서 하나 이상의 돌연변이를 함유하였고, 이는 올리고뉴클레오티드 합성 및 PCR 증폭 동안 도입된 오류일 수 있다. 또한, 렌티바이러스 벡터를 사용하여 고처리량 평가를 수행할 때, 두 개의 거리가 먼 요소가 섞일 수 있다. 세포 라이브러리에서 pegRNA 암호화 서열과 바코드-표적 서열 간의 비결합율을 측정한 결과, 4.2%로 나타났다. 이러한 돌연변이체 또는 비결합 서열에서 프라임에디팅이 거의 발생하지 않을 것으로 예상한다면, 관찰된 PE2 효율은 실제 PE2 효율의 87% (= 100% - 8.5% - 4.2%)일 것이다. 예를 들어, 실제 PE2 효율이 25%라면, 관찰된 PE2 효율은 25% x 87% = 22%일 것이다.

[0223] 도 5는 두 개의 다른 실험에 의해 독립적으로 PE2 암호화 플라스미드로 형질감염된 반복실험에서 PE 효율의 상관관계를 나타낸 것이다.

[0224] 도 5에 나타낸 바와 같이, 두 개의 다른 실험에 의해 독립적으로 형질감염된 반복실험 사이의 강한 상관관계를 관찰하였다. 후속 분석을 위해 두 반복실험의 데이터를 결합했다.

[0226] 다음으로, 고처리량 접근법을 사용하여, 통합된 서열에서 측정된 편집 효율과 개별 시험에 의해 평가된 내인성 부위에서의 편집 효율 사이의 상관관계를 결정하였다.

[0227] 도 6은 내인성 부위에서 측정된 PE 효율과 상응하는 통합된 표적 서열에서의 PE 효율 간의 상관관계를 나타낸 것이다.

[0228] 도 6에 나타낸 바와 같이, 초기 연구의 데이터 세트에서 Spearman의 상관계수 (R)=0.59, Pearson의 상관계수 (r)=0.69으로 나타나, 강한 상관관계가 있었다.

[0229] 또한, 라이브러리 1 및 2의 54,836개 pegRNA에서 무작위로 선별된 20 내지 31의 내인성 부위에서 PE2 효율의 새로운 6개 데이터 세트를 생성하였다. 생성된 데이터 세트는 Endo-BR1-TR1, Endo-BR1-TR2, Endo-BR2-TR1, Endo-BR2-TR2, Endo-BR2-TR3, Endo-BR3이다. 이들 실험에서, pegRNA 및 PE2를 암호화하는 플라스미드를 일시적으로 형질감염시켰다.

[0230] 도 7은 내인성 부위에서 측정된 PE 효율과 상응하는 통합된 표적 서열에서의 PE 효율 간의 상관관계를 나타낸 것이다.

[0231] 도 7에 나타낸 바와 같이, 내인성 부위에서의 PE2 효율 및 상응하는 통합된 표적 서열에서 PE2 효율 간의 높은 상관관계가 관찰되었다.

[0233] **실험예 2: 프라임에디팅 효율 데이터의 분석**

[0234] 상기 수집된 프라임에디팅 효율 데이터를 분석하였다.

[0235] 프라임에디팅을 위해, Cas9은 표적 서열과 결합하여 닉(nick)을 만들어야 한다. 따라서, PE2-pegRNA 및 Cas9-sgRNA의 활성은 높은 상관관계가 있을 것으로 예상되었다. 본 발명자들은 이전에 2,000개의 표적 서열에서 Cas9-sgRNA 활성과 관련된 indel 빈도를 평가했다(Kim, H.K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci Adv* **5**, eaax9249 (2019)).

[0236] 도 8은 SpCas9 유도 indel 빈도 및 동일한 표적 서열에서 결정된 PE2 효율 간의 상관관계를 나타낸 것이다.

[0237] 도 9는 라이브러리 1을 사용하여 SpCas9 유도 indel 빈도 및 동일한 표적 서열에서 결정된 PE2 효율 간의 상관관계를 나타낸 것이다.

[0238] 도 8 및 도 9에 나타낸 바와 같이, 동일한 표적 서열에서 PE2-pegRNA 및 Cas9-sgRNA의 활성의 연관성을 평가하였을 때, 적당한 상관관계가 관찰되었다. 강한 상관관계가 아닌 적당한 상관관계가 관찰된 이유는 프라임에디팅이 Cas9의 indel 생성 활성과 관련이 없는 추가적인 과정을 필요로 하기 때문일 것으로 생각되었다. 예를 들어, 이러한 과정은 pegRNA의 역전사, 5' 플랩 절단, 및 DNA 복구를 포함한다.

[0240] 주어진 표적 서열에서의 프라임에디팅의 경우, PBS 및 RT 주형 길이의 다양한 조합이 선택될 수 있고, pegRNA에서 이들 두 영역의 길이는 프라임에디팅 효율에 상당한 영향을 미친다. 그러므로, 다음으로 2,000개의 표적 서열에서 PE2 효율에 대한 상이한 PBS 및 RT 주형 길이의 영향을 평가하였다.

[0241] 도 10은 PBS 및 RT 주형 길이의 PE2 효율에 대한 영향을 나타낸 것이다. 히트맵은 주어진 길이의 PBS 및 RT 주형에서 평균 편집 효율을 나타낸다.

[0242] 도 11은 PBS 및 RT 주형 길이의 프라임에디팅 효율에 대한 영향을 나타낸 것이다. (A) RT 주형의 길이는 12 nt로 고정된 경우, 다양한 길이의 PBS에서의 PE 효율; (B) PBS의 길이는 13 nt로 고정된 경우, 다양한 길이의 RT 주형에서의 PE 효율.

[0243] 도 10 및 11에 나타낸 바와 같이, PBS 및 RT 주형 길이의 각각의 조합에 대해 평균 편집 효율을 계산하였을 때, 단봉분포를 보여주었고; 11 내지 13 nt PBS 및 10 내지 12 nt RT 주형을 갖는 pegRNA가 사용될 때, 최고 평균 효율 (13.4%)이 관찰되었다.

[0244] 도 12는 주어진 PBS 길이 및 RT 주형 길이에 대해 5% 이상의 PE2 효율을 갖는 pegRNA의 빈도이다.

[0245] 도 13은 (A) 주어진 PBS 길이 및 RT 주형 길이에 대해 5% 미만의 편집 효율을 갖는 pegRNA의 빈도; (B) 주어진 PBS 길이 및 RT 주형 길이에 대해 5% 이상의 편집 효율을 갖는 pegRNA의 빈도이다.

[0246] 도 12 및 13에 나타낸 바와 같이, PBS 및 RT 주형 길이에 따라 5% 미만의 PE2 효율을 갖는 것을 좋지 않은 pegRNA로 정의할 경우, pegRNA의 28%~81% (평균 43%)가 이 카테고리에 속하였다. 다시 말해, pegRNA의 19%~72% (평균 57%)는 PE2 효율이 5% 이상이었다.

[0247] 본 발명자들은 PBS 및 RT 주형 길이의 최적 조합은 표적 서열에 따라 가변적임을 발견하였다. 따라서, 다음으로 PBS 및 RT 주형 길이의 각 조합이 주어진 표적 서열 당 가장 높은 편집 효율을 얼마나 자주 유도하는지를 평가하였다.

[0248] 도 14는 주어진 표적 서열 당 가장 높은 편집 효율을 유도하는 PBS 및 RT 주형 길이 조합의 빈도를 나타낸 것이다.

[0249] 도 14에 나타낸 바와 같이, 이들 값도 단봉분포를 보여주었고, 가장 높은 편집 효율은 9 내지 13 nt PBS 및 10 내지 12 nt RT 주형이 사용되었을 때 가장 빈번하게 관찰되었다.

[0250]

[0251] 본 발명자들은 또한 각 표적에서 가장 효율적인 pegRNA를 선택할 때 PBS 및 RT 주형 길이의 각 조합의 평균 편집 효율을 비교하였다.

[0252] 도 15는 각 표적에서 가장 높은 편집 효율을 나타낸 PBS 및 RT 주형 길이의 조합을 선택할 때 평균 편집 효율을 나타낸 것이다.

[0253] 도 15에 나타낸 바와 같이, PBS 및 RT 주형 길이의 이러한 최적 조합에서 평균 편집 효율은 PBS 및 RT 주형의 길이가 짧을 때 가장 높았고(예를 들어, 7 nt PBS 및 10 내지 12 nt RT 주형), PBS 및 RT 주형 길이가 증가함에 따라 감소하였다.

[0254] 종합하면, 이러한 결과에 따르면 PE2 효율의 초기 테스트에 13 nt PBS 및 12 nt RT 주형을 사용하고, 두 번째 테스트에 9 내지 15 nt PBS 및 10 내지 15 nt RT 주형으로 확장하는 것이 권장된다는 결론을 얻을 수 있었다.

[0256] **실험예 3: 특징 중요도 평가**

[0257] 보다 체계적인 방식으로 PE2 효율과 관련된 다른 인자들을 평가하기 위해, 다음으로 pegRNA에서 다양한 영역의 융해 온도, GC 수, GC 함량, 및 최소 자가폴딩 자유 에너지, PBS 및 RT 주형의 길이, DeepSpCas9 점수(주어진 표적 서열에서 계산적으로 예측된 Cas9 뉴클레아제 활성)(Kim, H.K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci Adv* **5**, eaax9249 (2019)), 및 모든 위치-의존적 및 위치-독립적인 모노- 및 디뉴클레오티드와 같은 직접적인 서열 정보를 포함하는 1,766개의 특징을 사용하여 Tree SHAP 방법(Lundberg, S.M. et al. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 56-67 (2020).)을 수행하였다. 높은 특징 값이 높은 프라임에디팅 효율과 연결되었을 때, 그 특징은 선호(favored) 특징으로 분류하고; 높은 특징 값이 낮은 프라임에디팅 효율과 연결되었을 때, 그 특징은 비선호(unfavored) 특징으로 분류하였다.

[0258] 도 16은 Tree SHAP (XGBoost classifier)에 의해 결정된 PE2 효율과 연관된 가장 중요한 10개의 특징을 나타낸 것이다.

[0259] 도 17은 Tree SHAP에 의해 결정된 PE2 효율과 연관된 가장 중요한 1 내지 51번째 특징을 나타낸 것이다.

[0260] 도 18은 Tree SHAP에 의해 결정된 PE2 효율과 연관된 가장 중요한 52 내지 100번째 특징을 나타낸 것이다.

[0261] 첫 번째로 중요한 특징은 상응하는 표적 서열에서 DeepSpCas9 점수(favored)였으며(도 16), 이는 상기에서 나타낸 SpCas9 유도된 indel 빈도 및 PE2 효율 사이의 상관관계와 일치한다.

[0262] PBS에서 GC 수(favored)은 두 번째로 중요한 특징이었다. 이 결과와 함께, PBS에서의 GC 함량(favored)도 11번째로 가장 중요한 특징이었다(도 17). GC 함량은 GC 수(G 또는 C 뉴클레오티드 수)를 관련 DNA 가닥의 길이로 나누어 계산할 수 있다. 이 결과에 따르면, PBS에서 높은 GC 함량이 pegRNA의 표적 DNA의 닉 가닥에 대한 강한 결합을 초래한다는 것을 이해할 수 있으며, 이는 역전사에 필요하다.

[0263] 도 19는 PBS 및 RT 주형에서 GC 함량 및 GC 수의 프라임에디팅 효율에 대한 영향을 나타낸 것이다.

[0264] 도 19에 나타낸 바와 같이, PE2 효율에 대한 PBS, RT 주형, 및 PBS 및 RT 주형의 조합에서 GC 함량 및 GC 수의 영향을 체계적으로 평가하였을 때, PBS의 GC 함량 및 GC 수가 증가함에 따라 PE2 효율이 더 높음이 분명히 관찰되었다. PBS의 GC 함량이 30% 미만일 때, 15 nt와 같은 긴 길이에서 상대적으로 높은 편집 효율이 나타났으나, PE2 효율이 모든 테스트된 PBS 길이에 대해 좋지 못하였다. 반대로, PBS의 GC 함량이 60% 이상일 때, PBS를 7 내지 11 nt의 길이로 단축시키는 것이 비교적 높은 PE2 효율을 초래하였다. 이러한 결과에 기초하여, GC 함량이 각각 40% 미만 또는 60% 이상일 경우 각각 길이가 15 또는 9 nt인 PBS를 사용하는 것이 권장된다.

[0265] 그러나, RT 주형의 GC 함량 및 GC 수는 PE2 효율에 약간의 영향만을 미쳤고, GC 관련 파라미터가 극히 높거나 낮을 때 PE2 효율이 낮아지는 경향이 있었다. 이러한 결과와 호환되게, RT 주형의 GC 함량 또는 GC 수는 40가지의 가장 중요한 특징에 포함되지 않았다.

[0266] 세 번째 및 다섯 번째로 중요한 특징은 각각 PBS의 융해 온도(favored) 및 RT 주형에 상응하는 표적 DNA 영역의 융해 온도였다(즉, 프로토스페이서 인접 모티프 (protospacer adjacent motif; PAM)를 함유하는 가닥과 반대 가닥 사이; 본원에서 "PAM-반대 가닥"이라고 함; 이 특징은 융해 온도가 35℃보다 높을 때만 disfavored함). 높은 PBS 융해 온도는 PBS에서 높은 GC 수와 연관될 가능성이 있고, 표적 DNA에 대한 pegRNA의 PBS 영역의 강한 결합과 연결되어, 역전사 반응을 촉진할 것이다.

[0267] 도 20은 PBS의 융해 온도 및 RT 주형에 상응하는 표적 DNA 영역의 프라임에디팅 효율에 대한 영향을 나타낸 것이다.

[0268] 도 20에 나타낸 바와 같이, PE2 효율과 PBS 융해 온도 간의 관계를 조사한 결과, PBS 융해 온도가 증가함에 따라 PE2 효율도 증가하는 것을 확인하였다. RT 주형에 상응하는 표적 DNA 영역의 융해 온도가 너무 높으면, 3' 플랩의 5' 플랩으로의 전환, 즉, 역전사된 DNA 서열을 게놈에 통합시키기 위해 필요한 과정이 방지될 수 있다. PE2 효율과 이 영역의 융해 온도 간의 관계를 분석하였고, 융해 온도가 35℃ 이상으로 증가할 때, 차이가 통계적으로 유의하지는 않지만 PE2 효율이 감소하는 경향이 있음을 확인하였다.

[0269] 네 번째로 중요한 특징은 RT + PBS 영역에서 UU의 수이다(disfavored). 이 특징은 pegRNA에서 다수의 U에 상응하는 pegRNA-암호화 서열에서 다수의 T에 기인하며, 이는 RNA 폴리머라제 III에 의한 전사 효율을 감소시켜, 세

포내 pegRNA 농도를 감소시킬 수 있다.

[0270] 여섯 번째 및 여덟 번째로 중요한 특징은 각각 넓은 표적 서열에서 위치 16에서의 T의 존재(disfavored) 및 위치 17에서의 C의 존재(favored)였다(위치 1은 NGG PAM으로부터 20번째 뉴클레오티드). 이전 연구에 따르면, 위치 16에서의 T는 감소된 Cas9 뉴클레아제 활성과 연관된다. 또한, 위치 16에서의 T는 PBS에서 GC 수를 감소시키며, 이는 역전사, 특히 PBS의 길이가 짧을 때에 바람직하지 않다. 이 두 가지 효과를 결합하면 위치 16에서의 T가 여섯 번째로 중요한 특징이 된다. 유사하게, 이전 연구에 따르면, A 또는 C가 위치 17에 있을 때, Cas9 뉴클레아제 활성이 증가하였다. 또한, 위치 17에서의 C는 PBS에서 GC 수를 증가시켜, 역전사를 용이하게 한다. 이 두 가지 효과의 조합은 위치 17에서의 C를 favored한 특징으로 만든다.

[0271] 일곱 번째, 아홉 번째, 및 12번째로 중요한 특징은 RT 및 PBS 길이(일반적으로 disfavored), RT 주형 길이(길이가 길 때만 disfavored), 및 PBS 길이(일반적으로 disfavored)였다.

[0272] 10번째로 중요한 특징은 넓은 표적 서열에서 위치 24의 G이다(disfavored). 의도된 편집(+5 G에서 C)은 위치 22에서 G를 교체할 것이고, 이는 PAM 편집을 야기하여, Cas9이 표적 서열에 재결합하는 것을 막을 것이다.

[0274] **실험예 4: 다양한 종류의 편집에 대한 프라임에디팅 효율 평가**

[0275] 다음으로, 라이브러리 2에서 6,800개 pegRNA와 표적 서열 쌍(= 200개 표적 서열 x 1 PBS/표적 서열 x 34개 RT 주형/표적 서열)을 사용하여 더욱 다양한 종류의 게놈 편집에 대해 PE2 효율을 평가하여, 게놈 편집의 유형(즉, indel vs. 치환의 생성), 편집된 위치, 및 삽입되거나 결실된 뉴클레오티드의 수의 상기 PE2 효율에 대한 영향을 결정하였다.

[0276] 도 21은 1-bp 삽입, 결실, 및 치환의 경우 PE2 효율을 나타낸 것이다.

[0277] 도 22는 삽입된 뉴클레오티드 유형 및 수의 PE2 효율에 대한 영향을 나타낸 것이다.

[0278] 도 23은 결실 길이의 PE2 효율에 대한 영향을 나타낸 것이다.

[0279] 먼저, 1-bp 삽입, 1-bp 결실, 및 1-bp 치환을 생성하는 효과를 평가하였다. 일반적인 효율은 삽입 ≥ 결실 ≥ 치환으로 순위를 매길 수 있으며, 삽입과 치환 효율 간의 차이는 통계적으로 유의함을 확인하였다(도 21).

[0280] 그 다음, 삽입된 뉴클레오티드의 유형 및 수의 프라임에디팅 유도 삽입에 대한 영향을 평가하였다. 삽입된 뉴클레오티드의 아이덴티티가 1-bp 삽입 효율에 영향을 미치지 않음을 확인하였다. 삽입된 뉴클레오티드의 수를 1 bp에서 2, 5, 및 10 bp로 증가시켰을 때, 삽입 효율은 1- 및 2-bp 삽입은 비슷하고, 5-bp 삽입에 대해서는 감소하였으며, 10-bp 삽입에 대해서는 크게 감소하였다(도 22).

[0281] 동시에, 1-, 2-, 5-, 및 10-bp 결실에 대한 PE 효율을 평가하였고, PE 효율이 1-, 2-, 및 5-bp 결실에 대해 비슷하고, 10-bp 결실에 대해서는 크게 감소하였다(도 23).

[0283] 다음으로, 치환된 뉴클레오티드 아이덴티티의 PE2 효율에 대한 영향을 조사하였다.

[0284] 도 24는 치환 유형의 PE2 효율에 대한 영향을 나타낸 것이다.

[0285] 도 24에 나타낸 바와 같이, 넓은 표적 서열에서 위치 17과 18 사이에 해당하는, 닉킹 부위로부터 위치 +1에서 모든 12개의 가능한 유형의 1-bp 치환을 테스트하였고, PE2 효율이 치환의 유형에 따라 약간 다르다는 것을 확인하였다; C에서 T로의 변환 및 T에서 G로의 변환은 각각 가장 높은 PE2 효율과 가장 낮은 PE2 효율을 보여주었다. 이러한 영향에 대한 기계적인 통찰력을 얻기 위해, RT 주형으로부터 생성된 cDNA에서 뉴클레오티드와 PAM-반대 가닥에서 상응하는 뉴클레오티드 사이의 임시 염기쌍을 고려하였다. 흥미롭게도, PE2 효율은 다음과 같이 순위화되었다: T (cDNA) - G (PAM-반대 가닥에서 상응하는 뉴클레오티드)와 G - T 쌍 ≥ C - T와 T - C 쌍 ≥ C - A와 A - C 쌍 ≥ A - G와 G - A 쌍. 여기에서, T - G와 G - T 쌍 그룹과 A - G와 G - A 쌍 그룹 간의 차이는 통계적으로 유의하였으므로, cDNA와 PAM-반대 가닥 사이의 임시 염기쌍이 PE2 효율에 영향을 줄 수 있음을 암시하였다. 동일한 뉴클레오티드 사이에 임시 염기쌍이 형성되었을 때, 예를 들어 T (cDNA) - T (PAM-반대 가닥에서 상응하는 뉴클레오티드), G - G, C- C, 및 A - A, 이는 각각 A에서 T로, C에서 G로, G에서 C로, 및 T에서 A로의 전환에 대응하는데, PE2 효율이 모두 비슷했다.

[0286] 또한, 닉킹 부위로부터 +9, +11, 및 +14와 같은 상이한 위치에서 동일한 뉴클레오티드 사이의 임시 염기쌍에 의

해 매개되는 이들의 4개의 변환에 대한 PE2 효율을 분석하였다.

[0287] 도 25는 치환의 유형의 프라임에디팅 효율에 대한 영향을 나타낸 것이다.

[0288] 도 25에 나타낸 바와 같이, 모든 3개의 테스트 된 위치에서 4개의 테스트된 변환에 대해 비슷하였고, 이는 닉킹 부위로부터 위치 +1에서의 분석과 유사하였다.

[0290] 또한, 1-bp 치환 효율에 대한 편집 위치의 영향을 조사하였다.

[0291] 도 26은 1-bp 변환 치환의 경우 편집 위치의 PE2 효율에 대한 영향을 나타낸 것이다.

[0292] 도 26에 나타낸 바와 같이, 편집 효율은 닉킹 부위로부터 +1 내지 +14 범위의 모든 테스트된 위치에서 위치 +3, +5, 및 +6을 제외하고 일반적으로 비슷하였다. 이 영향에 대한 기본 메커니즘은 명확하지 않지만, 위치 +3에서 가장 낮은 편집 효율이 관찰되었다. 가장 높은 편집 효율은 위치 +5 및 +6, GG PAM의 위치에서 관찰되었다; 전술한 바와 같이, PAM이 편집되지 않으면, Cas9은 표적 서열에 재결합하고 상보적 가닥의 수리 전에 역전사된 DNA 가닥을 닉킹하여, PE 효율을 감소시킬 수 있다.

[0293] PE 효율에 대한 PAM 편집의 이 영향은 2-bp 치환 효율이 평가되었을 때에도 관찰될 수 있다.

[0294] 도 27는 2개 위치에서 1-bp 변환 치환의 경우 편집 위치의 프라임에디팅 효율에 대한 영향을 나타낸 것이다.

[0295] 도 27에 나타낸 바와 같이, 다양한 위치에서 2-bp 치환을 생성하였고, PAM이 그대로 남았을 때 (위치 1 및 2, 위치 1 및 10, 위치 2 및 3, 위치 2 및 10, 또는 위치 10 및 11이 편집됨)보다 PAM에서 하나 또는 둘 모두의 뉴클레오티드 (위치 5 및 6)가 편집되었을 때(예: 위치 1 및 5, 위치 2 및 5, 위치 5 및 6, 위치 5 및 10), 편집 효율이 높았다.

[0296] 도 28은 도 27에 설명된 두 개의 편집 위치 간의 거리에 따른 일부 편집의 상대적 빈도를 나타낸 것이다.

[0297] 도 29는 두 개의 뉴클레오티드가 치환의 목적일 때 프라임에디팅 분석 결과를 나타낸 것이다.

[0298] 편집 위치가 PE2 효율에 영향을 미치는 경우, 야생형 SpCas9 대신 다른 PAM을 인식하는 SpCas9 변이체를 사용하면 동일 표적 서열에서 PE2 효율을 향상시킬 수 있다. 흥미롭게도, 두 개의 의도된 편집 중 적어도 하나가 도입된 서열의 최대 20%까지의 중앙값이 오직 1개의 편집만 가졌다(도 28 및 도 29). 이러한 부분 편집률은 닉킹 부위와 가까운 위치에서보다 먼 위치에서 더 높았고, 두 위치 간의 거리가 증가함에 따라 증가하는 경향을 보였다.

[0300] **실험예 5: 딥러닝 기반 예측 모델 검증 1**

[0301] **(1) 특정 유형의 편집에서 PBS 및 RT 주형 길이에 따른 PE2 효율을 예측하기 위한 모델 DeepPE의 생성**

[0302] 실시예 2-6에 따라, 가변 PBS 및 RT 주형 길이를 갖는 24개의 서로 다른 pegRNA와 쌍을 이루는 주어진 표적 서열에서 PE2 효율을 예측하는 계산 모델을 개발하였다.

[0303] 48,000 쌍의 pegRNA와 표적 서열을 갖는 라이브러리 1을 사용하여 얻은 PE 효율은 무작위 샘플링에 의해 2개의 데이터 세트로 나누고, 각각 HT-Training (n = 38,692) 및 HT-Test (n = 4,457)로 명명하였다. 이때, 두 개의 데이터 세트 간에 동일한 표적 서열을 공유하지 않도록 하였다. HT-training을 훈련 데이터로 사용하여, 프라임 에디팅이 위치 +5에서 G에서 C로의 변환을 위해 설계된 경우 PBS와 RT 주형 길이의 서로 다른 조합을 갖는 24개의 pegRNA와 쌍을 이루는 주어진 표적 서열에서 PE2 효율을 예측하기 위한 계산 모델을 생성하였다.

[0305] **(2) 성능 검증**

[0306] 도 30은 사용된 기계 학습 프레임워크에 따른 예측 모델의 교차 검증 결과를 나타낸 것이다.

[0307] 도 30에 나타낸 바와 같이, 교차 검증 결과, 딥러닝 프레임워크가 두 번째로 우수한 프레임워크인 boosted RT와의 차이가 통계적으로 유의하지는 않았으나 가장 높은 성능을 가짐을 보여주었다.

[0308] 도 31은 데이터 세트 HT-Test (pegRNA 및 표적 서열 쌍의 수 n = 4,457) 및 Endo-BR1-TR1 (n = 26)를 사용한

DeepPE의 평가 결과를 나타낸 것이다.

[0309] 도 32는 DeepPE를 데이터 세트 HT-Test를 사용한 다른 예측 모델과 성능 비교한 결과이다.

[0310] 도 33은 pegRNA 및 PE2를 암호화하는 플라스미드를 HEK293T 세포로 일시적 트랜스펙션 한 후, 내인성 부위에서 PE2 효율을 측정하여 얻은 6개의 데이터 세트를 사용한 DeepPE의 평가 결과를 나타낸 것이다.

[0311] 도 31 내지 33에 나타낸 바와 같이, 테스트 데이터 세트로서 HT-test를 사용하여 평가한 결과, 딥러닝 기반 모델인 DeepPE는 기존 기계 학습을 기반으로 한 다른 모델을 능가하였다. 테스트 데이터 세트로서 내인성 부위에서의 PE2 효율의 6개 반복실험을 사용하여 테스트한 결과, Spearman 및 Pearson 상관계수(R 및 r)는 각각 R = 0.67~0.77 (평균 0.73) 및 r = 0.63~0.74 (평균 0.69)였고, 이는 DeepPE의 내인성 부위에서 PE2 효율을 예측하는 성능이 우수함을 나타낸다.

[0312] DeepPE 훈련에 사용된 적 없는 표적 서열에서의 두 개의 추가적인 세포 유형, HCT116 및 MDA-MB-231에서 DeepPE를 평가하였다.

[0313] 도 34는 HCT116 및 MDA-MB-231 세포를 사용한 DeepPE의 평가 결과를 나타낸 것이다.

[0314] 도 34에 나타낸 바와 같이, 생물학적 및 기술적 반복실험에 걸쳐 DeepPE는 우수한 성능을 나타내었다. HCT116, R = 0.70~0.77 (평균 0.74), r = 0.57~0.61 (평균 0.59); MDA-MB-231, R = 0.76~0.81 (평균 0.79), r = 0.62~0.65 (평균 0.64)로 나타났다.

[0316] 주어진 표적 서열에 대해 PBS 및 RT 주형 길이의 가장 효율적인 조합 (24개의 가능한 조합 중에서)을 선택하기 위한 DeepPE의 유용성을 확인하였다.

[0317] 도 35은 주어진 표적 서열에서 PBS 및 RT 주형 길이의 가능한 24개 조합 중에서 가장 효율적인 조합을 선택하기 위한 DeepPE 및 방법의 성능 비교를 나타낸 것이다. 예를 들어, "13-nt PBS & 12 nt-PT template"란 표적 서열에 관계 없이 이러한 길이의 조합을 선택하는 것을 의미한다. 초기 연구 권장사항 A 및 B는 13-nt PBS 및 12-nt RT 주형(RTT)을 사용하고, 필요에 따라 RTT 길이를 변경하는 것에 의해 마지막 주형 뉴클레오티드로서 G를 사용하지 않는 것을 기반으로 한다. 권장사항 A에서, 마지막 주형 뉴클레오티드가 G이면, 12-nt 보다 10-nt RTT가 선택된다. 이러한 변경 후 마지막 주형 뉴클레오티드가 다시 G이면, 15-nt RTT가 선택된다. 권장사항 B에서, 마지막 주형 뉴클레오티드가 G이면, 12-nt 보다 15-nt RTT가 선택된다. 이러한 변경 후에 마지막 주형 뉴클레오티드가 다시 G이면, 10-nt RTT가 선택된다. 대조군으로서, pegRNA를 무작위로 선택하였다(Random 1 및 Random 2).

[0318] 도 35에 나타낸 바와 같이, DeepPE를 사용하였을 때 평균 절대 및 상대 PE2 효율은 각각 1.2% 및 8.3%였다. 이는 초기 연구를 기반으로 한 권장사항(즉, 13 nt PBS 및 12 nt RT 주형을 사용하며, 마지막 주형 뉴클레오티드에 G를 사용하지 않음)을 사용하여 얻은 효율보다 유의적으로 높았다.

[0319] 또한, 의도된 편집을 위해, 다수의 표적 서열이 있을 수 있다; 이 경우, DeepPE는 가장 높은 효율로 편집될 수 있는 표적 서열을 선택하기 위해 유용할 것이다.

[0321] **실험예 6: 딥러닝 기반 예측 모델 검증 2**

[0322] **(1) 편집 유형 및 위치에 따른 PE2 효율을 예측하기 위한 모델 PE_Type 및 PE_position의 생성**

[0323] 실시예 2-6에 따라, 라이브러리 2를 사용하여 얻은 데이터 세트를 사용하여 편집 유형에 따른 PE2 효율을 예측하기 위한 계산 모델 PE_Type 및 편집 위치에 따른 PE2 효율을 예측하기 위한 계산 모델 PE_position을 개발하였다.

[0324] 라이브러리 2를 사용하여 얻은 데이터는 Type-training, Type-test, Position-training, 및 Position-test로 나누어 훈련 데이터 세트와 테스트 데이터 세트 간에 표적 서열이 공유되지 않도록 하였다.

[0326] **(2) 성능 검증**

[0327] 도 36은 사용된 기계 학습 프레임워크에 따른 PE_type의 교차 검증 결과를 나타낸 것이다.

[0328] 도 37은 사용된 기계 학습 프레임워크에 따른 PE_position의 교차 검증 결과를 나타낸 것이다.

[0329] 도 36 및 37에 나타낸 바와 같이, Type-training 및 Position-training을 사용한 교차 검증 결과, 랜덤 포레스트가 가장 우수한 성능을 가졌으나, 두 번째로 우수한 프레임워크와의 차이는 통계적으로 유의하지 않았다. 두 가지 경우에서, 딥러닝은 상대적으로 적은 수의 표적 서열 및 pegRNA로 인해 제한된 성능을 나타냈다. Type-test 및 Position-test를 사용하여 평가하였을 때, PE_type 및 PE_position, 랜덤 포레스트-기반 모델은 유용한 성능을 나타냈다. PE_type, R = 0.47, r = 0.48; PE_position, R = 0.56, r = 0.56.

[0330] 따라서, 모든 가능한 PBS 및 RT 주형 길이를 갖는 pegRNA 및 더욱 다양한 의도된 편집을 사용하여 더 많은 수의 표적 서열에서 프라임에디팅 효율을 평가하는 것이 더욱 유용한 모델을 생산할 수 있을 것이다.

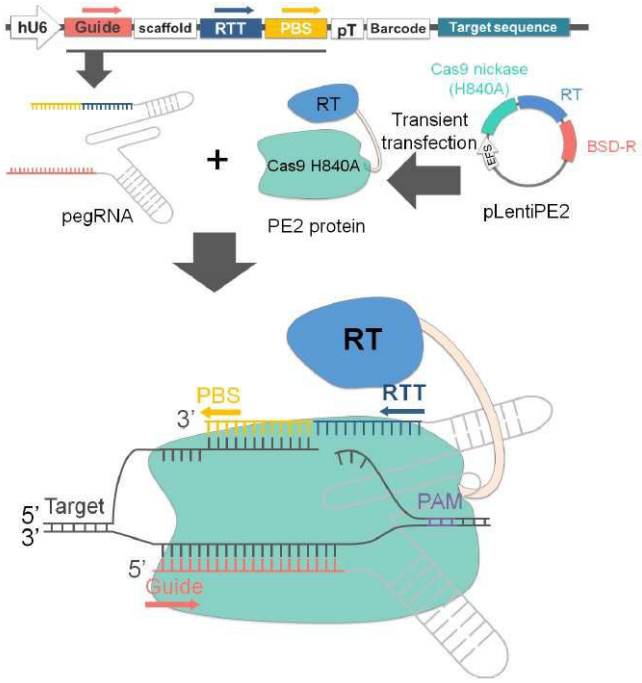
[0332] 본 발명자들은 주어진 표적 서열에 대한 DeepPE, PE_type, 및 PE_position의 결과를 제공하는 웹 툴을 http://deepcrispr/DeepPE에서 제공한다. 표적 서열을 포함하는 서열을 입력하면, 상기 웹 툴은 후보 표적 서열을 식별하고, 표적 서열 당 총 57개의 pegRNA (DeepPE에서 24개 pegRNA, PE_type에서 23개 pegRNA, 및 PE_position에서 10개 pegRNA)에 대해 예상되는 PE2 효율을 제공한다.

[0334] 프라임에디팅은 donor DNA를 사용하지 않고도 상당히 효율적인 방식으로 작은 유전적 돌연변이가 도입될 수 있다는 점에서 혁명적이다. DeepPE, PE_type, 및 PE_positin과 함께, 고처리량 분석을 기반으로 한 본 연구에서 확인된 PE2 효율에 영향을 미치는 인자에 대한 정보는 프라임에디팅을 촉진시킬 것으로 기대한다.

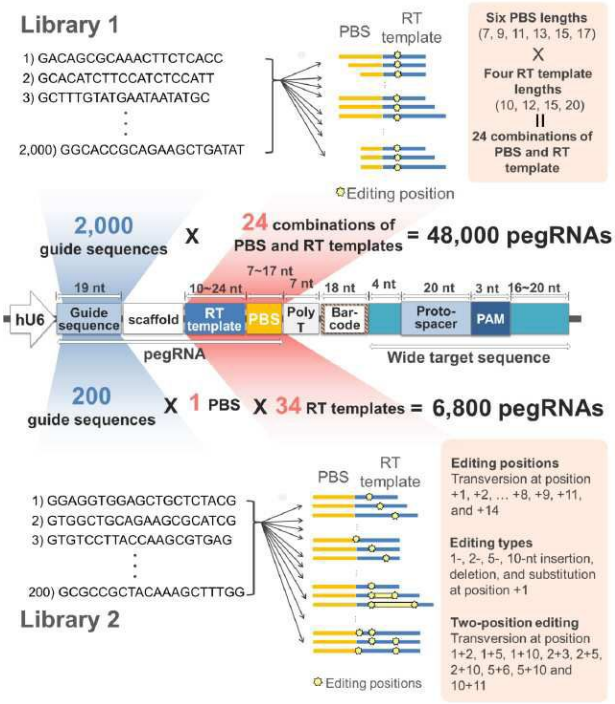
[0335] 상기와 같이, 본 발명자들은 인간 세포에서 54,836쌍의 pegRNA 및 표적 서열을 사용하여 프라임에디터2(PE2) 활성의 고처리량 평가를 수행하였다. PE2 효율의 큰 데이터 세트를 통해 i) 주어진 표적 서열에서 상이한 길이의 PBS 및 RT 주형을 갖고, 상이한 위치에서 다양한 유형의 의도된 편집을 생성하도록 지정된 총 57개 pegRNA에 대해 PE2 효율을 예측하는 계산 모델을 개발하였고, ii) 고도로 체계적인 방식으로 PE2 효율에 영향을 미치는 다수 인자를 식별하였다. 상기 계산 모델 및 PE2 효율에 대한 정보는 프라임에디팅을 촉진시킬 것이다.

도면

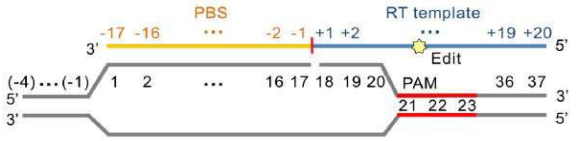
도면1



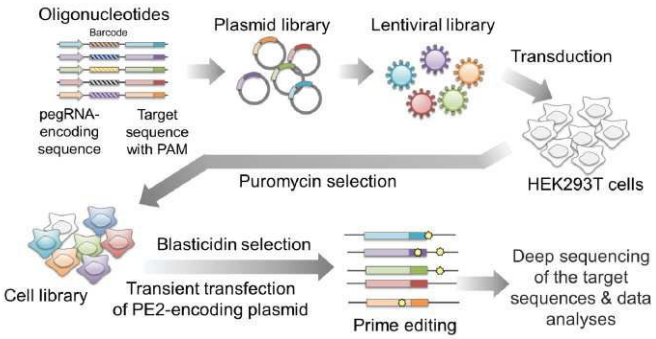
도면2


Library 1
1) GACAGCGCAAACTTCTCACC
2) GCACATCTTCCATCTCCATT
3) GCTTTGTATGAATAATATGC
2,000) GGCACCGCAGAAGCTGATAT
PBS
RT template
Editing position
Six PBS lengths
(7, 9, 11, 13, 15, 17)
X
Four RT template lengths
(10, 12, 15, 20)
=
24 combinations of PBS and RT template
2,000 guide sequences
X
24 combinations of PBS and RT templates
= 48,000 pegRNAs
19 nt
10~24 nt
7~17 nt
7 nt
18 nt
4 nt
20 nt
3 nt
16~20 nt
hU6
Guide sequence
scaffold
RT template
PBS
Poly T
Bar-code
Proto-spacer
PAM
pegRNA
Wide target sequence
200 guide sequences
X 1 PBS X 34 RT templates = 6,800 pegRNAs
1) GGAGGTGGAGCTGCTCTACG
2) GTGGCTGCAGAAGCGCATCG
3) GTGTCCTTACCAAGCGTGAG
200) GCGCCGCTACAAAGCTTTGG
Library 2
Editing positions
Editing positions
Transversion at position +1, +2, … +8, +9, +11, and +14
Editing types
1-, 2-, 5-, 10-nt insertion, deletion, and substitution at position +1
Two-position editing
Transversion at position 1+2, 1+5, 1+10, 2+3, 2+5, 2+10, 5+6, 5+10 and 10+11

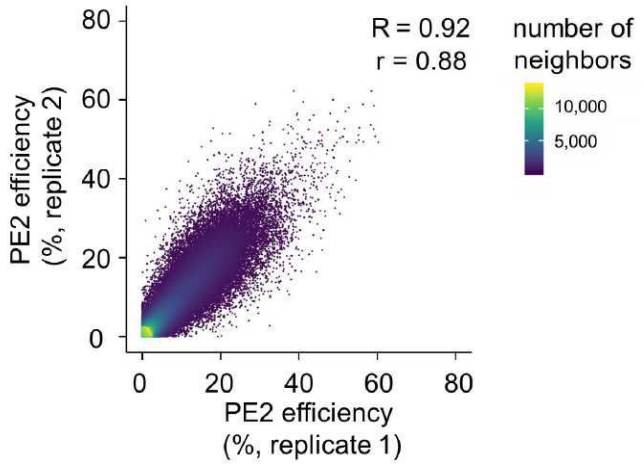
도면3


PBS
RT template
-17 -16
-2 -1 +1 +2
+19 +20
3'
5'
Edit
(-4) (-1)
1 2
16 17 18 19 20
PAM
36 37
21 22 23
5'
3'
3'
5'

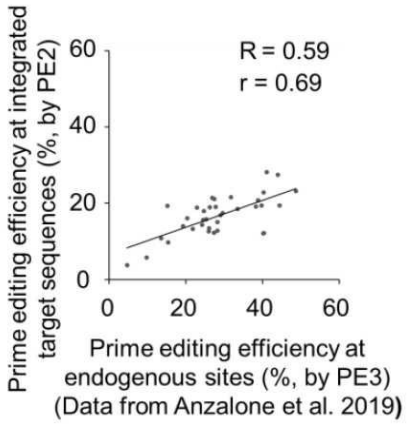
도면4


Oligonucleotides
Barcode
pegRNA-encoding sequence
Target sequence with PAM
Plasmid library
Lentiviral library
Transduction
HEK293T cells
Puromycin selection
Cell library
Blasticidin selection
Transient transfection of PE2-encoding plasmid
Prime editing
Deep sequencing of the target sequences & data analyses

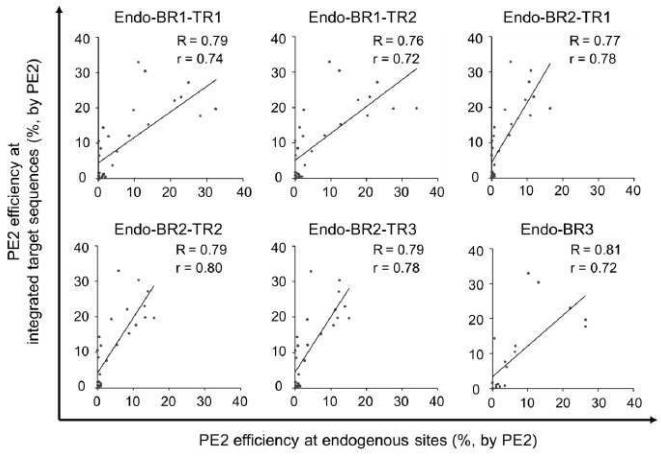
도면5


R = 0.92
r = 0.88
number of neighbors
10,000
5,000
PE2 efficiency (%, replicate 2)
PE2 efficiency (%, replicate 1)
0
20
40
60
80

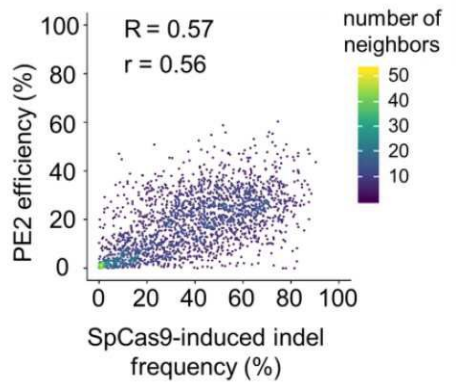
도면6


R = 0.59
r = 0.69
Prime editing efficiency at integrated target sequences (%, by PE2)
Prime editing efficiency at endogenous sites (%, by PE3)
(Data from Anzalone et al. 2019)
0
20
40
60

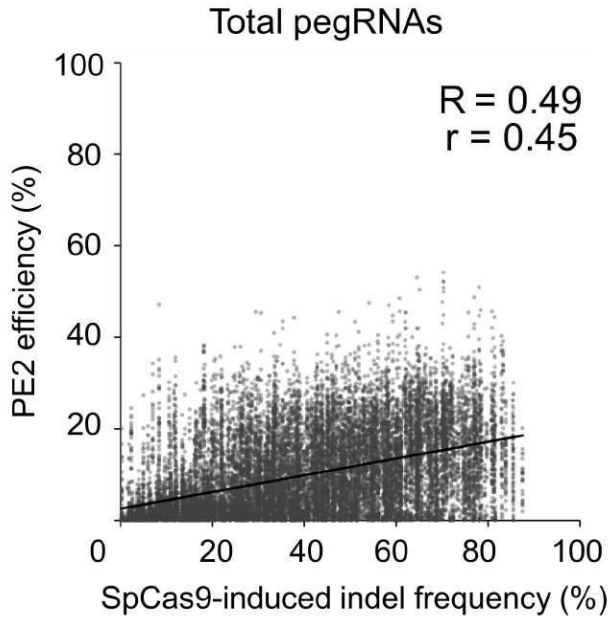
도면7


Endo-BR1-TR1
R = 0.79
r = 0.74
Endo-BR1-TR2
R = 0.76
r = 0.72
Endo-BR2-TR1
R = 0.77
r = 0.78
Endo-BR2-TR2
R = 0.79
r = 0.80
Endo-BR2-TR3
R = 0.79
r = 0.78
Endo-BR3
R = 0.81
r = 0.72
PE2 efficiency at integrated target sequences (%, by PE2)
PE2 efficiency at endogenous sites (%, by PE2)

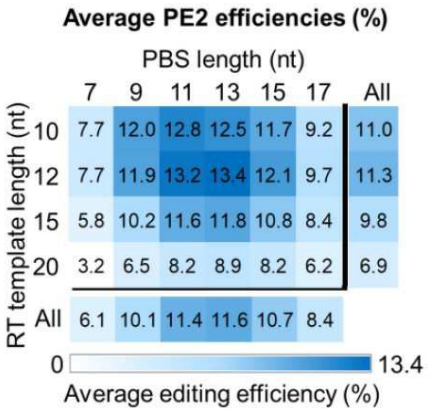
도면8


R = 0.57
r = 0.56
number of neighbors
50
40
30
20
10
PE2 efficiency (%)
0
20
40
60
80
100
SpCas9-induced indel frequency (%)

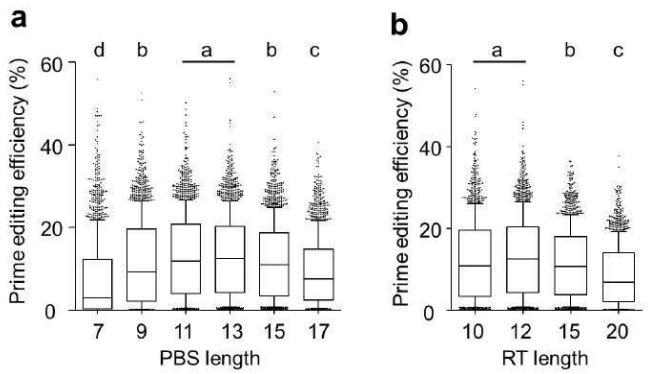
도면9


Total pegRNAs
R = 0.49
r = 0.45
PE2 efficiency (%)
SpCas9-induced indel frequency (%)

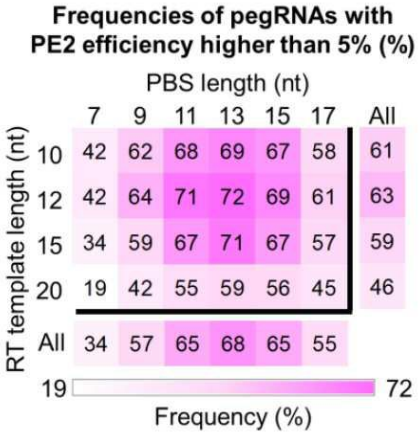
도면10


Average PE2 efficiencies (%)
PBS length (nt)
7 9 11 13 15 17 All
RT template length (nt)
10 7.7 12.0 12.8 12.5 11.7 9.2 11.0
12 7.7 11.9 13.2 13.4 12.1 9.7 11.3
15 5.8 10.2 11.6 11.8 10.8 8.4 9.8
20 3.2 6.5 8.2 8.9 8.2 6.2 6.9
All 6.1 10.1 11.4 11.6 10.7 8.4
0
13.4
Average editing efficiency (%)

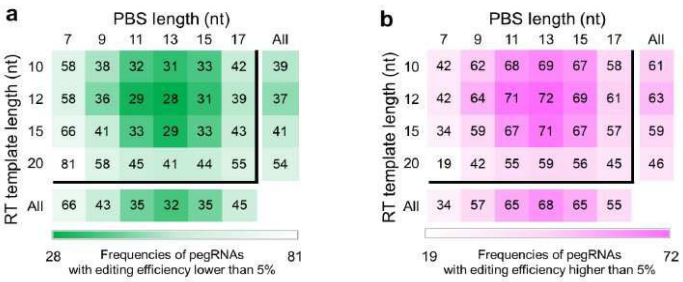
도면11



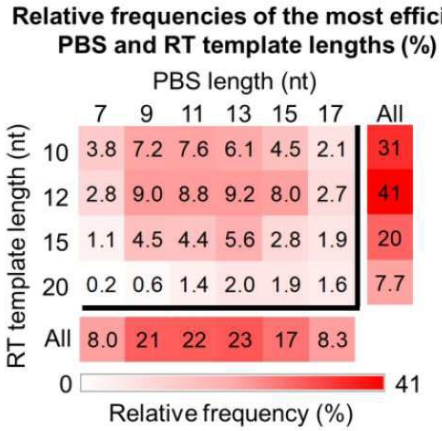
도면12



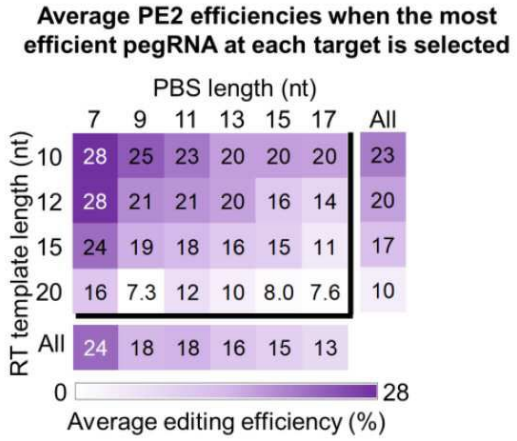
도면13



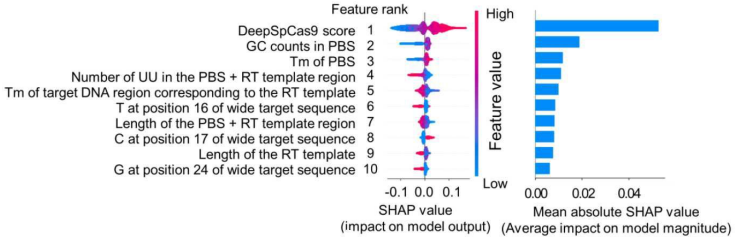
도면14



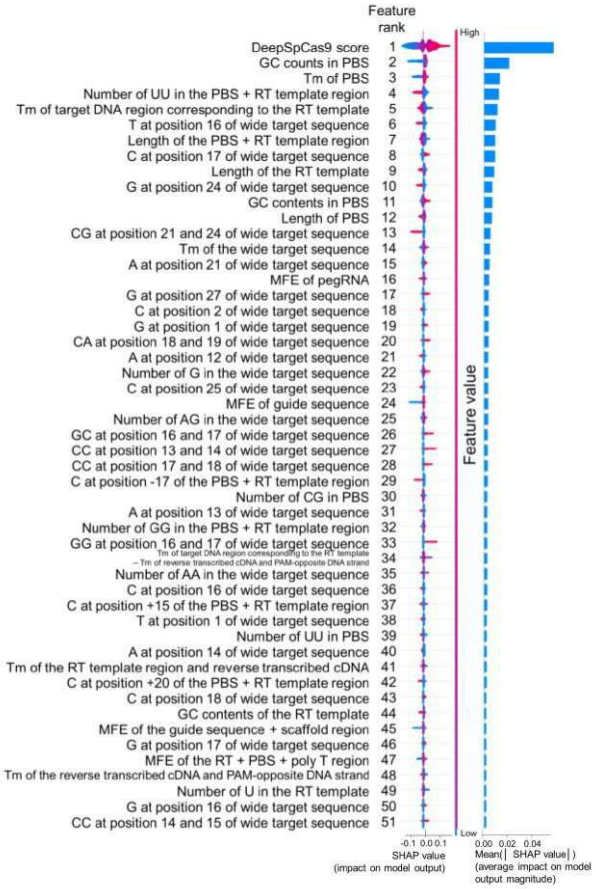
도면15



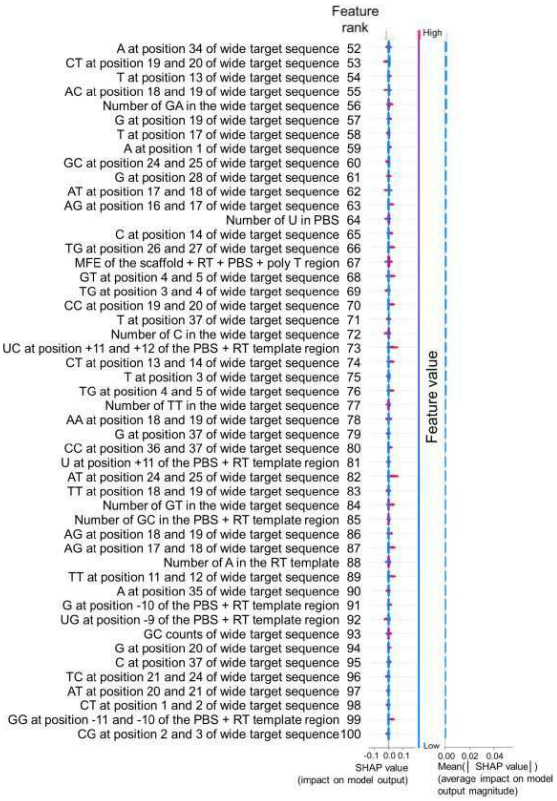
도면16



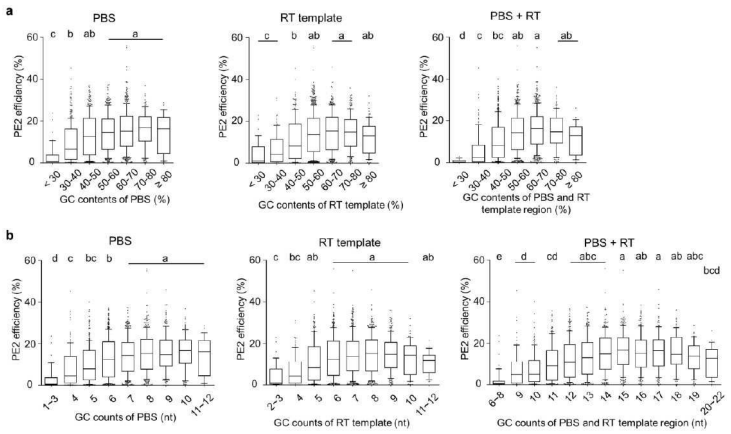
도면17


Feature rank
DeepSpCas9 score 1
GC counts in PBS 2
Tm of PBS 3
Number of UU in the PBS + RT template region 4
Tm of target DNA region corresponding to the RT template 5
T at position 16 of wide target sequence 6
Length of the PBS + RT template region 7
C at position 17 of wide target sequence 8
Length of the RT template 9
G at position 24 of wide target sequence 10
GC contents in PBS 11
Length of PBS 12
CG at position 21 and 24 of wide target sequence 13
Tm of the wide target sequence 14
A at position 21 of wide target sequence 15
MFE of pegRNA 16
G at position 27 of wide target sequence 17
C at position 2 of wide target sequence 18
G at position 1 of wide target sequence 19
CA at position 18 and 19 of wide target sequence 20
A at position 12 of wide target sequence 21
Number of G in the wide target sequence 22
C at position 25 of wide target sequence 23
MFE of guide sequence 24
Number of AG in the wide target sequence 25
GC at position 16 and 17 of wide target sequence 26
CC at position 13 and 14 of wide target sequence 27
CC at position 17 and 18 of wide target sequence 28
C at position -17 of the PBS + RT template region 29
Number of CG in PBS 30
A at position 13 of wide target sequence 31
Number of GG in the PBS + RT template region 32
GG at position 16 and 17 of wide target sequence 33
Tm of target DNA region corresponding to the RT template – Tm of reverse transcribed cDNA and PAM-opposite DNA strand 34
Number of AA in the wide target sequence 35
C at position 16 of wide target sequence 36
C at position +15 of the PBS + RT template region 37
T at position 1 of wide target sequence 38
Number of UU in PBS 39
A at position 14 of wide target sequence 40
Tm of the RT template region and reverse transcribed cDNA 41
C at position +20 of the PBS + RT template region 42
C at position 18 of wide target sequence 43
GC contents of the RT template 44
MFE of the guide sequence + scaffold region 45
G at position 17 of wide target sequence 46
MFE of the RT + PBS + poly T region 47
Tm of the reverse transcribed cDNA and PAM-opposite DNA strand 48
Number of U in the RT template 49
G at position 16 of wide target sequence 50
CC at position 14 and 15 of wide target sequence 51
High
Feature value
Low
-0.1 0.0 0.1
SHAP value
(impact on model output)
0.00 0.02 0.04
Mean(| SHAP value|)
(average impact on model output magnitude)

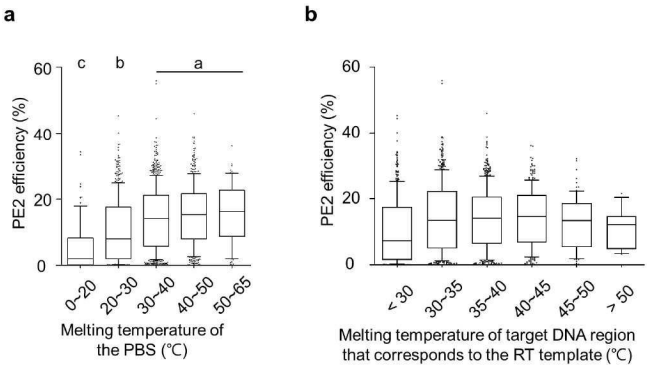
도면18


Feature rank
A at position 34 of wide target sequence 52
CT at position 19 and 20 of wide target sequence 53
T at position 13 of wide target sequence 54
AC at position 18 and 19 of wide target sequence 55
Number of GA in the wide target sequence 56
G at position 19 of wide target sequence 57
T at position 17 of wide target sequence 58
A at position 1 of wide target sequence 59
GC at position 24 and 25 of wide target sequence 60
G at position 28 of wide target sequence 61
AT at position 17 and 18 of wide target sequence 62
AG at position 16 and 17 of wide target sequence 63
Number of U in PBS 64
C at position 14 of wide target sequence 65
TG at position 26 and 27 of wide target sequence 66
MFE of the scaffold + RT + PBS + poly T region 67
GT at position 4 and 5 of wide target sequence 68
TG at position 3 and 4 of wide target sequence 69
CC at position 19 and 20 of wide target sequence 70
T at position 37 of wide target sequence 71
Number of C in the wide target sequence 72
UC at position +11 and +12 of the PBS + RT template region 73
CT at position 13 and 14 of wide target sequence 74
T at position 3 of wide target sequence 75
TG at position 4 and 5 of wide target sequence 76
Number of TT in the wide target sequence 77
AA at position 18 and 19 of wide target sequence 78
G at position 37 of wide target sequence 79
CC at position 36 and 37 of wide target sequence 80
U at position +11 of the PBS + RT template region 81
AT at position 24 and 25 of wide target sequence 82
TT at position 18 and 19 of wide target sequence 83
Number of GT in the wide target sequence 84
Number of GC in the PBS + RT template region 85
AG at position 18 and 19 of wide target sequence 86
AG at position 17 and 18 of wide target sequence 87
Number of A in the RT template 88
TT at position 11 and 12 of wide target sequence 89
A at position 35 of wide target sequence 90
G at position -10 of the PBS + RT template region 91
UG at position -9 of the PBS + RT template region 92
GC counts of wide target sequence 93
G at position 20 of wide target sequence 94
C at position 37 of wide target sequence 95
TC at position 21 and 24 of wide target sequence 96
AT at position 20 and 21 of wide target sequence 97
CT at position 1 and 2 of wide target sequence 98
GG at position -11 and -10 of the PBS + RT template region 99
CG at position 2 and 3 of wide target sequence 100
High
Low
Feature value
-0.1 0.0 0.1
SHAP value
(impact on model output)
0.00 0.02 0.04
Mean(| SHAP value|)
(average impact on model output magnitude)

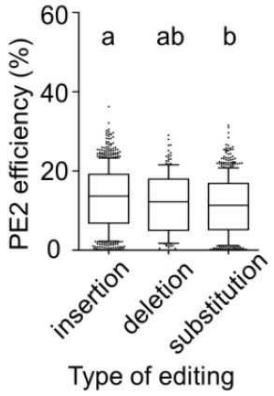
도면19


a
PBS
RT template
PBS + RT
PE2 efficiency (%)
GC contents of PBS (%)
GC contents of RT template (%)
GC contents of PBS and RT template region (%)
b
PBS
RT template
PBS + RT
PE2 efficiency (%)
GC counts of PBS (nt)
GC counts of RT template (nt)
GC counts of PBS and RT template region (nt)

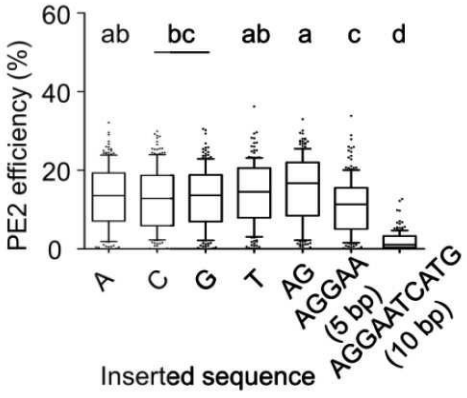
도면20



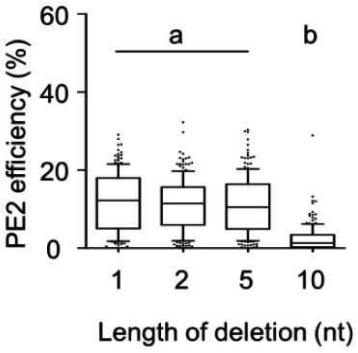
도면21



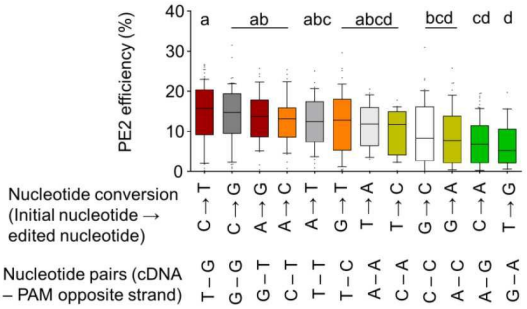
도면22



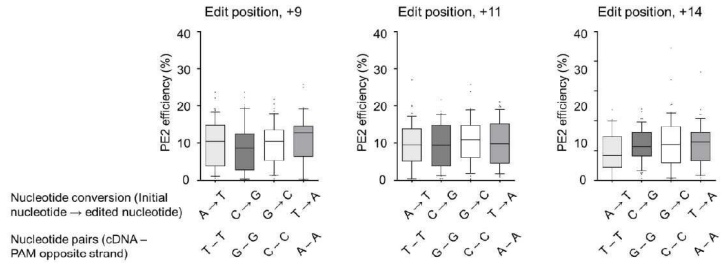
도면23



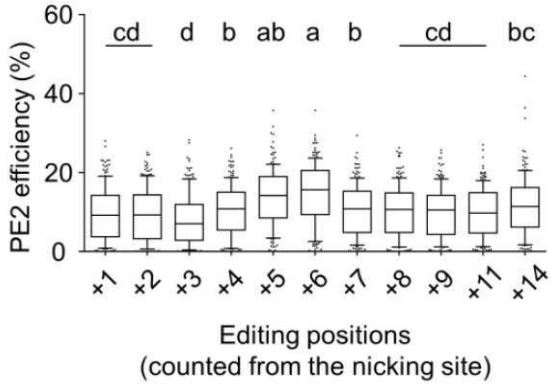
도면24



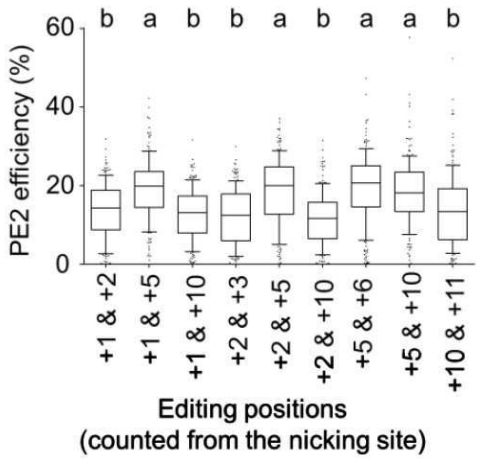
도면25



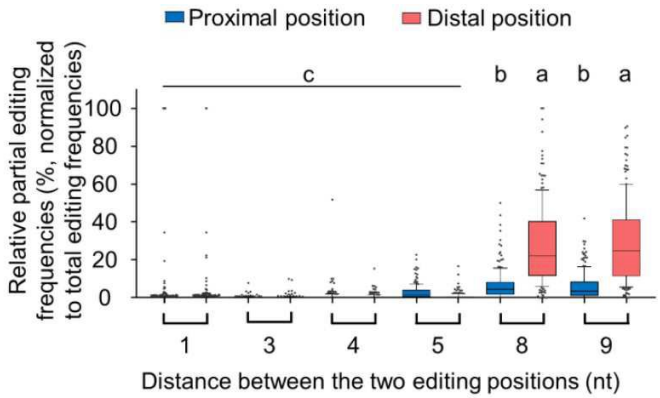
도면26



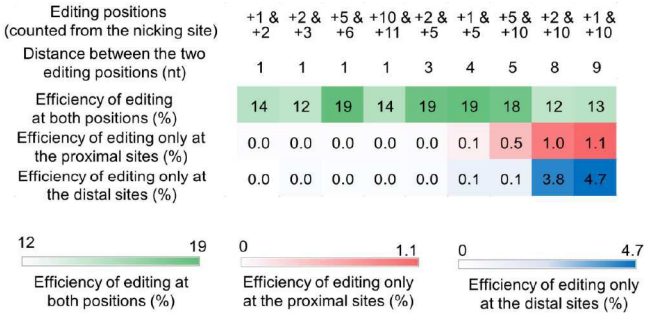
도면27



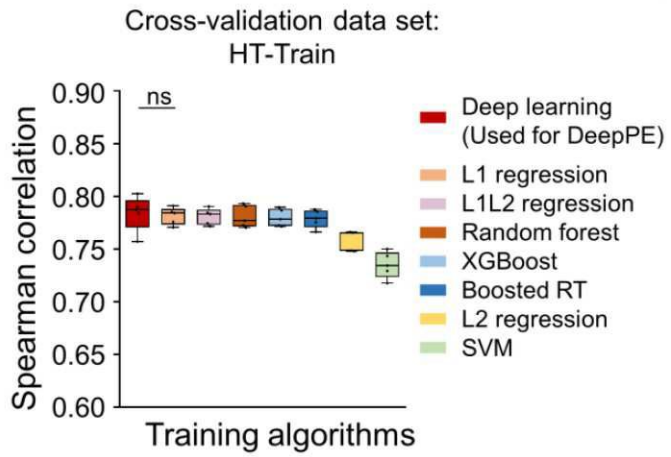
도면28



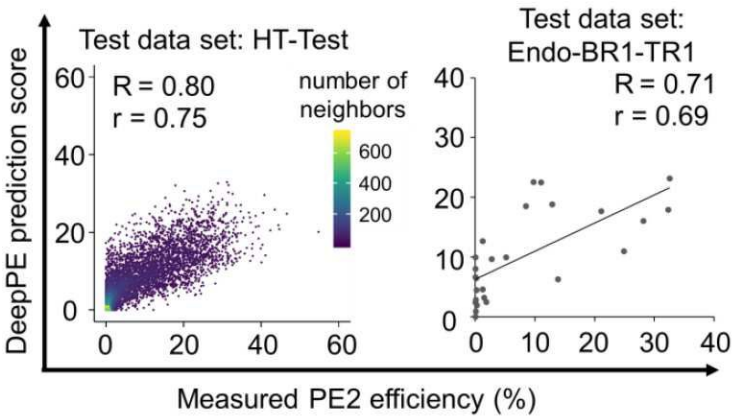
도면29



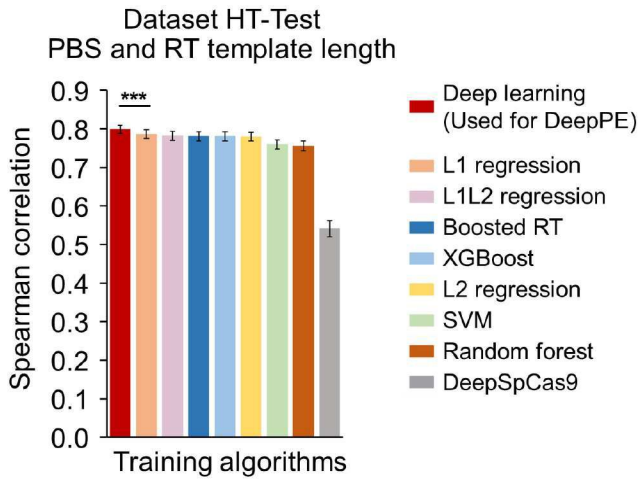
도면30



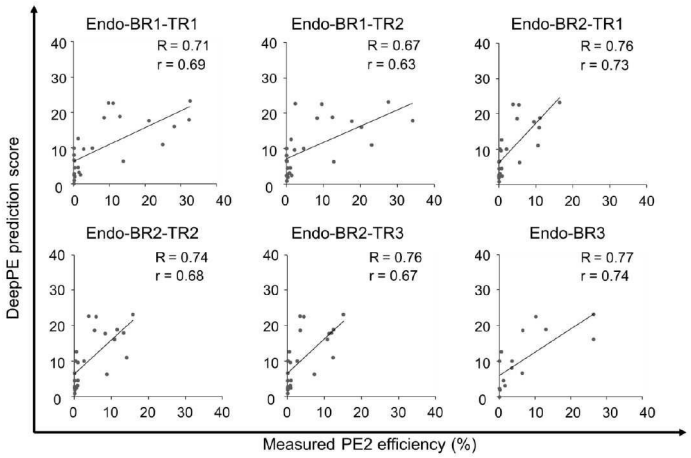
도면31



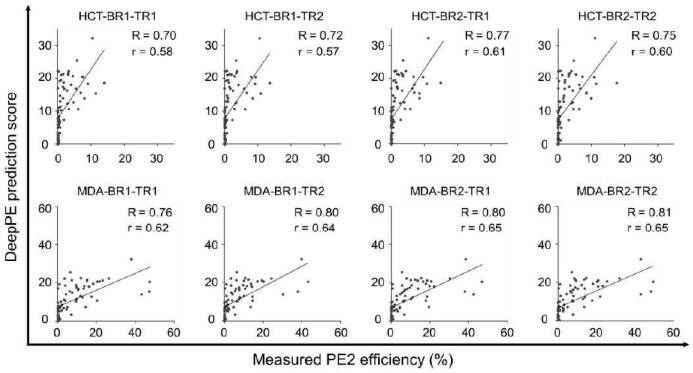
도면32



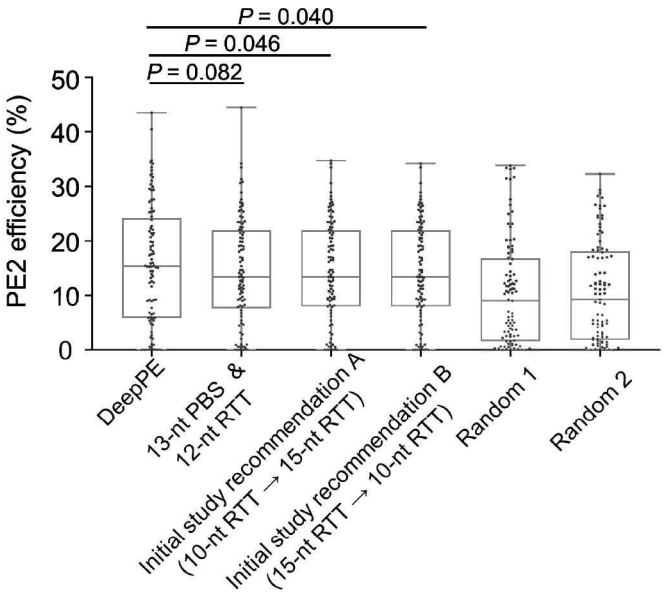
도면33



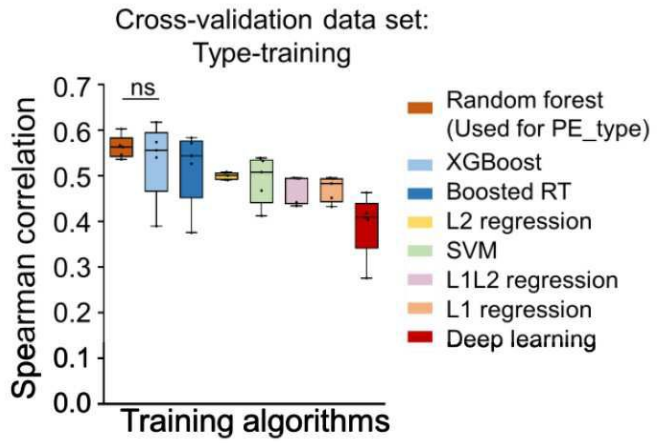
도면34



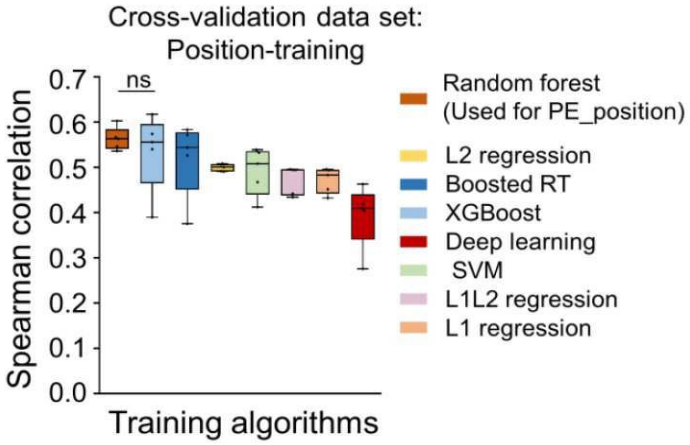
도면35



도면36


Cross-validation data set:
Type-training
ns
Spearman correlation
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Training algorithms
Random forest
(Used for PE_type)
XGBoost
Boosted RT
L2 regression
SVM
L1L2 regression
L1 regression
Deep learning

도면37


Cross-validation data set:
Position-training
ns
Spearman correlation
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Training algorithms
Random forest
(Used for PE_position)
L2 regression
Boosted RT
XGBoost
Deep learning
SVM
L1L2 regression
L1 regression