(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2020년11월16일
(11) 등록번호 10-2179034
(24) 등록일자 2020년11월10일

(51) 국제특허분류(Int. Cl.)
C12N 15/10 (2017.01) *C12N 15/113* (2010.01)
C12N 9/22 (2006.01) *C12N 9/78* (2006.01)
(52) CPC특허분류
C12N 15/1065 (2013.01)
C12N 15/113 (2013.01)
(21) 출원번호 10-2019-0078841
(22) 출원일자 2019년07월01일
심사청구일자 2019년07월01일
(65) 공개번호 10-2020-0002704
(43) 공개일자 2020년01월08일
(30) 우선권주장
1020180075668 2018년06월29일 대한민국(KR)
(56) 선행기술조사문헌
KR1020110061572 A*
NATURE BIOTECHNOLOGY. 2017, Vol.35, No.5, pp.438-439 1부.*
' '유전자 가위+DNA 바코드' 결합으로 암 성장 추적', 네이버 포스트, [online], 2018.04.05., 인터넷:<http://naver.me/FuyPZ6Ml>*
JP2017500038 A
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
연세대학교 산학협력단
서울특별시 서대문구 연세로 50 (신촌동, 연세대학교)
(72) 발명자
방두희
서울특별시 서대문구 연세로 50, 과학관 431비(신촌동, 연세대학교)
이욱재
서울특별시 서대문구 연세로 50, 과학관 431비(신촌동, 연세대학교)
황병진
서울특별시 서대문구 연세로 50, 과학관 431비(신촌동, 연세대학교)
(74) 대리인
특허법인다나

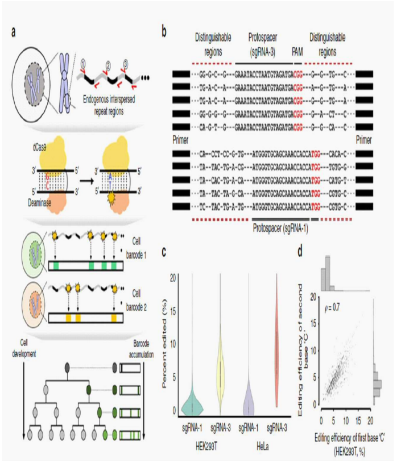
전체 청구항 수 : 총 4 항 심사관 : 한지혜

(54) 발명의 명칭 **세포 내재 반복 서열을 대상으로 탈아미노효소를 이용한 세포 바코딩 기술 개발**

(57) 요 약

본 발명은 세포 내재 반복 서열에 대한 nCas9-디아미나제 융합 단백질을 이용한 세포 바코딩 방법에 관한 것이다. 본 발명에 따른 표적화된 디아미나제는 상기 반복 서열을 표적으로 하여 다양한 유전자 바코드 패턴을 세포 내에 생성시킬 수 있어 세포 간의 관계 및 계통을 추적하는 것이 용이할 뿐만 아니라, 대상 세포의 범위가 넓다.

대 표 도 - 도2



(52) CPC특허분류
C12N 9/22 (2013.01)
C12N 9/78 (2013.01)
C12Y 305/04005 (2013.01)
C12N 2310/20 (2017.05)

이 발명을 지원한 국가연구개발사업
과제고유번호 2018M3A9H3024850
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 바이오의료기술개발사업
연구과제명 엔지니어드 박테리아 기반 정밀 면역 치료법 개발(1/3, 1단계)
기 여 율 1/3
과제수행기관명 연세대학교 산학협력단
연구기간 2018.04.01 ~ 2018.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 2015R1A2A1A10055972
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 중견연구자지원사업
연구과제명 종양 합성생물공학을 위한 차세대 플랫폼기술 개발(3/3)
기 여 율 1/3
과제수행기관명 연세대학교 산학협력단
연구기간 2017.11.01 ~ 2018.10.31
이 발명을 지원한 국가연구개발사업
과제고유번호 2016M3A9B6948494
부처명 과학기술정보통신부
과제관리(전문)기관명 한국연구재단
연구사업명 바이오의료기술개발사업
연구과제명 CRISPR 기반 유전자 진단기술 개발(2/3, 1단계)
기 여 율 1/3
과제수행기관명 연세대학교 산학협력단
연구기간 2017.08.29 ~ 2018.06.28

명 세 서

청구범위

청구항 1

삭제

청구항 2

삭제

청구항 3

삭제

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

삭제

청구항 8

삭제

청구항 9

세포 내재 반복 서열(endogenous repeat sequnce)을 타겟으로 하는 sgRNA, 및 nCas9(nickase Cas9) 및 시티딘 디아미나제(cytidine deaminase)를 발현하는 벡터를 제조하는 단계;

상기 제조된 벡터를 대상 세포 내로 형질감염(transfection)시켜 세포 내재 반복 서열 상에 유전자 바코드를 생성시키는 단계;

상기 세포를 배양하여 세포 분열 및 발달시키는 단계;

세포 분열 후 각 세대의 세포들에 대하여 서열 분석을 수행하여 유전적 바코드를 검출하는 단계; 및

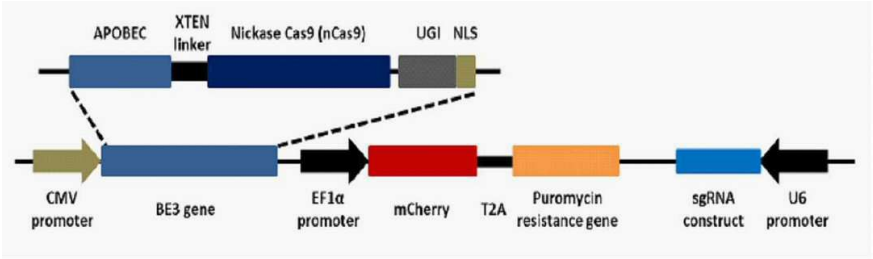
상기 세포들에서 검출된 유전자 바코드를 비교하여 정렬하는 단계;를 포함하고,

상기 유전자 바코드는 세포 분열 및 발달 단계에서 지속적으로 생성되어 누적되는 것을 특징으로 하는, 세포 내 유전자 바코드를 이용하여 세포 계통을 추적하는 방법.

청구항 10

제 9항에 있어서,

상기 벡터는 하기의 구조를 갖는 것을 특징으로 하는 방법:



청구항 11

제9항에 있어서,

상기 세포 내재 반복 서열은 레트로트랜스포존인 것을 특징으로 하는 방법.

청구항 12

제11항에 있어서,

상기 레트로트랜스포존은 LINE(Long interspersed nuclear element)인 것인 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 세포 내재 반복 서열에 대한 nCas9-디아미나제 융합 단백질을 이용한 세포 바코딩 방법에 관한 것이다.

배 경 기 술

[0002] 세포 유형간의 계통(lineage) 관계를 측정하는 것은 생명체의 발달과 질병에서 세포 분화의 기본 메카니즘을 이해하는데 중요하다.

[0003] 발달생물학의 핵심목표는 어떻게 단일세포가 여러 종류의 세포 유형으로 분화되어 성체로 변형되는지를 이해하는 것이다. 발생초기 및 세포의 지속적인 변형이 많은 성체시기에는 단일 세포 RNA 시퀀싱으로 전사체 유사성을 통해 일시적으로 궤도를 추적하여 세포 계통도를 유추 가능하다.

[0004] 그러나 이 접근법만으로는 성체 세포의 발달기원을 추적할 수 없어 계통 추적을 위한 단백질 마커에 형광물질을 부착하거나 바이러스를 이용하여 바코드를 붙이거나 트랜스포존 삽입 부위를 이용하는 방법 등과 같은 방법이 전통적으로 사용되어 왔다. 이러한 방법들은 단일 세포 전사체 분석이 어려우므로 세포 유형에 대한 정보를 제공하진 않는다.

[0005] 최근에는 상술한 바와 같은 단순하게 외인성 바코드를 도입하는 방법이 아닌 도입한 외인성 바코드에 지속적으로 직접 돌연변이를 유도하여 세포를 표지하는 세포 바코딩 전략이 사용되고 있다.

[0006] 현재 다양한 세포(유전자) 바코딩 전략이 개발되어 있으며, 특히, CRISPR/Cas9 시스템은 유전자 편집을 통한 세포 바코딩 기술로 세포 계통 추적에 이용되고 있다.

[0007] 다만, CRISPR/Cas9 시스템을 사용하는 경우, 반복 서열을 가진 외인성 바코드를 세포에 도입해야 하기 때문에 배열 요소의 반복된 복사본을 생성하고 세포에 도입하는 것이 어렵고 이와 같은 외부 DNA 서열을 추가로 유전자에 도입해야한다. 또한 타겟 DNA 이중나선을 모두 절단하기 때문에 세포에 치명적일 우려가 있다는 문제점이 있다.

[0008] 따라서, 다양한 종의 세포의 계통을 추적을 위한 보다 보편적이고 효율성이 좋은 새로운 세포 바코딩 전략이 필요한 실정이다.

선행기술문헌

비특허문헌

[0009] (비특허문헌 0001) Spanjaard, B. et al. Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. Nat. Biotechnol. 36, 469-473 (2018)

발명의 내용

해결하려는 과제

[0010] 본 발명의 목적은 상기의 기술적 과제를 해결하기 위한 것으로, 세포 반복 내재 서열을 대상으로 nCas9-디아미나제 융합 단백질를 이용하는 세포 바코딩 방법 및 상기 방법에 의해 생성된 세포 바코드를 이용한 세포 계통 추적 방법을 제공함에 있다.

과제의 해결 수단

[0011] 세포 계통을 추적하기 위해 내재적 유전자에 돌연변이를 유발시키거나 외인성 바코드를 도입하는 등의 방법으로 새겨진 세포 내 흔적(scar)을 이용하고 있다. 또한, 최근 연구된 세포 계통 추적 방법으로, CRISPR-Cas9 시스템을 이용하여 제브라피쉬 세포 내 RFP 트랜스 유전자에 흔적을 남겨 이를 추적하는 방법이 공개되었다. 그러나, 상기 방법을 RFP 트랜스 유전자가 도입된 제브라피쉬가 아닌 다른 종의 세포에 적용하는 데는 한계가 존재하며 또한. CRISPR-Cas9 시스템의 유전자 절단에 의해 돌연변이가 유발되어도 세포 발달에 영향을 미치지 않는 안전한 타겟 유전자의 발굴 및 CRISPR-Cas9 시스템의 안전성 문제가 해결되어야 한다.

[0012] 이에 본 발명자들은 상기 문제를 해결하기 위하여 모든 포유류 세포에 다량 분포하면서, 용이하게 다양한 패턴의 유전자 바코드를 도입할 수 있도록 세포 내재 반복 서열을 표적으로 하여 바코딩할 수 있는 변형 CRISPR-Cas9 시스템을 고안하였고, 상기 시스템을 이용하는 경우, 세포에 발달 과정에서 세포에 누적된 유전자 바코드를 형성시킴을 확인하였고, 세대마다 세포 간 누적된 유전자 바코드를 비교함으로써 세포 계통 추적이 가능함을 확인함으로써 본 발명을 완성하였다.

[0013] 이에 본 발명은 세포 내재 반복 서열(endogenous repeat sequence)을 타겟으로 하는 sgRNA, 및 nCas9(nickase Cas9) 및 시티딘 디아미나제(cytidine deaminase)로 이루어진 융합 단백질을 포함하는 세포 내 유전적 바코드 도입을 위한 조성물을 제공한다.

[0014] 본 발명에서 사용된 용어 “내재 반복 서열(endogenous repeat sequence)”은 유전체에 존재하는 반복된 서열을 의미하며, 포유류 세포들은 다양한 내재 반복 서열을 갖는데, 대표적으로 트랜스포존(transposon), 레트로트랜스포존(retrotransposon), 마이크로세틀레이드(microsatellite), 미니세틀레이트(minisatellite), 텔로미어(telomere) 등이 있으며, 이들은 서열의 반복 수 및 반복된 형태에 따라 분류된다.

[0015] 본 발명의 조성물이 타겟으로 하는 세포 내재 반복 서열은 레트로트랜스포존일 수 있으며, 보다 구체적으로는 LINE(Long interspersed nuclear element)일 수 있다. 상기 LINE은 많은 진핵세포의 게놈에 널리 산재하고 있는 비-LTR(long terminal repeat) 레트로트랜스포존의 일종으로, 인간 게놈에는 약 21%가 존재하고, 약 7000개의 긴 염기쌍을 갖는다. 본 발명의 일 구체예에서는, 상기 LINE은 포유류 세포 내에 다수 분포하고 있으며, 반복 서열이 상당히 길기 때문에 이를 대상으로 하는 경우, 다양한 조합의 바코드를 타겟 영역에 형성시킬 수 있을 것으로 가정하여 이를 타겟 유전자 영역으로 선택하였다.

[0016] 본 발명의 상기 “sgRNA(single guide RNA)”는 표적 유전자에 특이적으로 결합하는 서열을 포함하는 다일 가닥 가이드 RNA로서, 상기 sgRNA는 표적 유전자에 결함으로써 Cas9 단백질을 해당 유전자로 유도하여 복합체를 형성할 수 있다.

[0017] 상기 sgRNA와 복합체를 형성하는 본 발명의 Cas9 단백질은 시티딘 디아미나제와 융합된 것으로, 상기 Cas9/시티딘 디아미나제 융합 단백질일 수 있다. 상기 융합 단백질은 기존 CRISPR-Cas9 시스템은 이중 나선을 모두 절단하는 방식이므로, 세포의 발달 및 분화에 영향을 줄 수 있는 단점을 보완하기 위하여 고안된 것이며, 타겟 유전자 부위의 C(cytosine)를 T(thymine)로 치환시킬 수 있다. 따라서 본 발명의 Cas9/시티딘 디아미나제 융합 단백질은 타겟 유전자 부위에 단일 염기쌍 변이를 형성시킨다.

[0018] 본 발명의 세포 내재 반복 서열(endogenous repeat sequence)을 타겟으로 하는 sgRNA, 및 nCas9(nickase Cas9)

및 시티딘 디아미나제(cytidine deaminase)로 이루어진 융합 단백질은 세포 내로 형질감염된 플라스미드 벡터에 의해 발현될 수 있다.

[0019] 상기 세포 내재 반복 서열(endogenous repeat sequence)을 타겟으로 하는 sgRNA, 및 nCas9(nickase Cas9) 및 시티딘 디아미나제(cytidine deaminase)로 이루어진 융합 단백질를 발현하는 플라스미드 벡터는 도 1로 나타낼 수 있고, 보다 구체적으로 상기 벡터는 서열번호 1로 이루어진 서열로 표시될 수 있다. 서열번호 1에서 "NNNNN..."와 같이 N으로 반복된 서열은 sgRNA를 의미한다.

[0020] 본 발명의 일 구체예에서, 하기 서열번호 2 및/또는 서열번호 3로 표시되는 염기서열로 이루어진 상기 세포 내재 반복 서열인 LINE-1을 타겟으로 하는 sgRNA을 사용하였으나, 이는 예시일 뿐 본 발명이 이에 제한되는 것은 아니다.

[0021] sgRNA-1(서열번호 2): 5'-ATGGGTGCAGCAAACCACCA-3'

[0022] sgRNA-3(서열번호 3): 5'-GAAATACCTAATGTAGATGA-3'

[0023] 본 발명의 일 구체예에서, 상기 플라스미드 벡터는 이하 제조예 1과 같은 방법을 이용하여 제조될 수 있다.

[0024] 본 명세서에서 상기 제조된 플라스미드 벡터(PB CMV-BE3 EF1α-mCherry-T2A-puro sgRNA)에 의해 발현되는 세포 내재 반복 서열(endogenous repeat sequnece)을 타겟으로 하는 sgRNA, 및 nCas9(nickase Cas9) 및 시티딘 디아미나제(cytidine deaminase)로 이루어진 융합 단백질은 "Cas9/시티딘 디아미나제", "표적화된 탈아미노효소" 또는 "표적화된 디아미나제"와 교환적으로 사용될 수 있다. 그리고 상기 용어들은 "표적화된 디아미나제 시스템"과 같이 "시스템"과 같이 병용될 수 있다.

[0025] 본 발명에서 사용된 용어 "유전자 바코드"는 개별 세포를 식별할 수 있는 세포 고유 또는 인위적으로 표지된 표시를 의미한다. 본 발명에서 상기 유전자 바코드는 상기 표적화된 디아미나제에 의해 세포 내재 반복 서열상에 형성된 단일 염기 변이(C > T)를 의미할 수 있다. 또한, 상기 표적화된 디아미나제에 의해 단일 염기 변이가 복수의 사이트에서 생성된 경우에는 이러한 단일 염기 변이 패턴도 유전자 바코드가 될 수 있다. 본 명세서에서의 유전자 바코드는 "세포 바코드"와 교환적으로 사용될 수 있다.

[0026] 또한, 본 발명은 상술한 바와 같은 표적화된 디아미나제를 발현하는 플라스미드 벡터를 제조하는 단계 및 이를 세포 내로 형질감염시키는 단계를 포함하는 세포 내 유전자 바코드를 생성시키는 방법을 제공한다.

[0027] 본 발명의 표적화된 디아미나제는 세포 내재 반복 서열을 타겟으로 하는 sgRNA에 의해 내재 반복 서열상으로 유도되고, 상기 내재 반복 서열상에 존재하는 C 들을 T로 치환시킴으로써 유전자 바코드를 생성시킬 수 있다.

[0028] 표적화된 디아미나제에 대한 설명은 본 발명에 따른 표적화된 디아미나제에 대해 위에서 기술한 내용이 그대로 적용 또는 준용될 수 있다.

[0029] 본 발명자들은 본 발명에 따라 상기 표적화된 디아미나제를 이용하여 유전자 바코드를 생성시킴으로써 세포를 식별할 수 있는 기술이 어느 하나의 조상(배아) 세포로부터 다양한 자손 세포로 분열/발달하는 경우, 상기 자손 세포들 간 또는 자손 세포들 및 그 상위 세대 세포들 간의 상관관계(즉, 계통)를 규명할 수 있을 것으로 예상하였다.

[0030] 본 발명의 구체예에서, 본 발명자들은 포유류 세포(예를 들면, HEK293T 인간 배아 신장 및 HeLa 인간 자궁 경부암 세포)를 분열/발달 시키면서 LINE을 대상으로 하는 상기 표적화된 디아미나제를 이용하는 경우, 세대를 거칠 때마다 유전자 바코드가 누적되는 것을 확인하였고, 이러한 누적된 유전자 바코드 패턴을 세대 간 또는 자손 세포 간 비교/분석을 통해서 계통을 추적할 수 있음을 확인하였다.

[0031] 이에 본 발명은 상기 표적화된 디아미나제를 이용하여 유전자 바코드를 세포 내재 반복 서열 상에 생성시키는 단계;

[0032] 상기 세포를 배양하여 세포 분열 및 발달시키는 단계;

[0033] 세포 분열 후 각 세대의 세포들에 대하여 서열 분석을 수행하여 유전적 바코드를 검출하는 단계; 및

[0034] 상기 세포들에서 검출된 유전자 바코드를 비교하여 정렬하는 단계를 포함하는 세포 내 유전자 바코드를 이용하여 세포 계통을 추적하는 방법을 제공한다.

[0035] 표적화된 디아미나제에 대한 설명은 본 발명에 따른 표적화된 디아미나제에 대해 위에서 기술한 내용이 그대로

적용 또는 준용될 수 있다.

[0036] 본 발명의 따른 세포 계통 추적 방법은 세포 내재 반복 서열인 LINE을 표적으로 하는 표적화된 디아미나제를 이용하기 때문에, 유전자 바코드가 형성될 수 있는 부위가 다양화되어 이러한 바코드가 누적되는 경우 다양한 패턴의 유전자 바코드를 형성시킬 수 있으므로, 세포 계통을 추적하는데 효율적이다.

[0037] 본 발명의 일 실시예에 따르면, 타겟 부위 및 편집 속도의 증가는 세포 계통 트리의 정확성을 향상시키므로, 상기 개시된 sgRNA 이외에 다른 부위를 타겟으로 하는 sgRNA와의 조합을 통해서 계통 추적의 정확성을 향상시킬 수 있다.

[0038] 또한, 상기 세포 내재 반복 서열은 포유류 세포에 많이 산재되어 있는 유전자 부위에 해당하므로, 포유류, 특히 인간 세포에 대한 세포 계통 추적이 가능하다.

발명의 효과

[0039] 본 발명에 따른 표적화된 디아미나제는 단일 염기쌍을 치환하는 방식으로 유전자 편집이 가능하기 때문에 기존 유전자 가위가 DNA 이중 나선을 모두 절단하는 방식보다 유전자 발달에 미치는 영향이 상대적으로 적습니다. 또한, 상기 세포 내재 반복 서열들은 특히 포유류 세포 유전자의 많은 부분을 차지하고 있기 때문에, 상기 반복 서열을 표적으로 하는 본 발명의 유전자 바코딩 방법을 이용하여 다양한 유전자 바코드 패턴을 세포 내에 생성시킬 수 있으므로, 세포 간의 관계 및 계통을 추적하는 것을 용이하며, 적용 가능한 세포의 범위를 확대시킬 수 있다.

도면의 간단한 설명

[0040] 도 1은 표적화된 디아미나제를 발현하는 벡터의 구조를 나타낸 것이다.

도 2a는 표적화된 디아미나제의 유전자 바코딩 원리의 개략도를 나타낸 것이다.

도 2b는 sgRNA-1 및 2에 의해 인식되는 대표적인 타겟 부위의 서열을 정렬하여 나타낸 것이다. 표적화된 디아미나제는 주로 상기 sgRNA에 의해 식별되는 프로토스페이서 옆의 PAM 서열로부터 13~17 뉴클레오타이드 떨어진 범위의 C를 T로 치환하게 된다.

도 2c는 HEK293T 및 HeLa 세포에서의 sgRNA-1 및 3의 유전자 편집 효율을 나타낸 것이다.

도 2d는 sgRNA-3의 프로토스페이서 서열 내 PAM 서열로부터 먼 순으로 제 1 및 제 2 염기 C에 대한 편집 효율을 비교한 결과를 나타낸 것이다.

도 3a는 지정된 시점에서의 sgRNA-1 및 3을 사용하여 누적 유전자 편집 효율을 측정한 결과를 나타낸다.

도 3b는 HEK293T 세포에서 sgRNA-3를 이용한 트리 실험의 모식도를 나타낸 것이다.

도 3c는 단일 세포 확장 실험의 저속 이미징 결과를 나타낸 것으로, 마지막 영상상의 단일 세포들을 계통도 구축을 위해 선별, 시퀀싱하였다.

도 3d는 도 3c의 7개의 단일 세포에 대한 트리 확장 실험예를 나타낸다.

도 4는 벌크 세포에서의 시험관 트리 확장 실험의 모식도이다.

도 5는 HeLa 세포에서의 sgRNA-3를 이용한 트리 확장 실험의 모식도를 나타낸 것이다. 빨간 화살표는 잘못 연결된 노드들 나타내고, 빨간선은 잘못 연결된 모-녀 노드 연결을 나타내며, 점선은 수정된 모-녀 노드 연결을 나타낸다.

도 6은 단일 세포에 대한 PCR 및 WGA 결과 간의 서열 커버리지 통계 결과를 나타낸 것이다.

도 7은 단일 세포 저속 이미징을 통한 추가 트리 확장 실험예를 나타낸 것이다.

도 8은 HEK293T 세포에서 상이한 시점에서 측정된 sgRNA-3의 편집역학 결과를 나타낸 것이다. 편집 효율을 측정하기 위해 두 개의 복제물의 평균값을 사용하여 곡선으로 나타내었다.

도 9는 다른 파라미터로 sgRNA-3를 이용한 트리 재구축 시뮬레이션 결과를 나타내는 것으로, Cophenetic Correlation을 이용한 가상(in silico) 실험에 의해 평가되었다.

발명을 실시하기 위한 구체적인 내용

[0041] 이하, 본 발명을 실시예를 통해 상세히 설명한다. 다만, 하기 실시예는 오로지 본 발명을 보다 구체적으로 설명하기 위한 것이고, 본 발명의 요지에 따라 본 발명의 범위가 이들 실시예에 의해 제한되지 않는다는 것은 당업계에서 통상의 지식을 가진 자에 있어서 자명할 것이다.

[0042] **[제조예 1] L1 요소를 표적으로 하는 유전자 바코딩를 위한 플라스미드 벡터 구축**

[0043] PB CMV-BE3 EF1α-mCherry-T2A-puro sgRNA는 gRNA_클로닝 벡터(Addgene plasmid #41824, Addgene, USA)의 U6-sgRNA 발현 카세트, pCMV-BE3 (Addgene plasmid #73021)의 CMV-BE3, lentiGuide-Puro (Addgene plasmid #52963)의 푸로마이신 유전자(puromycin gene) 및 PB_tet_attB-mCherry (Seung Hyeok Seok in Seoul National University College of Medicine)의 mCherry를 PB-CA (Addgene plasmid #20960)의 PiggyBac 트랜스포존 백본(transposon backbone) 상에 PCR 조립을 이용하여 구축하였다. 표적 sgRNA 서열은 제공된 hCRISPR gRNA 합성 프로토콜에 따라 클론되었다. pCy43 PiggyBac 트랜스포존 벡터는 Sanger Institute (Hinxton, UK)로부터 제공받았다. 모든 복제된 플라스미드를 생거 시퀀싱(Sanger sequencing)을 이용하여 확인하였다. 플라스미드들을 제조자의 프로토콜에 따라 EndoFree Plasmid Maxi Kit (QIAGEN, USA) 및 Exprep Plasmid SV kit (GeneAll, Korea)를 사용하여 제조하였다.

[0044] 제조된 플라스미드 벡터(PB CMV-BE3 EF1α-mCherry-T2A-puro sgRNA)는 도 1의 구조를 갖고 서열은 서열번호 1과 같다:

[0045] 서열번호 1:

[0046] ACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCA
TTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAA
ATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTT
TTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCA
ACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTGGTTTAGTGAA
CCGTCAGATCCGCTAGAGATCCGCGGCCGCTAATACGACTCACTATAGGGAGAGCCGCCACCATGAGCTCAGAGACTGGCCCAGTGGCTGTGGACCCCACAT
TGAGACGGCGGATCGAGCCCCATGAGTTTGAGGTATTCTTCGATCCGAGAGAGCTCCGCAAGGAGACCTGCCTGCTTTACGAAATTAATTGGGGGGGCCGGC
ACTCCATTTGGCGACATACATCACAGAACACTAACAAGCACGTCGAAGTCAACTTCATCGAGAAGTTCACGACAGAAAGATATTTCTGTCCGAACACAAGGT
GCAGCATTACCTGGTTTCTCAGCTGGAGCCCATGCGGCGAATGTAGTAGGGCCATCACTGAATTCCTGTCAAGGTATCCCCACGTCACTCTGTTTATTTACA
TCGCAAGGCTGTACCACCACGCTGACCCCCGCAATCGACAAGGCCTGCGGGATTTGATCTCTTCAGGTGTGACTATCCAAATTATGACTGAGCAGGAGTCAG
GATACTGCTGGAGAAACTTTGTGAATTATAGCCCGAGTAATGAAGCCCACTGGCCTAGGTATCCCCATCTGTGGGTACGACTGTACGTTCTTGAACTGTACT
GCATCATACTGGGCCTGCCTCCTTGTCTCAACATTCTGAGAAGGAAGCAGCCACAGCTGACATTCTTTACCATCGCTCTTCAGTCTTGTCATTACCAGCGAC
TGCCCCCACACATTCTCTGGGCCACCGGGTTGAAAAGCGGCAGCGAGACTCCCGGGACCTCAGAGTCCGCCACACCCGAAAGTGATAAAAAGTATTCTATTG
GTTTAGCCATCGGCACTAATTCCGTTGGATGGGCTGTCATAACCGATGAATACAAAGTACCTTCAAAGAAATTTAAGGTGTTGGGGAACACAGACCGTCATT
CGATTAAAAAGAATCTTATCGGTGCCCTCCTATTCGATAGTGGCGAAACGGCAGAGGCGACTCGCCTGAAACGAACCGCTCGGAGAAGGTATACACGTCGCA
AGAACCGAATATGTTACTTACAAGAAATTTTTAGCAATGAGATGGCCAAAGTTGACGATTCTTTCTTTCACCGTTTGGAAGAGTCCTTCCTTGTCGAAGAGG
ACAAGAAACATGAACGGCACCCCATCTTTGGAAACATAGTAGATGAGGTGGCATATCATGAAAAGTACCCAACGATTTATCACCTCAGAAAAAAGCTAGTTG
ACTCAACTGATAAAGCGGACCTGAGGTTAATCTACTTGGCTCTTGCCCATATGATAAAGTTCCGTGGGCACTTTCTCATTGAGGGTGATCTAAATCCGGACA
ACTCGGATGTCGACAAACTGTTCATCCAGTTAGTACAAACCTATAATCAGTTGTTTGAAGAGAACCCTATAAATGCAAGTGGCGTGGATGCGAAGGCTATTC
TTAGCGCCCGCCTCTCTAAATCCCGACGGCTAGAAAACCTGATCGCACAATTACCCGGAGAGAAGAAAAATGGGTTGTTCGGTAACCTTATAGCGCTCTCAC
TAGGCCTGACACCAAATTTTAAGTCGAACTTCGACTTAGCTGAAGATGCCAAATTGCAGCTTAGTAAGGACACGTACGATGACGATCTCGACAATCTACTGG
CACAAATTGGAGATCAGTATGCGGACTTATTTTTGGCTGCCAAAAACCTTAGCGATGCAATCCTCCTATCTGACATACTGAGAGTTAATACTGAGATTACCA
AGGCGCCGTTATCCGCTTCAATGATCAAAAGGTACGATGAACATCACCAAGACTTGACACTTCTCAAGGCCCTAGTCCGTCAGCAACTGCCTGAGAAATATA
AGGAAATATTCTTTGATCAGTCGAAAAACGGGTACGCAGGTTATATTGACGGCGGAGCGAGTCAAGAGGAATTCTACAAGTTTATCAAACCCATATTAGAGA
AGATGGATGGGACGGAAGAGTTGCTTGTAAAACTCAATCGCGAAGATCTACTGCGAAAGCAGCGGACTTTCGACAACGGTAGCATTCCACATCAAATCCACT
TAGGCGAATTGCATGCTATACTTAGAAGGCAGGAGGATTTTTATCCGTTCCTCAAAGACAATCGTGAAAAGATTGAGAAAATCCTAACCTTTCGCATACCTT
ACTATGTGGGACCCCTGGCCCGAGGGAACTCTCGGTTCGCATGGATGACAAGAAAGTCCGAAGAAACGATTACTCCATGGAATTTTGAGGAAGTTGTCGATA
AAGGTGCGTCAGCTCAATCGTTCATCGAGAGGATGACCAACTTTGACAAGAATTTACCGAACGAAAAAGTATTGCCTAAGCACAGTTTACTTTACGAGTATT
TCACAGTGTACAATGAACTCACGAAAGTTAAGTATGTCACTGAGGGCATGCGTAAACCCGCCTTTCTAAGCGGAGAACAGAAGAAAGCAATAGTAGATCTGT
TATTCAAGACCAACCGCAAAGTGACAGTTAAGCAATTGAAAGAGGACTACTTTAAGAAAATTGAATGCTTCGATTCTGTCGAGATCTCCGGGGTAGAAGATC
GATTTAATGCGTCACTTGGTACGTATCATGACCTCCTAAAGATAATTAAAGATAAGGACTTCCTGGATAACGAAGAGAATGAAGATATCTTAGAAGATATAG

TGTTGACTCTTACCCTCTTTGAAGATCGGGAAATGATTGAGGAAAGACTAAAAACATACGCTCACCTGTTCGACGATAAGGTTATGAAACAGTTAAAGAGGC
GTCGCTATACGGGCTGGGGACGATTGTCGCGGAAACTTATCAACGGGATAAGAGACAAGCAAAGTGGTAAAACTATTCTCGATTTTCTAAAGAGCGACGGCT
TCGCCAATAGGAACTTTATGCAGCTGATCCATGATGACTCTTTAACCTTCAAAGAGGATATACAAAAGGCACAGGTTTCCGGACAAGGGGACTCATTGCACG
AACATATTGCGAATCTTGCTGGTTCGCCAGCCATCAAAAAGGGCATACTCCAGACAGTCAAAGTAGTGGATGAGCTAGTTAAGGTCATGGGACGTCACAAAC
CGGAAAACATTGTAATCGAGATGGCACGCGAAAATCAAACGACTCAGAAGGGGCAAAAAAACAGTCGAGAGCGGATGAAGAGAATAGAAGAGGGTATTAAAG
AACTGGGCAGCCAGATCTTAAAGGAGCATCCTGTGGAAAATACCCAATTGCAGAACGAGAAACTTTACCTCTATTACCTACAAAATGGAAGGGACATGTATG
TTGATCAGGAACTGGACATAAACCGTTTATCTGATTACGACGTCGATCACATTGTACCCCAATCCTTTTTGAAGGACGATTCAATCGACAATAAAGTGCTTA
CACGCTCGGATAAGAACCGAGGGAAAAGTGACAATGTTCCAAGCGAGGAAGTCGTAAAGAAAATGAAGAACTATTGGCGGCAGCTCCTAAATGCGAAACTGA
TAACGCAAAGAAAGTTCGATAACTTAACTAAAGCTGAGAGGGGTGGCTTGTCTGAACTTGACAAGGCCGGATTTATTAAACGTCAGCTCGTGGAAACCCGCC
AAATCACAAAGCATGTTGCACAGATACTAGATTCCCGAATGAATACGAAATACGACGAGAACGATAAGCTGATTCGGGAAGTCAAAGTAATCACTTTAAAGT
CAAAATTGGTGTCGGACTTCAGAAAGGATTTTCAATTCTATAAAGTTAGGGAGATAAATAACTACCACCATGCGCACGACGCTTATCTTAATGCCGTCGTAG
GGACCGCACTCATTAAGAAATACCCGAAGCTAGAAAGTGAGTTTGTGTATGGTGATTACAAAGTTTATGACGTCCGTAAGATGATCGCGAAAAGCGAACAGG
AGATAGGCAAGGCTACAGCCAAATACTTCTTTTATTCTAACATTATGAATTTCTTTAAGACGGAAATCACTCTGGCAAACGGAGAGATACGCAAACGACCTT
TAATTGAAACCAATGGGGAGACAGGTGAAATCGTATGGGATAAGGGCCGGGACTTCGCGACGGTGAGAAAAGTTTTGTCCATGCCCCAAGTCAACATAGTAA
AGAAAACTGAGGTGCAGACCGGAGGGTTTTCAAAGGAATCGATTCTTCCAAAAAGGAATAGTGATAAGCTCATCGCTCGTAAAAAGGACTGGGACCCGAAAA
AGTACGGTGGCTTCGATAGCCCTACAGTTGCCTATTCTGTCCTAGTAGTGGCAAAAGTTGAGAAGGGAAAATCCAAGAAACTGAAGTCAGTCAAAGAATTAT
TGGGGATAACGATTATGGAGCGCTCGTCTTTTGAAAAGAACCCCATCGACTTCCTTGAGGCGAAAGGTTACAAGGAAGTAAAAAAGGATCTCATAATTAAAC
TACCAAAGTATAGTCTGTTTGAGTTAGAAAATGGCCGAAAACGGATGTTGGCTAGCGCCGGAGAGCTTCAAAAGGGGAACGAACTCGCACTACCGTCTAAAT
ACGTGAATTTCCTGTATTTAGCGTCCCATTACGAGAAGTTGAAAGGTTCACCTGAAGATAACGAACAGAAGCAACTTTTTGTTGAGCAGCACAAACATTATC
TCGACGAAATCATAGAGCAAATTTCGGAATTCAGTAAGAGAGTCATCCTAGCTGATGCCAATCTGGACAAAGTATTAAGCGCATACAACAAGCACAGGGATA
AACCCATACGTGAGCAGGCGGAAAATATTATCCATTTGTTTACTCTTACCAACCTCGGCGCTCCAGCCGCATTCAAGTATTTTGACACAACGATAGATCGCA
AACGATACACTTCTACCAAGGAGGTGCTAGACGCGACACTGATTCACCAATCCATCACGGGATTATATGAAACTCGGATAGATTTGTCACAGCTTGGGGGTG
ACTCTGGTGGTTCTACTAATCTGTCAGATATTATTGAAAAGGAGACCGGTAAGCAACTGGTTATCCAGGAATCCATCCTCATGCTCCCAGAGGAGGTGGAAG
AAGTCATTGGGAACAAGCCGGAAAGCGATATACTCGTGCACACCGCCTACGACGAGAGCACCGACGAGAATGTCATGCTTCTGACTAGCGACGCCCCTGAAT
ACAAGCCTTGGGCTCTGGTCATACAGGATAGCAACGGTGAGAACAAGATTAAGATGCTCTCTGGTGGTTCTCCCAAGAAGAAGAGGAAAGTCTAACAGCAGA
GATCCAGTTTATCGATGAGTAATTCATACAAAAGGACTCGCCCCTGCCTTGGGGAATCCCAGGGACCGTCGTTAAACTCCCACTAACGTAGAACCCAGAGAT
CGCTGCGTTCCCGCCCCCTCACCCGCCCGCTCTCGTCATCACTGAGGTGGAGAAGAGCATGCGTGAGGCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCG
CCCACAGTCCCCGAGAAGTTGGGGGGAGGGGTCGGCAATTGAACCGGTGCCTAGAGAAGGTGGCGCGGGGTAAACTGGGAAAGTGATGTCGTGTACTGGCTC
CGCCTTTTTCCCGAGGGTGGGGGAGAACCGTATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTCGCAACGGGTTTGCCGCCAGAACACAGGTAAGTGCC
GTGTGTGGTTCCCGCGGGCCTGGCCTCTTTACGGGTTATGGCCCTTGCGTGCCTTGAATTACTTCCACGCCCCTGGCTGCAGTACGTGATTCTTGATCCCGA
GCTTCGGGTTGGAAGTGGGTGGGAGAGTTCGAGGCCTTGCGCTTAAGGAGCCCCTTCGCCTCGTGCTTGAGTTGAGGCCTGGCTTGGGCGCTGGGGCCGCCG
CGTGCGAATCTGGTGGCACCTTCGCGCCTGTCTCGCTGCTTTCGATAAGTCTCTAGCCATTTAAAATTTTTGATGACCTGCTGCGACGCTTTTTTTCTGGCA
AGATAGTCTTGTAAATGCGGGCCAAGATCTGCACACTGGTATTTCGGTTTTTGGGGCCGCGGGCGGCGACGGGGCCCGTGCGTCCCAGCGCACATGTTCGGC
GAGGCGGGGCCTGCGAGCGCGGCCACCGAGAATCGGACGGGGGTAGTCTCAAGCTGGCCGGCCTGCTCTGGTGCCTGGCCTCGCGCCGCCGTGTATCGCCCC
GCCCTGGGCGGCAAGGCTGGCCCGGTCGGCACCAGTTGCGTGAGCGGAAAGATGGCCGCTTCCCGGCCCTGCTGCAGGGAGCTCAAAATGGAGGACGCGGCG
CTCGGGAGAGCGGGCGGGTGAGTCACCCACACAAAGGAAAAGGGCCTTTCCGTCCTCAGCCGTCGCTTCATGTGACTCCACGGAGTACCGGGCGCCGTCCAG
GCACCTCGATTAGTTCTCGAGCTTTTGGAGTACGTCGTCTTTAGGTTGGGGGGAGGGGTTTTATGCGATGGAGTTTCCCCACACTGAGTGGGTGGAGACTGA
AGTTAGGCCAGCTTGGCACTTGATGTAATTCTCCTTGGAATTTGCCCTTTTTGAGTTTGGATCTTGGTTCATTCTCAAGCCTCAGACAGTGGTTCAAAGTTT
TTTTCTTCCATTTCAGGTGTCGTGAGGATCTATTTCCGGTGAATTCCTCGAGACTAGTTCTAGATGGTGAGCAAGGGCGAGGAGGATAACATGGCCATCATC
AAGGAGTTCATGCGCTTCAAGGTGCACATGGAGGGCTCCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAGGGCGAGGGCCGCCCCTACGAGGGCACCCAG
ACCGCCAAGCTGAAGGTGACCAAGGGTGGCCCCCTGCCCTTCGCCTGGGACATCCTGTCCCCTCAGTTCATGTACGGCTCCAAGGCCTACGTGAAGCACCCC
GCCGACATCCCCGACTACTTGAAGCTGTCCTTCCCCGAGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGACGGCGGCGTGGTGACCGTGACCCAGGAC
TCCTCCCTGCAGGACGGCGAGTTCATCTACAAGGTGAAGCTGCGCGGCACCAACTTCCCCTCCGACGGCCCCGTAATGCAGAAGAAGACCATGGGCTGGGAG
GCCTCCTCCGAGCGGATGTACCCCGAGGACGGCGCCCTGAAGGGCGAGATCAAGCAGAGGCTGAAGCTGAAGGACGGCGGCCACTACGACGCTGAGGTCAAG
ACCACCTACAAGGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTCAACATCAAGTTGGACATCACCTCCCACAACGAGGACTACACCATCGTGGAA
CAGTACGAACGCGCCGAGGGCCGCCACTCCACCGGCGGCATGGACGAGCTGTACAAGGAGGGCCGGGGCAGCCTGCTGACCTGCGGCGACGTGGAGGAGAAC
CCCGGCCCCATGACCGAGTACAAGCCCACGGTGCGCCTCGCCACCCGCGACGACGTCCCCAGGGCCGTACGCACCCTCGCCGCCGCGTTCGCCGACTACCCC
GCCACGCGCCACACCGTCGATCCGGACCGCCACATCGAGCGGGTCACCGAGCTGCAAGAACTCTTCCTCACGCGCGTCGGGCTCGACATCGGCAAGGTGTGG
GTCGCGGACGACGGCGCCGCGGTGGCGGTCTGGACCACGCCGGAGAGCGTCGAAGCGGGGGCGGTGTTCGCCGAGATCGGCCCGCGCATGGCCGAGTTGAGC
GGTTCCCGGCTGGCCGCGCAGCAACAGATGGAAGGCCTCCTGGCGCCGCACCGGCCCAAGGAGCCCGCGTGGTTCCTGGCCACCGTCGGCGTCTCGCCCGAC
CACCAGGGCAAGGGTCTGGGCAGCGCCGTCGTGCTCCCCGGAGTGGAGGCGGCCGAGCGCGCCGGGGTGCCCGCCTTCCTGGAGACCTCCGCGCCCCGCAAC

CTCCCCTTCTACGAGCGGCTCGGCTTCACCGTCACCGCCGACGTCGAGGTGCCCGAAGGACCGCGCACCTGGTGCATGACCCGCAAGCCCGGTGCCTGAACG
CGTTAAGTCACCCAGCTTTCTTGTACAAAGTGGTGATAACTCTAGAGAATTCACTCCTCAGGTGCAGGCTGCCTATCAGAAGGTGGTGGCTGGTGTGGCCAA
TGCCCTGGCTCACAAATACCACTGAGATCTTTTTCCCTCTGCCAAAAATTATGGGGACATCATGAAGCCCCTTGAGCATCTGACTTCTGGCTAATAAAGGAA
ATTTATTTTCATTGCAATAGTGTGTTGGAATTTTTTGTGTCTCTCACTCGGAAGGACATATGGGAGGGCAAATCATTTAAAACATCAGAATGAGTATTTGGT
TTAGAGTTTGGCAACATATGCCATATGCTGGCTGCCATGAACAAAGGTGGCTATAAAGAGGTCATCAGTATATGAAACAGCCCCCTGCTGTCCATTCCTTAT
TCCATAGAAAAGCCTTGACTTGAGGTTAGATTTTTTTTATATTTTGTTTTGTGTTATTTTTTTCTTTAACATCCCTAAAATTTTCCTTACATGTTTTACTAG
CCAGATTTTTCCTCCTCTCCTGACTACTCCCAGTCATAGCTGTCCCTCTTCTCTTATGAAGATCCCTCGACCTGCAGCCCAAAAAAAAGCACCGACTCGGTG
CCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACNNNNNNNNNNNNNNNNNNNNCGGTGTTTCGTCCTTTCCAC
AAGATATATAAAGCCAAGAAATCGAAATACTTTCAAGTTACGGTAAGCATATGATAGTCCATTTTAAAACATAATTTTAAAACTGCAAACTACCCAAGAAAT
TATTACTTTCTACGTCACGTATTTTGTACTAATATCTTTGTGTTTACAGTCAAATTAATTCTAATTATCTCTCTAACAGCCTTGTATCGTATATGCAAATAT
GAAGGAATCATGGGAAATAGGCCCTCTTCCTGCCCGACCTT

[0048] 상기 서열번호 1에서, “NNNNNNNNNNNNNNNNNNNN”은 L1 요소(element)를 표적으로 하는 sgRNA를 나타내는 염기서열로, 상기 서열번호 1 에 표시된 N의 수는 예시적인 것으로, 반드시 20개의 염기서열로 한정되는 것은 아니며, 표적하는 L1 요소에 따라 특이적으로 변경될 수 있다.

[0049] 우리는 L1 요소(element)를 표적으로 하는 최적 sgRNA를 선별하기 위해 다음과 같이 실험을 수행하였다.

[0050] 우리는 먼저, 인간 L1 레트로트랜스포존 영역으로 알려진 증폭된 영역을 평가했다. 단일 프라이머 쌍을 사용하여 L1 레트로트랜스포존 영역을 증폭시켰다 (도 2a). 이 영역은 구별되는 서열의 수를 최대화하고 균일한 증폭 가능성을 증가시키기 위해 선택되었다. 예상한 바와 같이, 앰플리콘 크기는 바이모달(Bimodal)로 분포되었으며, 99 %의 영역이 알려진 L1 서브 패밀리와 중첩되어 있었다.

[0051] 그 다음 우리는 표적화된 디아미나제를 사용하여 유전자 편집 후 바코드 역할을 할 수 있는 내재 영역을 검색하였다. 기존의 CRISPR/Cas9 시스템과 마찬가지로 표적화된 디아미나제는 'NGG'PAM 서열에 인접한 20bp 프로토스페이서 (protospacer)를 표적으로 삼아 특정 프로토스페이서 서열 창 내부에서 C > T 전환을 유도한다. 개념 증명을 위해, 우리는 레트로트랜스포존 영역에서 'NGG' PAM 서열을 갖는 것들을 동정하기 위해 모든 표적 가능한 프로토스페이서 영역을 스크리닝하고, PAM 서열 옆의 프로토스페이서의 4-8 뉴클레오티드 윈도우상에 C를 가진 후보 프로토스페이서 영역을 컴파일하였다. 우리는 우리의 확립된 조건을 만족시키는 동일한 스페이서 서열을 갖는 여러 sgRNA 후보 물질이 존재함을 확인했다. 우리는 표적화된 반복 부위에서 가장 일치하는 부위 수가 가장 많은 두 개의 프로토스페이서 서열을 선택했다. sgRNA-2와 sgRNA-1의 서열은 거의 동일했다 (하나의 염기 차이만). 따라서, sgRNA-2는 추후 실험에서 생략하였다. 우리는 하나의 sgRNA가 L1 영역 내 다중 표적 부위에서 치환을 도입하고 C > T 치환으로 '세포 바코드'를 정의할 것으로 예상했다. 각 sgRNA의 다중 표적 부위는 동일한 프로토스페이서 서열을 가지고 있지만, 증폭 후에 주변 서열을 통해 표적 부위를 구별할 수 있었고, 특정 유전자 위치 (도 2b)에 유일하게 정렬될 수 있었다.

[0052] 상기 제조된 플라스미드 벡터를 형질감염시키기 위한 세포주들은 다음과 같이 준비되었다: 모든 세포주는 KCLB (Korean Cell Line Bank)에서 얻은 후 37 ℃에서 5 % CO_2로 유지 하였다. HEK293T 인간 배아 신장 및 HeLa 인간 자궁 경부암 세포주를 10 % 소 태아 혈청 (FBS, Gibco, USA) 및 1 % 페니실린/스트렙토마이신 (P/L; Thermo Fisher Scientific, USA)을 첨가한 Dulbecco 's Modified Eagle 's Medium (DMEM; Gibco, USA)에서 배양하였다.

[0054] **[실험예 1] 내재 유전자 바코드 증폭-벌크 세포**

[0055] 세포로부터 추출한 gDNA를 내재 유전자 바코드의 증폭에 사용하였다. Kapa High Fidelity 중합 효소 (Kapa BioSystems, USA)를 모든 바코드 증폭에 사용하였다. 최대 500ng의 gDNA를 10μM 정방향 및 역방향 프라이머 (표 1의 L1 사이트 정방향(서열번호 4) 및 L1 사이트 역방향(서열번호 5)) 각각 1㎕, 10㎕ KAPA DNA 중합 효소 및 8㎕ 뉴클레아제(nuclease) 없는 물을 포함하는 20㎕ 시작 PCR 반응물에 로딩하고 다음 프로토콜에 따라 시퀀싱 어댑터를 갖는 프라이머를 사용하여 증폭하였다: 98℃에서 120초 후, 98℃에서 10초, 57℃에서 120초 및 72℃에서 120초 2 사이클, 및 72℃에서 최종 10분. Sera-Mag SpeedBeads (6515-2105-050350, Thermo Scientific, USA)를 사용하여 만든 홈메이드 AMPure XP 비즈 (이후 AMPure 비즈)를 사용하여 시작 PCR 산물을 정제하였다. 시작 PCR 산물을 단일 20 ㎕ 두 번째 인덱스 PCR 반응물에 넣고 다음 프로토콜을 사용하여 인덱스 프라이머로 증폭시켰다: 98℃에서 30 초 후, 98℃에서 10초, 60℃에서 30초 및 72℃에서 30초 15 사이클, 및 72℃에서 최종 10분. 그 다음 두 번째 PCR 산물을 1.2 × AMPure beads를 사용하여 정제하였다. 모든 프라이머는 IDT (Integrated DNA Technologies, USA)에 의해 제조하였다. 시퀀싱은 NextSeq 500/550 High Output v2 키트

(300 사이클) (Illumina, USA)를 사용하여 Illumina NextSeq 500 시스템에서 수행하였다.

표 1

Primers	Sequences (5' to 3')
L1 site for	ACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNNNNNNNNACACAGGGAGGGGAACAT (서열번호 4)
L1 site rev	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTTGCCATGGTGGTTTGCT (서열번호 5)
sgRNA-1 sgRNA for	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGATGGGTGCAGCAAACCACCA (서열번호 6)
sgRNA-1 sgRNA rev	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACTGGTGGTTTGCTGCACCCATC (서열번호 7)
sgRNA-3 sgRNA for	TTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGAAATACCTAATGTAGATGA (서열번호 8)
sgRNA-3 sgRNA rev	GACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAACTCATCTACATTAGGTATTTCC (서열번호 9)

[0057]

[0058] **[실험예 2] 내재 유전자 바코드 증폭-단일 세포**

[0059] illustra™ Ready-To-Go GenomiPhi V3 DNA 증폭 키트 (GE Healthcare, USA)를 사용하여 단일 세포의 전체 게놈 다중 변이 증폭 (Whole-genome multiple displacement amplification, MDA)을 수행하였다. MDA는 제조사의 프로토콜에 따라 1 시간 30 분에서 3 시간으로 반응 시간을 약간 변경하여 수행하였다. 다음으로, 단일 세포 MDA 생성물 5 ㎕를 10 μM 정방향 및 역방향 프라이머 (표 1의 L1 사이트에 대한 정방향(서열번호 4) 및 역방향(서열번호 5)) 각각 1 ㎕, 10 ㎕ KAPA DNA 중합효소, 3 ㎕의 뉴클레아제가 없는 물을 함유하는 시작 어댑터 PCR 반응물 20 ㎕에 첨가하였다. PCR 반응은 다음 프로토콜을 사용하여 수행하였다: 98℃에서 2분 후, 98℃에서 10초, 57℃에서 2분 및 72℃에서 2분 10사이클, 및 72℃에서 최종 2분. AMPure 비드를 사용하여 정제한 후 초기 PCR 산물을 두 번째 인덱스 PCR 반응에 넣고 다음의 프로토콜을 사용하여 PCR을 수행하였다 : 98℃에서 30초 후, 98℃에서 10초, 60℃에서 30초 및 72℃에서 30초 15사이클, 및 72℃에서 최종 10분. 이어서, AMPure 비드를 사용하여 최종 생성물을 정제하였다.

[0060] 수동으로 채취한 단일 세포의 단일 세포 PCR 증폭은 위에서 설명한 PCR 반응 조성 및 프로토콜을 사용하여 수행하였다: 1. 어댑터 (시작) PCR : 98℃에서 2분 후, 98℃에서 10초, 57℃에서 2분 및 72℃에서 2분 30 사이클, 및 72℃에서 최종 2분. 이어서, 생성물을 AMPure 비드를 사용하여 정제하였다. 2. 인덱스 (두번째) PCR: 98℃에서 30초 후, 98℃에서 10초, 60℃에서 30초 및 72℃에서 30초 15사이클, 및 72℃에서 최종 10분. 이어서, 2차 PCR의 생성물을 2% 아가로오스겔 상에 로딩하고 겔 전기 영동으로 분리하였다. 예상된 크기의 밴드를 제조사의 프로토콜에 따라 QIAquick Gel Extraction Kit (QIAGEN, USA)를 사용하여 정제하였다. 마찬가지로, 시퀀싱은 NextSeq 500/550 High Output v2 키트 (300 사이클)를 사용하여 Illumina NextSeq 500에서 수행되었다.

[0062] **[실험예 3] 증폭 후 정렬 과정의 분석**

[0063] 리드들을 BWA (v0.7.12-r1039)를 사용하여 hg19에 정렬하였고 indels (RealignerTargetCreator, IndelRealigner) 주위의 재정렬은 GATK (v3.3-0)를 사용하여 수행하였다. 위치별 염기 검출은 SAMtools (v1.1) *mpilup* 함수를 사용하여 수행하였으며 *pileup* 파일은 사용자 정의 변형 검출에 사용하였다. 정렬된 영역에는 RepeatMasker (http://www.repeatmasker.org)를 사용하여 주석을 달고 증폭된 영역의 크기를 플롯하여 겹쳐진 부분을 계산하였다.

[0065] **[실험예 4] 계통 재구축을 위한 확실한 사이트의 동정**

[0066] 우리는 계통 재건축을 위한 확실한 마커 (C > T 치환)를 추출하기 위해 FreeBayes (v1.1.0-3-g961e5f3)를 사용하는 변형 검출 방법을 채택하였다. 상기 변형 검출은 FreeBayes (indel 재정렬 후 BAM으로부터 입력)를 사용하

였고 필터링된 위치 (깊이>10)는 후보 마커를 고려하였으며 빈 벡터를 사용하여 배경 대조군에 대해 계산된 값보다 높은 대립 유전자 빈도를 포함하는 마커만을 포함시켰다. HeLa 세포를 포함한 벌크 및 단일 세포 계통 추적 실험에서 변형된 파라미터(--ploidy 3, --pooled-discrete)를 사용하여 변형 검출을 수행하였다. 벌크 데이터 및 단일 세포 데이터를 효율적으로 처리하기 위해 표적화된 디아미나제 시스템을 기반으로 한 변형 검출을 위한 맞춤 알고리즘을 개발하였다. 우리는 이전 연구에서 설명한 바와 같이 조건부 확률을 갖는 이항 혼합 모델을 사용하는 확률론적 접근 방식을 채택했다(predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26,** 730-6 (2010)). 예상 최대화 알고리즘(expectation-maximization algorithm)은 불안정한 게놈 (예를 들어, 상이한 원형질을 갖는 게놈)에서 대립 유전자 빈도의 고유 편차를 설명하는 모델 파라미터를 측정하기 위해 사용되었다. 표적 부위의 모든 후보 위치, 깊이>10배, 변이 대립 유전자 수>2, 및 사후 확률≥0.95가 최종 마커로 선택되었다. 벌크 노드에 있는 모든 마커에 대해 합집합 연산을 수행한 후, 다음과 같은 기준을 사용하여 확실한 마커를 선택하였다. 1) 0.1 및 2의 최소 편집 효율을 갖는 부위 및 2) 매우 높은 관련성을 갖는 마커(≥0.9) 제거. 각 세포에 대한 이러한 마커들의 합(sum)은 최종 '세포 바코드'를 나타낸다.

[0068] **[실험예 5] 벌크 및 단일 세포 실험에 대한 세포 계통 트리 구축**

[0069] 벌크 세포 실험의 경우 이전 방법 (https://bitbucket.org/Bastiaanspanjaard/linnaeus)과 유사한 방법을 사용하여 트리 건축을 수행하였다. 우리는 염기 편집 패턴을 사용하여 치환 그래프를 작성하였다. 단순화를 위해, 노드는 CIGAR 문자열 형 시퀀스로 식별되었다 (예를 들면, 1E10E는 완벽한 온-타겟 영역의 첫 번째 및 열 번째 C 위치가 편집되었음을 의미한다).

[0070] 그래프 재구축 전략은 먼저 DFS(depth-first search)을 사용하여 편집된 부위를 사용하여 가장 강력한 연결 구성 요소를 식별한다. 연결된 구성 요소를 식별하는데 사용되는 기존 알고리즘과는 달리, 우리는 처음에 연결된 구성 요소의 가중치를 최대화하기 위해 DFS 방식을 사용하였다 (구성 요소의 시퀀싱 깊이의 합이 최대화되었다). 그래프 검색(graph search)은 깊이가 큰 구성 요소를 우선시하므로 DFS 기반 알고리즘은 공유 편집의 깊이가 비정상적으로 높기 때문에 여러 클래이드(clade)에 배치되는 노드를 발생시킨다. 따라서 노드들 간에 편집된 바코드의 겹치는 부분을 기반으로 연결된 구성 요소를 식별하는 알고리즘을 수정하였다 (동일한 클래이드의 노드는 다른 클래이드의 노드보다 더 많은 바코드를 공유함). 그렇게 함으로써, 클래이드 구별에 오류가 발생하지 않았다. 클래이드 내에 모-녀(mother-daughter) 관계를 지정하는 데 최소한의 오차만 발생하였다 (HEK293T 및 HeLa 세포 모두를 포함하는 실험의 정확도가 97 %로 증가). 올바른 클래이드 내 모녀 관계를 지정하는 것은 편집된 바코드의 연속 누적에 크게 의존한다. 따라서, 나머지 오류는 일부 PCR 또는 시퀀싱 오류가 특정 노드의 최종 바코드 조합에 기여하여 미묘한 모-녀 관계의 잘못된 지정을 초래한다는 점에서 대량 시퀀싱 결과의 특성에 부분적으로 기인한 것 같다. 예를 들어 한 모(mother) 노드로부터의 녀(daughter) 노드가 다른 모 노드로부터의 녀 노드에 속해야 하는 편집 결과를 우연히 다른 모 노드와 겹치게 하면 알고리즘이 노드를 잘못 배치한다. 왜냐하면 바코트 조합이 조상과 얼마나 공유되어 있느냐에 따라 모-녀 관계가 지정되기 때문이다.

[0071] 본질적으로 첫 번째 조상 노드 (벌크 세포)는 다른 노드들과 최대한의 연결점을 가진다. 반복적으로 이 노드를 제거하고 나머지 연결된 구성 요소에서 조상 노드를 식별하였다. 이 절차를 모든 노드가 지정될 때까지 반복하였다. 모든 세포 네트워크가 구축되면, 세포를 그래프에 배치되었다. 이 연구에서는 scRNA-seq가 사용되지 않았기 때문에 '세포 이중렛 (cell doublet)'검출 임계값을 사용하지 않았다.

[0072] 단일 세포 기반 계통 추적을 위해 저속 이미징 실험을 다음과 같이 수행하였다. 1 % P/S, 1 % 비 필수 아미노산 (NEAA) (Gibco, USA), 100 mM 2-메르캅토에탄올 (2-Mercaptoethanol, 2-ME) (Sigma-Aldrich) 및 10 % FBS 를 첨가한 2 mm DMEM 35 mm 접시에서 계대 배양하여 HeLa 세포를 준비하였다. 세포를 전기 천공 시스템 (Neon, Invitrogen, 전압: 1140v, 펄스폭 범위 : 40 ms, 펄스 번호 : 1)을 사용하여 플라스미드로 형질감염시켰다. 형질감염 4 시간 후, 배양 배지를 새로운 배지로 교체하여 죽은 세포를 제거하였다.

[0073] 모든 단일 세포 조작은 반사 현미경 하에서 관찰하는 동안 마이크로 조작 장치 (Nikon-Narishige, Tokyo, Japan)를 사용하여 수행되었다. 형질 감염 후 1 일째, 세포를 트립신 처리하고 인산 완충 식염수 (PBS; 미국, Gibco, USA)로 세척하였다. 수작업으로 단일 세포를 선택하기 위해, 세포 현탁액을 0.5 % FBS가 함유된 PBS 방울에 넣고 미네랄 오일 (Sigma, USA)로 덮었다. 단일 RFP-양성 세포만을 형광 노출 하에 마이크로 인젝션 피펫 (직경: 20 ㎛, ORIGIO, Charlottesville, VA)을 사용하여 흡인시켰다. 흡인된 단일 세포를 100 mm 접시에서 1 % P/S, 1 % NEAA, 100 mM 2-ME 및 10 % FBS가 보충된 DMEM의 4 μl 액적(droplet)으로 옮기고 미네랄 오일을 각 액적 당 하나의 세포 비율로 오버레이드시켰다. 단일 세포를 37 ℃의 CO_2배양기에서 4 시간 동안 배양

한 후, 라이브, 저속 이미징을 사용하여 세포 성장을 관찰할 수 있는 JuLI™ Stage 실시간 세포 역사 기록 장치 (NanoEnTek)가 장착된 배양기로 옮겼다.

[0074] 단일 세포 기반 계통 추적의 경우 부위 편집 여부에 관계없이 정보가 제한되었다. 확실한 마커를 확인하기 위해 블랙리스트 후보 지역 (mCherry 신호가 없거나 비히클 대조군 단일 세포를 나타내는 단일 세포 결과들의 통합)도 필터링하였다. 벌크 세포 계통 구축과는 달리, 저속 이미지 촬영 기반 단일 세포 실험은 마지막 확장 깊이로부터의 세포를 포함하였다. 따라서 계통 추적은 다른 논리를 사용하여 수행되었다. 세포 사이의 거리는 Jaccard 인덱스를 사용하여 계산되었고, 계층형 클러스터링은 R에서 *pvclust* 함수를 사용하여 수행되었다. 대략 편차 없는 확률값 (p-값)은 1,000 번의 반복을 기반으로 계산되었다.

[0076] **[실험예 6] 편집 효율성 역학 추정**

[0077] 세포를 상술한 바와 같이 표적화된 디아미나제 벡터로 형질감염시켰다. 4, 8, 12, 16, 20, 24, 30, 36, 42 및 48 시간 후, 벌크 세포를 수집하고 시퀀싱을 위해 증폭시켰다 (2개의 복제물). t = 0에서 1의 야생형 비율 (100 %)을 가정하여, 지수 함수를 피팅함으로써 편집 속도 (λ)를 계산하였다. 우리는 완벽한 표적-대상 영역에서 편집되지 않은 부위 (C> T 후보 위치)의 비율로 야생형 비율을 결정하였다.

[0079] **[실시예 1] 유전자 바코딩 시스템을 보유하는 세포의 생성**

[0080] 먼저, 우리는 HEK293T 및 HeLa 세포의 L1 레트로트랜스포존 부위에 표적화된 디아미나제 시스템을 적용하여 계통 추적 실험에 사용될 수 있는지를 확인하기 위하여 다음과 같이 수행하였다.

[0081] 유전자 바코드를 보유하는 세포를 생성하기 위해 HEK293T 및 HeLa 세포를 Lipofectamine™ 3000 (Life Technologies, USA.)을 사용하여 2 : 1의 트랜스포존 (PB CMV-BE3 EF1α-mCherry-T2A-puro sgRNA) 및 트랜스포사제(transposase) 비율로 제조사의 프로토콜에 따라 유전자 바코딩 시스템 플라스미드 벡터로 개별적으로 형질감염시켰다. 형질 감염된 세포를 약 3일 동안 배양하고 DNeasy Blood & Tissue Kit (QIAGEN, USA)를 사용하여 게놈 DNA (gDNA)를 수확하였다.

[0082] 바코드 편집 효율을 측정하기 위하여 Lipofectamine™ 3000을 사용하여 트랜스포사제 없이 유전자 바코드 시스템만 형질감염시켰다. 형질감염된 세포를 4, 8, 12, 16, 20, 24, 30, 36, 42, 48 시간마다 수확한 다음, gDNA를 추출하였다.

[0083] 동일한 프로토스페이서 영역을 갖는 다수의 표적 부위 (sgRNA-1에 의해 표적화된)를 분석한 결과, HEK293T 및 HeLa 세포에 대한 평균 편집 효율 (C > T 치환 수의 분율)은 각각 1.5% 및 2.3%였는데 (114 및 143 세포 바코드), 이는 알려진 4-8 뉴클레오타이드 프로토스페이서 서열 창 (도 2c)의 다중 표적에 대해, 편집 세포의 수가 적음을 의미한다. 대조적으로, sgRNA-3는 sgRNA-1과 비교하여 평균 4배 높은 편집 효율 (HEK293T 및 HeLa 세포에서 각각 6.3 % 및 9.3 %)을 나타내었고, 프로토스페이서 서열 창에서 2개의 C 사이의 편집 효율의 상관관계는 다중 표적에서 높았다 (도 2d). 우리는 또한 표적 부위에 대해 두 세포주 간의 편집 효율의 차이를 관찰했다. 표적화된 디아미나제 및 sgRNA가 없는 비히클 대조군과 비교하여 유의한 배경 돌연변이는 검출되지 않았다 (P 값 <2.2e-16, Mann-Whitney U test). 뿐만 아니라, PAM 옆의 프로토스페이서에서 알려진 4-8 뉴클레오티드 창에서 C들을 제외한 비 표적 C들에 대해서는 매우 낮은 편집 빈도가 관찰되었다.

[0084] 다음으로, 우리는 표적 디아미나제 시스템이 표적 반복 영역에 유전 바코드를 연속적으로 도입할 수 있는지 여부를 조사했다. 정해진 시간 지점에서 연속 편집 전략의 범위를 탐구하기 위해 약 3일 간격으로 표적 디아미나제 시스템을 반복적으로 상술한 바와 같이 형질감염시켜 sgRNA-1및 sgRNA-3를 비교하였다 (도 3a). 평균 형질 전환 효율은 연속 형질 감염 후 편집된 부위가 점진적으로 축적됨에 따라 선형적으로 증가하였으며, 관찰된 편집 속도는 sgRNA-1과 비교하여 sgRNA-3을 사용하는 것이 더 빨랐다. 따라서 우리는 유전자 세포 바코드의 지속적인 도입이 우리의 방법을 사용하여 실현 가능하다고 결론지었다.

[0086] **[실시예 2] 조절된 시험관 트리 실험(tree experiment)을 이용한 벌크 수준에서 계통 추적**

[0087] 우리는 세포 바코드가 트리 구축(tree construction)을 위해 각 세대에 적절하게 도입되었는지 여부를 조사하기 위해 시험관내 세포 확장 실험을 다음과 수행하였다.

[0088] HEK293T 및 HeLa 세포를 제조예 1에 따라 제조된 PiggyBac™ 트랜스포존 시스템을 사용하여 형질감염시켰다. 바코드 생성에 사용된 절차는 위에서 설명한 프로토콜과 동일하며 푸로마이신 (2μg/ml)을 사용하여 성공적으로 형질감염된 세포를 선별하였다. 선별된 세포는 mCherry 형광 (형질감염의 마커)에 기초한 Aria II FACS 장치를 사용하여 단일 세포 FACS에 의해 96-웰 플레이트로 분류되었다. 분류된 단일 세포를 20 % FBS 및 1 % P/S가 보

충된 DMEM에서 배양하였다. 단일 세포 유래 개체군은 3주까지 배양하여 클론 확장시켰다. 그 다음 mCherry-양성 단일 세포를 분류하고 다른 웰로 옮겨 배양하는 과정을 반복했다 (도 4). 그 다음 상기 실험예 1 내지 4와 같은 방법으로 확장된 세포로부터의 표적 영역을 단일 프라이머 쌍을 사용하여 증폭시키고 차세대 시퀀싱 (NGS)을 수행한 후 정렬 및 변형 검출을 수행하였다. 이 시스템은 HEK293T 세포에 처음 적용되었다. 벌크 세포로 대표되는 각 노드(하나의 노드는 세포 바코드의 합계를 나타냄) 내의 알고 있는 트리 토폴로지는 계통 추적의 유효성을 검사할 수 있게 하였다. 평균적으로 고유 리드(read)들의 95%가 정렬되었고 이 리드들은 인접 영역 간의 상동성으로 인해 다중 정렬이 발생할 수 있으므로 더 처리되었다. 정렬 및 변형 검출(variant calling) 후, 우리는 트리의 1 세대에서 노드 당 평균 5개의 세포 바코드를 발견했으며 대상 부위의 ~93%가 sgRNA-1에 의해 편집되지 않았음을 발견했다.

[0089] 그 다음, 트리 구축은 실험예 5와 같이 반복적인 그래프 접근법(iterative graph approach) 및 추가적인 post-hoc 세포 바코드 선택 단계(post-hoc cell barcode selection step)을 사용하여 수행되었다. sgRNA-1의 경우 정보를 제공하는 세포 바코드 수가 적기 때문에 트리를 올바르게 식별할 수 없었다. 반대로, sgRNA-3는 1 세대 노드에 대해 평균 29 개의 세포 바코드를 나타냈다. 재구성된 트리의 정확도는 알고 있는 트리의 깊이(depth) 및 위치상에 올바른 노드 배치 비율에 따라 정의되었다. sgRNA-3의 경우 변형 검출 방식(variant calling approach)을 기반으로 재구성된 트리는 81% 정확성 (29/36)을 나타냈다. 기존의 변형 검출 방식은 작은 변형은 반환하기 때문에 확실한 세포 바코드를 얻어 잘못된 연결을 제거하고 재구축된 트리를 개선하기 위해 우리는 맞춤 알고리즘을 개발하였다. 트리 구축을 위한 최종 후보 세포 바코드를 선택하기 위해 사후 계산(posterior calculation)을 포함한 확률론적 접근법을 사용했다. 기존의 변형 검출 방식(variant calling approach)과 비교하여 sgRNA-3에 대해 평균 70개의 세포 바코드를 관찰했으며 재구축 정확도가 97 % (35/36)로 향상되었다(도 3b). HeLa 세포의 경우, 정확도면에서 약간 개선된 성능을 보였다 (기존 변형 검출 대 맞춤 알고리즘의 경우 각각 88 % 및 97 % (59/61)) (도 5).

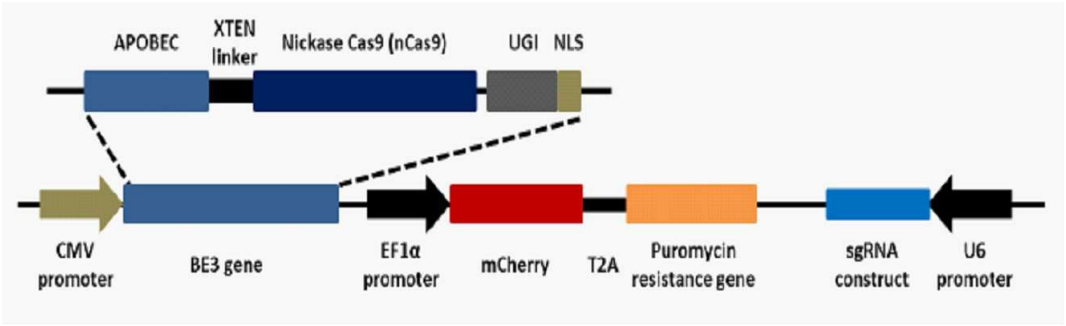
[0091] **[실시예 3] 단일 세포 수준에서 계통 관계의 재구축**

[0092] 우리는 표적화된 디아미나제 시스템이 단일 세포 수준에서 계통 관계의 재구축이 가능한지 여부를 확인하였다. sgRNA-3이 sgRNA-1보다 트리 재구축 효율성이 더 좋았기 때문에 sgRNA-3에만 초점을 맞추고 단일 세포 분리의 용이성을 위해 단일 세포 수준 계통 추적에 HeLa 세포를 선택했다. HeLa 세포는 mCherry 형광 단백질, 표적화된 디아미나제 및 sgRNA-3를 함유하는 제조예 1의 PiggyBac™ 트랜스포존 시스템으로 형질감염시켰다. 우리는 그라운드 트루 트리 데이터(ground truth tree data)를 생성하기 위해 저속 이미징(time-lapse imaging)을 사용하고 매뉴얼에 의해 mCherry마커 양성인 개별 세포 ([3개의 다른 트리]로 분석된 총 n=32의 단일 세포)를 선택했다(도 3c). 시퀀싱 실험을 준비하기 위해 먼저 선택된 단일 세포의 전체 게놈 증폭(whole-genome amplification, WGA)과 후속 PCR 증폭을 수행했다. 그러나 시퀀싱 리드들의 불규칙한 분포는 세포 바코드 식별을 방해하고 WGA 동안 높은 변성 온도로 인해 증가된 백그라운드 C>T 돌연변이가 발생할 수 있다고 보고되어 있다. 따라서 단일 세포 PCR 조건을 실험예 2와 같이 최적화하여 실시하였다. 최적화 후, 우리는 표적 영역에 대해보다 균일한 깊이 분포를 달성했다 (도 6). 확장된 3 내지 4개의 분류(8-16 세포)에 대한 세 가지 다른 실험 트리에 대해 표준 응집형 계층적 클러스터링 접근법(standard agglomerative hierarchical clustering approach)이 사용되었다. 확실한 세포 바코드는 이진 상태(binary state)로 인코딩되었고 트리 재구축을 위해 세포 대 세포 사이의 거리가 계산되었다. 평균 연계법을 사용한 계층적 클러스터링은 이미징 실험에서 검증된 알고 있는 트리들을 고도화된 코펜네식 상관관계(cophenetic correlation)(0.92, 0.91 및 0.81)로 일관되게 복원하였다(도 3d 및 도 7).

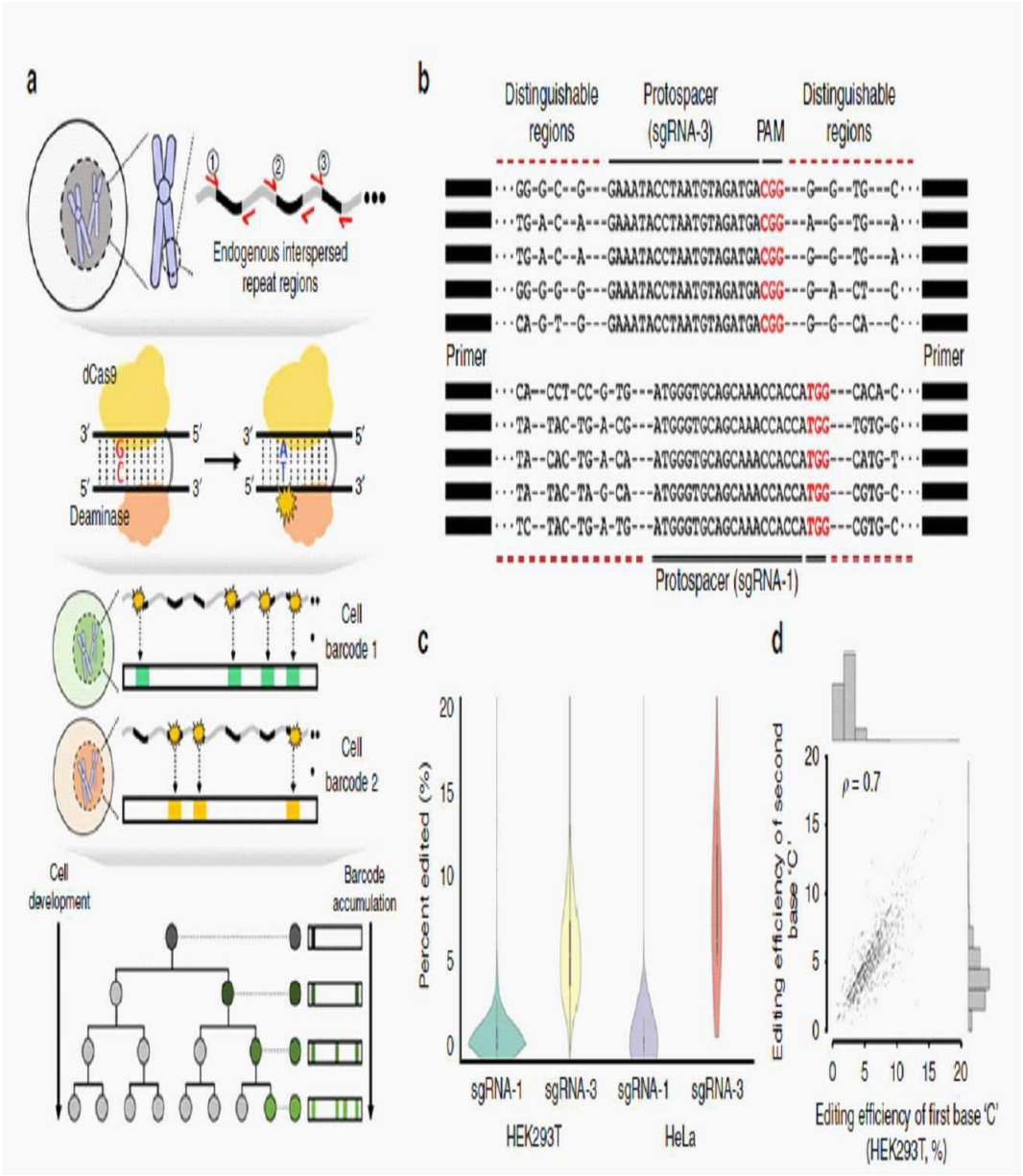
[0093] 우리의 플랫폼을 사용하여 생성된 계통 재구축의 정확성에 편집 속도가 영향을 미칠 수 있는지를 결정하기 위해, 실험예 6에 따라, sgRNA-3의 편집 속도를 근사치화함으로써 누적 돌연변이률을 측정하였다 (도 8). 지수 피팅(exponential fitting) 후 시간당 0.06의 편집 속도는 실험적 편집 속도를 정확하게 반영한다. 이 파라미터를 편집 효율로 사용하여 실험 결과를 바탕으로 실험 대상 부위의 수를 사용하여 세포 확장 당 세포 바코드 수를 예측하는 시뮬레이션을 수행했다. 시뮬레이션 결과는 트리의 각 깊이에서 생성된 바코드 수가 단일 세포 수준에서 얻은 실험 결과를 정확하게 반영한다는 것을 나타내었다 (세포 당 평균 세포 바코드 수 비교, 단일 세포 실험 대 시뮬레이션, 트리 1: 6.4 대 6.1, 트리 2: 20.3 대 18, 트리 3: 9 대 9.5, 도 3d 및 도 7). 대상 부위의 수 및 편집 속도를 변수로 포함하는 모델을 사용하여 우리는 정확성 (Cophenetic correlation)이 대상 부위 및 편집 속도의 증가와 함께 향상된다는 것을 확인하였다 (도 9).

도면

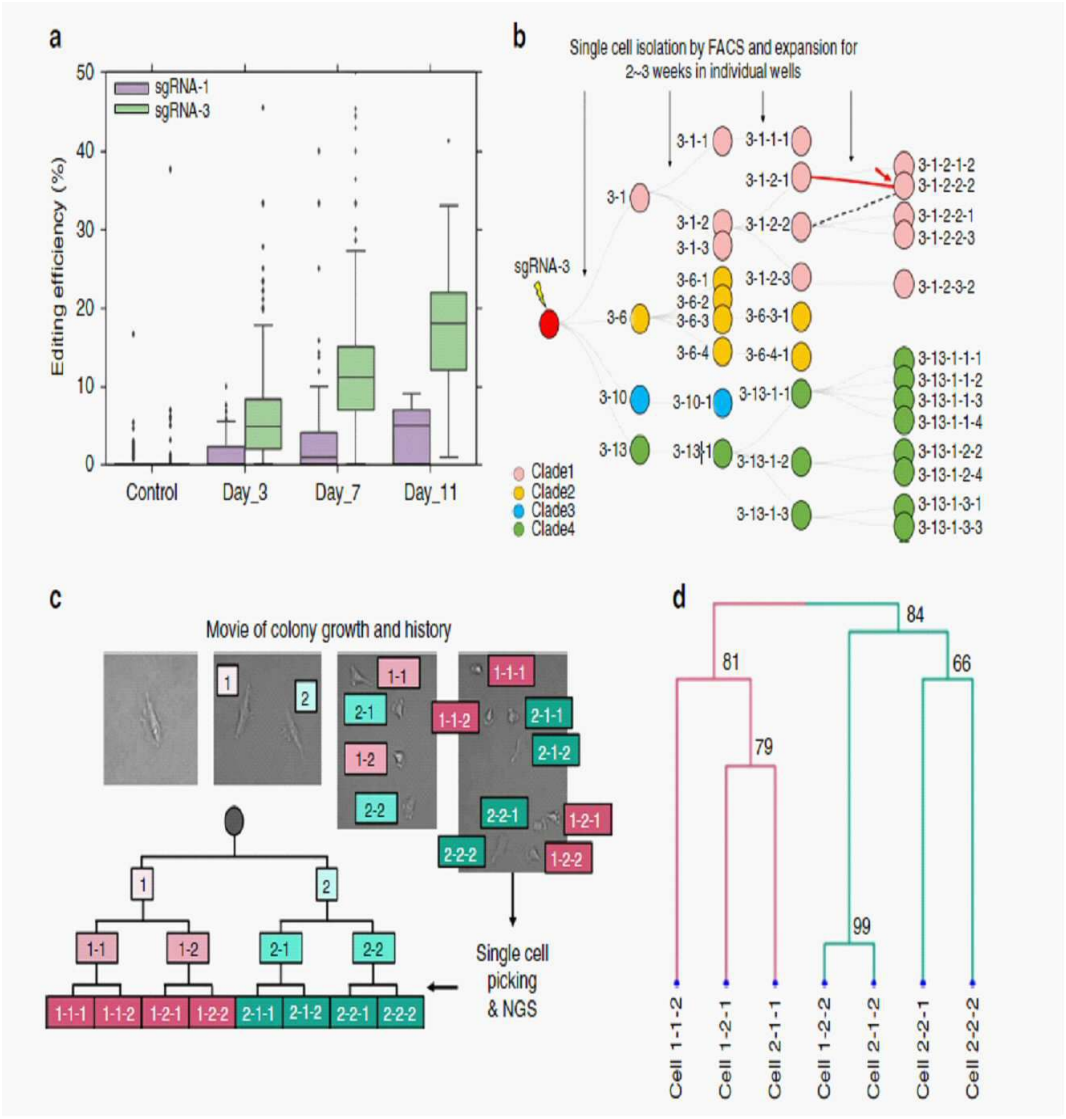
도면1


APOBEC
XTEN linker
Nickase Cas9 (nCas9)
UGI
NLS
CMV promoter
BE3 gene
EF1α promoter
mCherry
T2A
Puromycin resistance gene
sgRNA construct
U6 promoter

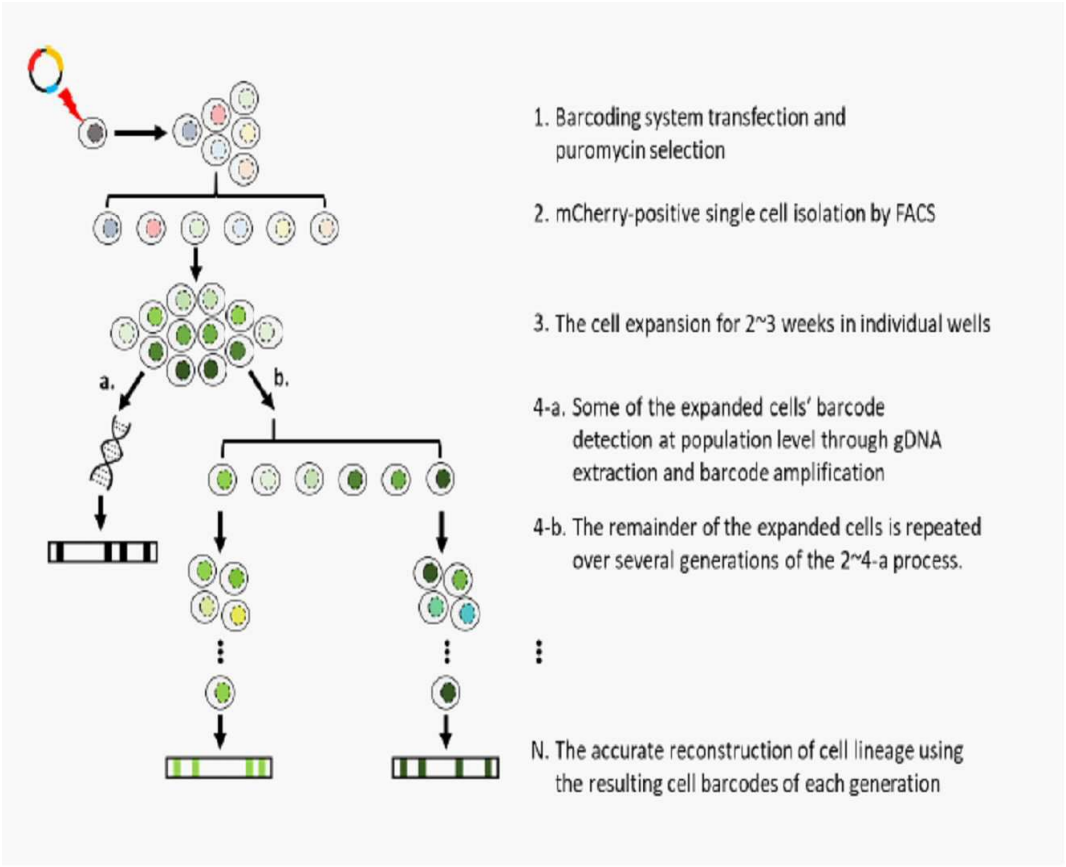
도면2


a
Endogenous interspersed repeat regions
dCas9
Deaminase
Cell barcode 1
Cell barcode 2
Cell development
Barcode accumulation
b
Distinguishable regions
Protospacer (sgRNA-3)
PAM
Distinguishable regions
Primer
Primer
Protospacer (sgRNA-1)
c
Percent edited (%)
sgRNA-1
sgRNA-3
sgRNA-1
sgRNA-3
HEK293T
HeLa
d
Editing efficiency of second base 'C'
ρ = 0.7
Editing efficiency of first base 'C' (HEK293T, %)

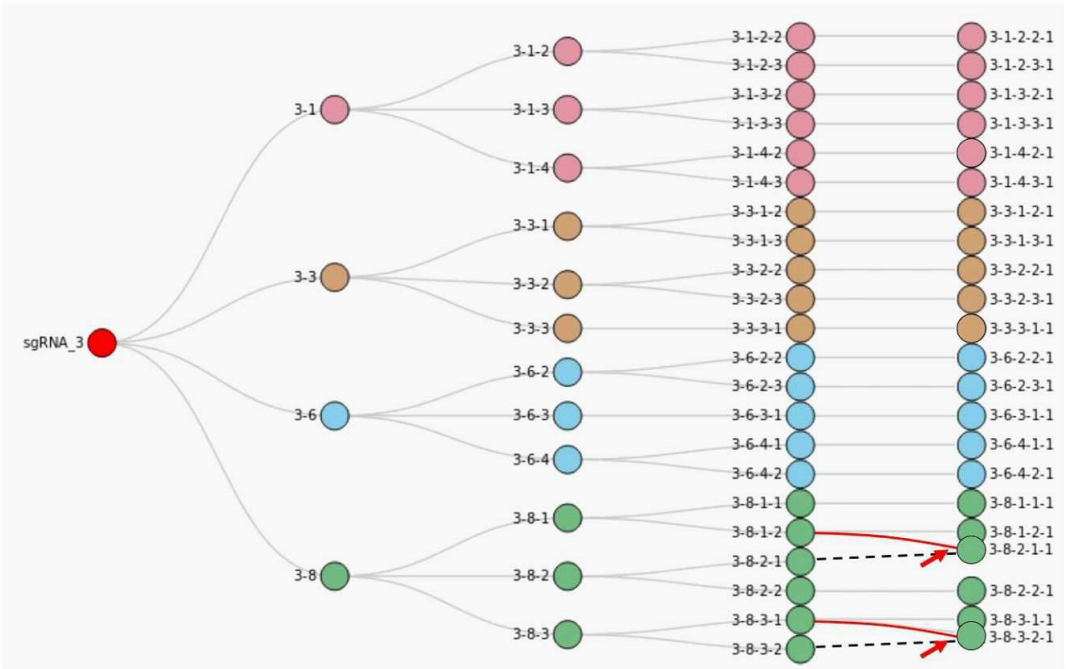
도면3


a
sgRNA-1
sgRNA-3
Editing efficiency (%)
Control
Day_3
Day_7
Day_11
b
Single cell isolation by FACS and expansion for
2~3 weeks in individual wells
sgRNA-3
Clade1
Clade2
Clade3
Clade4
c
Movie of colony growth and history
Single cell
picking
& NGS
d
Cell 1-1-2
Cell 1-2-1
Cell 2-1-1
Cell 1-2-2
Cell 2-1-2
Cell 2-2-1
Cell 2-2-2

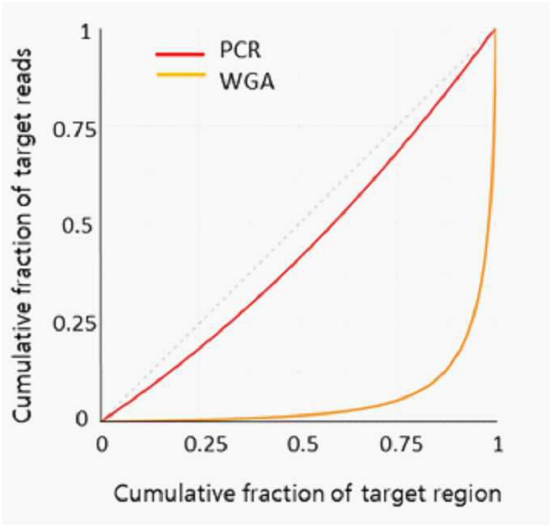
도면4


a.
b.
1. Barcoding system transfection and puromycin selection
2. mCherry-positive single cell isolation by FACS
3. The cell expansion for 2~3 weeks in individual wells
4-a. Some of the expanded cells' barcode detection at population level through gDNA extraction and barcode amplification
4-b. The remainder of the expanded cells is repeated over several generations of the 2~4-a process.
N. The accurate reconstruction of cell lineage using the resulting cell barcodes of each generation

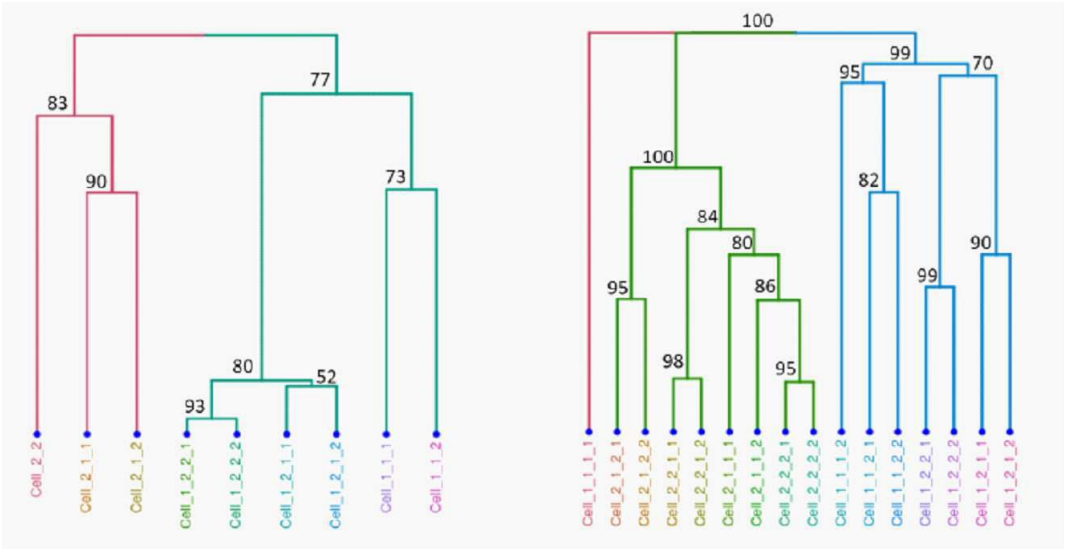
도면5


sgRNA_3
3-1
3-3
3-6
3-8
3-1-2
3-1-3
3-1-4
3-3-1
3-3-2
3-3-3
3-6-2
3-6-3
3-6-4
3-8-1
3-8-2
3-8-3
3-1-2-2
3-1-2-3
3-1-3-2
3-1-3-3
3-1-4-2
3-1-4-3
3-3-1-2
3-3-1-3
3-3-2-2
3-3-2-3
3-3-3-1
3-6-2-2
3-6-2-3
3-6-3-1
3-6-4-1
3-6-4-2
3-8-1-1
3-8-1-2
3-8-2-1
3-8-2-2
3-8-3-1
3-8-3-2
3-1-2-2-1
3-1-2-3-1
3-1-3-2-1
3-1-3-3-1
3-1-4-2-1
3-1-4-3-1
3-3-1-2-1
3-3-1-3-1
3-3-2-2-1
3-3-2-3-1
3-3-3-1-1
3-6-2-2-1
3-6-2-3-1
3-6-3-1-1
3-6-4-1-1
3-6-4-2-1
3-8-1-1-1
3-8-1-2-1
3-8-2-1-1
3-8-2-2-1
3-8-3-1-1
3-8-3-2-1

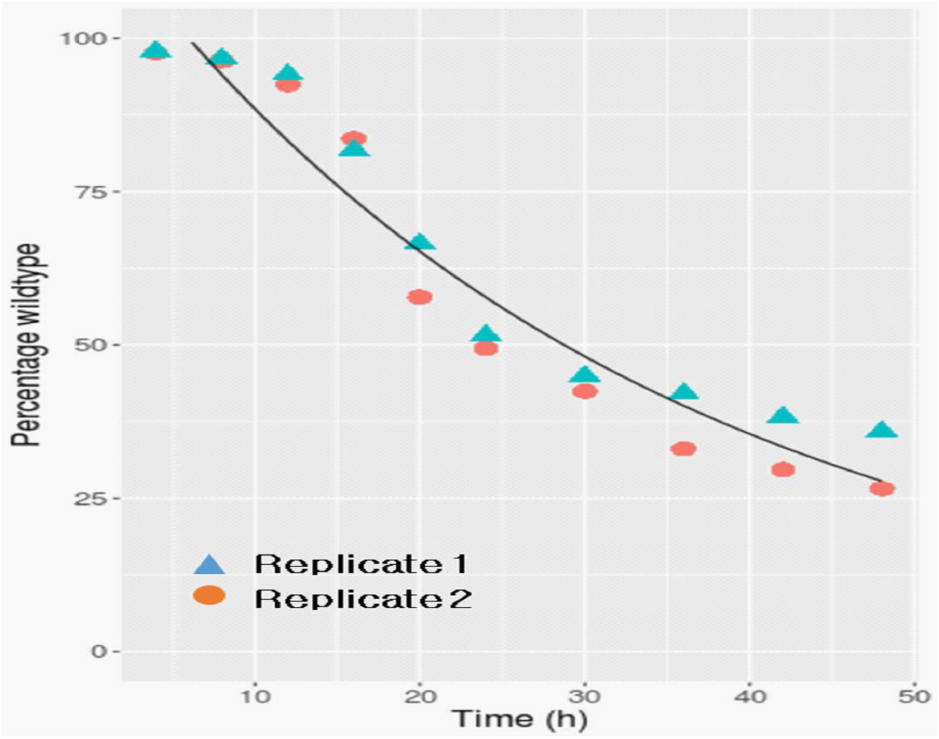
도면6


PCR
WGA
Cumulative fraction of target reads
Cumulative fraction of target region
0
0.25
0.5
0.75
1

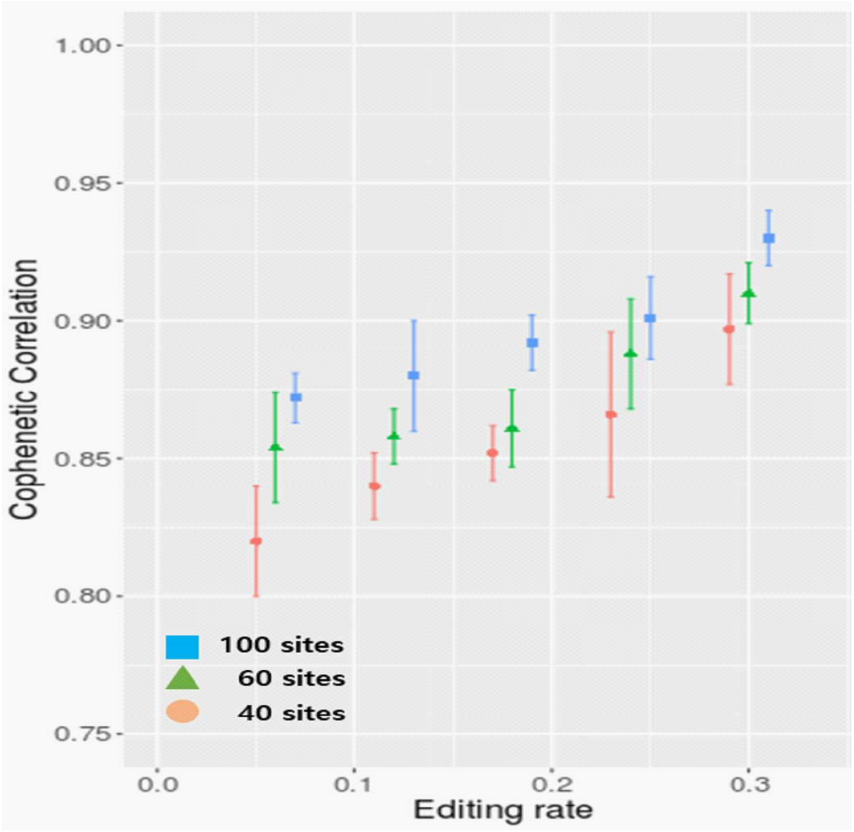
도면7


83
90
77
73
80
52
93
100
99
95
70
100
82
84
80
90
86
99
95
98
95

도면8



도면9



서 열 목 록

<110> Industry-Academic Cooperation Foundation, Yonsei University

<120> Development of cell barcoding technology using deaminase for endogenous repeat sequences

<130> P19U16C0595

<150> KR 0-2018-0075668

<151> 2018-06-29

<160> 9

<170> KoPatentIn 3.0

<210> 1

<211> 9425

<212> DNA

<213> Artificial Sequence

<220><223> PB CMV-BE3 EF1a-mCherry-T2A-puro sgRNA

<400> 1

acttacggta aatggcccgc ctggctgacc gcccaacgac ccccgcccat tgacgtcaat 60

aatgacgtat gttcccatag taacgccaat agggactttc cattgacgtc aatgggtgga 120

gtatttacgg taaactgccc acttggcagt acatcaagtg tatcatatgc caagtacgcc 180

ccctattgac gtcaatgacg gtaaatggcc cgcctggcat tatgcccagt acatgacctt 240

atgggacttt cctacttggc agtacatcta cgtattagtc atcgctatta ccatggtgat 300

gcggttttgg cagtacatca atgggcgtgg atagcggttt gactcacggg gatttccaag 360

tctccacccc attgacgtca atgggagttt gttttggcac caaaatcaac gggactttcc 420

aaaatgtcgt aacaactccg ccccattgac gcaaatgggc ggtaggcgtg tacggtggga 480

ggtctatata agcagagctg gtttagtgaa ccgtcagatc cgctagagat ccgcggccgc 540

taatacgact cactataggg agagccgcca ccatgagctc agagactggc ccagtggctg 600

tggaccccac attgagacgg cggatcgagc cccatgagtt tgaggtattc ttcgatccga 660

gagagctccg caaggagacc tgcctgcttt acgaaattaa ttgggggggc cggcactcca 720

tttggcgaca tacatcacag aacactaaca agcacgtcga agtcaacttc atcgagaagt 780

tcacgacaga aagatatttc tgtccgaaca caaggtgcag cattacctgg tttctcagct 840

ggagcccatg cggcgaatgt agtagggcca tcactgaatt cctgtcaagg tatccccacg 900

tcactctgtt tatttacatc gcaaggctgt accaccacgc tgacccccgc aatcgacaag 960

gcctgcggga tttgatctct tcaggtgtga ctatccaaat tatgactgag caggagtcag 1020

gatactgctg gagaaacttt gtgaattata gcccgagtaa tgaagcccac tggcctaggt 1080

atccccatct gtgggtacga ctgtacgttc ttgaactgta ctgcatcata ctgggcctgc 1140

ctccttgtct caacattctg agaaggaagc agccacagct gacattcttt accatcgctc 1200

ttcagtcttg tcattaccag cgactgcccc cacacattct ctgggccacc gggttgaaaa 1260

gcggcagcga gactcccggg acctcagagt ccgccacacc cgaaagtgat aaaaagtatt 1320

ctattggttt agccatcggc actaattccg ttggatgggc tgtcataacc gatgaataca 1380

aagtaccttc aaagaaattt aaggtgttgg ggaacacaga ccgtcattcg attaaaaaga 1440

atcttatcgg tgccctccta ttcgatagtg gcgaaacggc agaggcgact cgcctgaaac 1500

gaaccgctcg gagaaggtat acacgtcgca agaaccgaat atgttactta caagaaattt 1560

ttagcaatga gatggccaaa gttgacgatt ctttctttca ccgtttggaa gagtccttcc 1620

ttgtcgaaga ggacaagaaa catgaacggc accccatctt tggaaacata gtagatgagg 1680

tggcatatca tgaaaagtac ccaacgattt atcacctcag aaaaaagcta gttgactcaa 1740

ctgataaagc ggacctgagg ttaatctact tggctcttgc ccatatgata aagttccgtg 1800

ggcactttct cattgagggt gatctaaatc cggacaactc ggatgtcgac aaactgttca 1860

tccagttagt acaaacctat aatcagttgt ttgaagagaa ccctataaat gcaagtggcg 1920

tggatgcgaa ggctattctt agcgcccgcc tctctaaatc ccgacggcta gaaaacctga 1980

tcgcacaatt acccggagag aagaaaaatg ggttgttcgg taaccttata gcgctctcac 2040

taggcctgac accaaatttt aagtcgaact tcgacttagc tgaagatgcc aaattgcagc 2100

ttagtaagga cacgtacgat gacgatctcg acaatctact ggcacaaatt ggagatcagt 2160

atgcggactt atttttggct gccaaaaacc ttagcgatgc aatcctccta tctgacatac 2220

tgagagttaa tactgagatt accaaggcgc cgttatccgc ttcaatgatc aaaaggtacg 2280

atgaacatca ccaagacttg acacttctca aggccctagt ccgtcagcaa ctgcctgaga 2340

aatataagga aatattcttt gatcagtcga aaaacgggta cgcaggttat attgacggcg 2400

gagcgagtca agaggaattc tacaagttta tcaaacccat attagagaag atggatggga 2460

cggaagagtt gcttgtaaaa ctcaatcgcg aagatctact gcgaaagcag cggactttcg 2520

acaacggtag cattccacat caaatccact taggcgaatt gcatgctata cttagaaggc 2580

aggaggattt ttatccgttc ctcaaagaca atcgtgaaaa gattgagaaa atcctaacct 2640

ttcgcatacc ttactatgtg ggacccctgg cccgagggaa ctctcggttc gcatggatga 2700

caagaaagtc cgaagaaacg attactccat ggaattttga ggaagttgtc gataaaggtg 2760

cgtcagctca atcgttcatc gagaggatga ccaactttga caagaattta ccgaacgaaa 2820

aagtattgcc taagcacagt ttactttacg agtatttcac agtgtacaat gaactcacga 2880

aagttaagta tgtcactgag ggcatgcgta aacccgcctt tctaagcgga gaacagaaga 2940

aagcaatagt agatctgtta ttcaagacca accgcaaagt gacagttaag caattgaaag 3000

aggactactt taagaaaatt gaatgcttcg attctgtcga gatctccggg gtagaagatc 3060
gatttaatgc gtcacttggt acgtatcatg acctcctaaa gataattaaa gataaggact 3120
tcctggataa cgaagagaat gaagatatct tagaagatat agtgttgact cttaccctct 3180
ttgaagatcg ggaaatgatt gaggaaagac taaaaacata cgctcacctg ttcgacgata 3240
aggttatgaa acagttaaag aggcgtcgct atacgggctg gggacgattg tcgcggaaac 3300
ttatcaacgg gataagagac aagcaaagtg gtaaaactat tctcgatttt ctaaagagcg 3360
acggcttcgc caataggaac tttatgcagc tgatccatga tgactcttta accttcaaag 3420

aggatataca aaaggcacag gtttccggac aaggggactc attgcacgaa catattgcga 3480
atcttgctgg ttcgccagcc atcaaaaagg gcatactcca gacagtcaaa gtagtggatg 3540
agctagttaa ggtcatggga cgtcacaaac cggaaaacat tgtaatcgag atggcacgcg 3600
aaaatcaaac gactcagaag gggcaaaaaa acagtcgaga gcggatgaag agaatagaag 3660
agggtattaa agaactgggc agccagatct taaaggagca tcctgtggaa aatacccaat 3720
tgcagaacga gaaactttac ctctattacc tacaaaatgg aagggacatg tatgttgatc 3780
aggaactgga cataaaccgt ttatctgatt acgacgtcga tcacattgta ccccaatcct 3840

ttttgaagga cgattcaatc gacaataaag tgcttacacg ctcggataag aaccgaggga 3900
aaagtgacaa tgttccaagc gaggaagtcg taaagaaaat gaagaactat tggcggcagc 3960
tcctaaatgc gaaactgata acgcaaagaa agttcgataa cttaactaaa gctgagaggg 4020
gtggcttgtc tgaacttgac aaggccggat ttattaaacg tcagctcgtg gaaacccgcc 4080
aaatcacaaa gcatgttgca cagatactag attcccgaat gaatacgaaa tacgacgaga 4140
acgataagct gattcgggaa gtcaaagtaa tcactttaaa gtcaaaattg gtgtcggact 4200
tcagaaagga ttttcaattc tataaagtta gggagataaa taactaccac catgcgcacg 4260

acgcttatct taatgccgtc gtagggaccg cactcattaa gaaatacccg aagctagaaa 4320
gtgagtttgt gtatggtgat tacaaagttt atgacgtccg taagatgatc gcgaaaagcg 4380
aacaggagat aggcaaggct acagccaaat acttctttta ttctaacatt atgaatttct 4440
ttaagacgga aatcactctg gcaaacggag agatacgcaa acgaccttta attgaaacca 4500
atggggagac aggtgaaatc gtatgggata agggccggga cttcgcgacg gtgagaaaag 4560
ttttgtccat gccccaagtc aacatagtaa agaaaactga ggtgcagacc ggagggtttt 4620
caaaggaatc gattcttcca aaaaggaata gtgataagct catcgctcgt aaaaaggact 4680

gggacccgaa aaagtacggt ggcttcgata gccctacagt tgcctattct gtcctagtag 4740

tggcaaaagt tgagaaggga aaatccaaga aactgaagtc agtcaaagaa ttattgggga 4800

taacgattat ggagcgctcg tcttttgaaa agaaccccat cgacttcctt gaggcgaaag 4860

gttacaagga agtaaaaaag gatctcataa ttaaactacc aaagtatagt ctgtttgagt 4920

tagaaaatgg ccgaaaacgg atgttggcta gcgccggaga gcttcaaaag gggaacgaac 4980

tcgcactacc gtctaaatac gtgaatttcc tgtatttagc gtcccattac gagaagttga 5040

aaggttcacc tgaagataac gaacagaagc aacttttgt tgagcagcac aaacattatc 5100

tcgacgaaat catagagcaa atttcggaat tcagtaagag agtcatccta gctgatgcca 5160

atctggacaa agtattaagc gcatacaaca agcacaggga taaacccata cgtgagcagg 5220

cggaaaatat tatccatttg tttactctta ccaacctcgg cgctccagcc gcattcaagt 5280

attttgacac aacgatagat cgcaaacgat acacttctac caaggaggtg ctagacgcga 5340

cactgattca ccaatccatc acgggattat atgaaactcg gatagatttg tcacagcttg 5400

ggggtgactc tggtggttct actaatctgt cagatattat tgaaaaggag accggtaagc 5460

aactggttat ccaggaatcc atcctcatgc tcccagagga ggtggaagaa gtcattggga 5520

acaagccgga aagcgatata ctcgtgcaca ccgcctacga cgagagcacc gacgagaatg 5580

tcatgcttct gactagcgac gcccctgaat acaagccttg ggctctggtc atacaggata 5640

gcaacggtga gaacaagatt aagatgctct ctggtggttc tcccaagaag aagaggaaag 5700

tctaacagca gagatccagt ttatcgatga gtaattcata caaaaggact cgcccctgcc 5760

ttggggaatc ccagggaccg tcgttaaact cccactaacg tagaacccag agatcgctgc 5820

gttcccgccc cctcacccgc ccgctctcgt catcactgag gtggagaaga gcatgcgtga 5880

ggctccggtg cccgtcagtg ggcagagcgc acatcgccca cagtccccga gaagttgggg 5940

ggaggggtcg gcaattgaac cggtgcctag agaaggtggc gcggggtaaa ctgggaaagt 6000

gatgtcgtgt actggctccg cctttttccc gagggtgggg gagaaccgta tataagtgca 6060

gtagtcgccg tgaacgttct ttttcgcaac gggtttgccg ccagaacaca ggtaagtgcc 6120

gtgtgtggtt cccgcgggcc tggcctcttt acgggttatg gcccttgcgt gccttgaatt 6180

acttccacgc ccctggctgc agtacgtgat tcttgatccc gagcttcggg ttggaagtgg 6240

gtgggagagt tcgaggcctt gcgcttaagg agcccccttcg cctcgtgctt gagttgaggc 6300

ctggcttggg cgctggggcc gccgcgtgcg aatctggtgg caccttcgcg cctgtctcgc 6360

tgctttcgat aagtctctag ccatttaaaa tttttgatga cctgctgcga cgctttttttt 6420

ctggcaagat agtcttgtaa atgcgggcca agatctgcac actggtattt cggtttttgg 6480

ggccgcgggc ggcgacgggg cccgtgcgtc ccagcgcaca tgttcggcga ggcggggcct 6540

gcgagcgcgg ccaccgagaa tcggacgggg gtagtctcaa gctggccggc ctgctctggt 6600

gcctggcctc gcgccgccgt gtatcgcccc gccctgggcg gcaaggctgg cccggtcggc 6660

accagttgcg tgagcggaaa gatggccgct tcccggccct gctgcaggga gctcaaaatg 6720

gaggacgcgg cgctcgggag agcgggcggg tgagtcaccc acacaaagga aaagggcctt 6780

tccgtcctca gccgtcgctt catgtgactc cacggagtac cgggcgccgt ccaggcacct 6840

cgattagttc tcgagctttt ggagtacgtc gtctttaggt tgggggggagg ggttttatgc 6900

gatggagttt ccccacactg agtgggtgga gactgaagtt aggccagctt ggcacttgat 6960

gtaattctcc ttggaatttg ccctttttga gtttggatct tggttcattc tcaagcctca 7020

gacagtggtt caaagttttt ttcttccatt tcaggtgtcg tgaggatcta tttccggtga 7080

attcctcgag actagttcta gatggtgagc aagggcgagg aggataacat ggccatcatc 7140

aaggagttca tgcgcttcaa ggtgcacatg gagggctccg tgaacggcca cgagttcgag 7200

atcgagggcg agggcgaggg ccgcccctac gagggcaccc agaccgccaa gctgaaggtg 7260

accaagggtg gccccctgcc cttcgcctgg gacatcctgt cccctcagtt catgtacggc 7320

tccaaggcct acgtgaagca ccccgccgac atccccgact acttgaagct gtccttcccc 7380

gagggcttca agtgggagcg cgtgatgaac ttcgaggacg gcggcgtggt gaccgtgacc 7440

caggactcct ccctgcagga cggcgagttc atctacaagg tgaagctgcg cggcaccaac 7500

ttcccctccg acggccccgt aatgcagaag aagaccatgg gctgggaggc ctcctccgag 7560

cggatgtacc ccgaggacgg cgccctgaag ggcgagatca agcagaggct gaagctgaag 7620

gacggcggcc actacgacgc tgaggtcaag accacctaca aggccaagaa gcccgtgcag 7680

ctgcccggcg cctacaacgt caacatcaag ttggacatca cctcccacaa cgaggactac 7740

accatcgtgg aacagtacga acgcgccgag ggccgccact ccaccggcgg catggacgag 7800

ctgtacaagg agggccgggg cagcctgctg acctgcggcg acgtggagga gaaccccggc 7860

cccatgaccg agtacaagcc cacggtgcgc ctcgccaccc gcgacgacgt ccccagggcc 7920

gtacgcaccc tcgccgccgc gttcgccgac taccccgcca cgcgccacac cgtcgatccg 7980

gaccgccaca tcgagcgggt caccgagctg caagaactct tcctcacgcg cgtcgggctc 8040

gacatcggca aggtgtgggt cgcggacgac ggcgccgcgg tggcggtctg gaccacgccg 8100

gagagcgtcg aagcgggggc ggtgttcgcc gagatcggcc cgcgcatggc cgagttgagc 8160

ggttcccggc tggccgcgca gcaacagatg gaaggcctcc tggcgccgca ccggcccaag 8220

gagcccgcgt ggttcctggc caccgtcggc gtctcgcccg accaccaggg caagggtctg 8280

ggcagcgccg tcgtgctccc cggagtggag gcggccgagc gcgccggggt gcccgccttc 8340

ctggagacct ccgcgccccg caacctcccc ttctacgagc ggctcggctt caccgtcacc 8400

gccgacgtcg aggtgcccga aggaccgcgc acctggtgca tgacccgcaa gcccggtgcc 8460

tgaacgcgtt aagtcaccca gctttcttgt acaaagtggt gataactcta gagaattcac 8520
tcctcaggtg caggctgcct atcagaaggt ggtggctggt gtggccaatg ccctggctca 8580
caaataccac tgagatcttt ttccctctgc caaaaattat ggggacatca tgaagcccct 8640
tgagcatctg acttctggct aataaaggaa atttattttc attgcaatag tgtgttggaa 8700
ttttttgtgt ctctcactcg gaaggacata tgggagggca aatcatttaa aacatcagaa 8760
tgagtatttg gtttagagtt tggcaacata tgccatatgc tggctgccat gaacaaaggt 8820
ggctataaag aggtcatcag tatatgaaac agccccctgc tgtccattcc ttattccata 8880

gaaaagcctt gacttgaggt tagatttttt ttatattttg ttttgtgtta tttttttctt 8940
taacatccct aaaattttcc ttacatgttt tactagccag atttttcctc ctctcctgac 9000
tactcccagt catagctgtc cctcttctct tatgaagatc cctcgacctg cagcccaaaa 9060
aaaagcaccg actcggtgcc acttttttcaa gttgataacg gactagcctt attttaactt 9120
gctatttcta gctctaaaac nnnnnnnnnn nnnnnnnnnn cggtgtttcg tcctttccac 9180
aagatatata aagccaagaa atcgaaatac tttcaagtta cggtaagcat atgatagtcc 9240
attttaaaac ataattttaa aactgcaaac tacccaagaa attattactt tctacgtcac 9300

gtattttgta ctaatatctt tgtgtttaca gtcaaattaa ttctaattat ctctctaaca 9360
gccttgtatc gtatatgcaa atatgaagga atcatgggaa ataggccctc ttcctgcccg 9420
acctt 9425

<210> 2
<211> 20
<212> DNA
<213> Artificial Sequence
<220><223> sgRNA-1
<400> 2
atgggtgcag caaaccacca 20
<210> 3
<211> 20
<212> DNA
<213> Artificial Sequence
<220><223> sgRNA-3

<400> 3

gaaataccta atgtagatga 20

<210> 4

<211> 67

<212> DNA

<213> Artificial Sequence

<220><223> L1 site for

<400> 4

acactctttc cctacacgac gctcttccga tctnnnnnnn nnnnnnnnna cacagggagg 60

ggaacat 67

<210> 5

<211> 49

<212> DNA

<213> Artificial Sequence

<220><223> L1 site rev

<400> 5

gactggagtt cagacgtgtg ctcttccgat cttgccatgg tggtttgct 49

<210> 6

<211> 61

<212> DNA

<213> Artificial Sequence

<220><223> sgRNA-1 sgRNA for

<400> 6

tttcttggct ttatatatct tgtggaaagg acgaaacacc gatgggtgca gcaaaccacc 60

a 61

<210> 7

<211> 61

<212> DNA

<213> Artificial Sequence

<220><223> sgRNA-1 sgRNA rev

<400> 7

gactagcctt attttaactt gctatttcta gctctaaaac tggtggtttg ctgcacccat 60

c 61

<210> 8

<211> 61

<212> DNA

<213> Artificial Sequence

<220><223> sgRNA-3 sgRNA for

<400> 8

tttcttggct ttatatatct tgtggaaagg acgaaacacc ggaaatacct aatgtagatg 60

a 61

<210> 9

<211> 61

<212> DNA

<213> Artificial Sequence

<220><223> sgRNA-3 sgRNA rev

<400> 9

gactagcctt attttaactt gctatttcta gctctaaaac tcatctacat taggtatttc 60

c 61